Equitable learning via dissimilar variable grouping for synthetic healthcare data

Digital Health
XX(X):1–17

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Hyeonggeun Yun[1], **Hanna Suominen**[1, 2, 3], **Amanda S. Barnard**[1]

Abstract

Machine learning models are increasingly applied in clinical and biomedical settings, yet their complexity can obscure the transparency in decision-making processes and risk propagating biases. This issue needs to be found and fixed early before applying the models in real patients. Synthetic healthcare data can be an efficient way to test preliminary models, yet prone to implicitly learning the bias from the original source. To address these concerns, this paper proposes "SHIELD: A SHapley and Information-theory based framework for Equitable Learning via Dissimilar variable grouping". SHIELD combines dissimilarity-driven variable grouping with transparent latent representations to mitigate proxy bias and enhance equitable learning of the resulting model. By constructing a dissimilarity matrix based on conditional mutual information, variables are grouped to weaken spurious correlations that may amplify unfairness. Group-specific autoencoders learn latent representations while preserving their decoders to map back to the original variables for interpretability. This automation is more effective than clustering similar variables, fixes problematic groups post-hoc, and produces a more equitable distribution of variable importance than ungrouped cases. Experiments on Australian synthetic healthcare data demonstrate the following two key findings: First, the three proposed grouping approaches (greedy, bicriterion, and K-plus) achieved notable improvements in various fairness metrics by 13.32% on average. Second, while an average reduction of 1.38% in predictive performance was observed, it remained within acceptable limits for clinical applications, demonstrating the feasibility of this fairness-performance trade-off. Overall, SHIELD integrates dissimilarity-based grouping, latent representation learning, and explanation-level auditing to promote equitable and explainable machine learning for health informatics.

Keywords

Conditional mutual information, Equitable learning, Explainable artificial intelligence, Health informatics, Machine learning, Shapley additive explanations, Synthetic data, Variable grouping

Introduction

Artificial Intelligence (AI) and Machine Learning (ML) systems are increasingly deployed in healthcare to assist with diagnosis, triage, and treatment planning. Here, AI refers to computational methods that emulate human cognitive tasks (e.g., pattern recognition, reasoning, and decision-making) in clinical workflows, whereas ML is a subset of AI that learns predictive functions directly from data rather than via explicit rule-based programming[1]. Their application in healthcare is significant, as model outputs can directly influence clinical decisions and resource allocation[2]. When the reasoning behind a model's prediction is opaque, clinicians may be unable to verify or contest its recommendations, and patients may lose trust in automated systems. Failures in transparency can also obscure insights about the fairness of models[3]; for instance, some models have been found to allocate fewer resources to patients with minority demographics, and some risk prediction tools have systematically underestimated disease risk for disadvantaged groups[3,4]. Explainable and interpretable approaches (i.e., objective of eXplainable AI (XAI)) are therefore increasingly advocated in healthcare, both to reduce diagnostic and treatment errors and to support meaningful human oversight of AI-assisted decisions[5–7]. As recent editorials in the *Journal of the American Medical Informatics Association* (JAMIA) note, explainability in clinical ML should be evaluated not only by interpretability and fidelity but also by clinical value[8,9]. Hence, this paper aims to make the explanations both understandable to humans and faithful to the underlying model, while supporting equitable and safe care.

In parallel with these developments, synthetic healthcare data have emerged as a promising complement to real-world datasets. Synthetic data can be shared without infringing patient privacy, allowing research and algorithmic benchmarking under fewer regulatory constraints. Recent reviews highlight that well-designed synthetic data resources can accelerate the development

[1] School of Computing, The Australian National University, Australia
[2] School of Medicine and Psychology, The Australian National University, Australia
[3] Department of Computing, University of Turku, Finland

Corresponding author:
Hyeonggeun Yun
Email: geun.yun@anu.edu.au

and validation of medical AI/ML systems while mitigating ethical and privacy risks inherent in the use of real patient records [10,11]. In particular, platforms such as Synthea produce detailed patient-level electronic health records that emulate real clinical encounters using generative rules based on medical logic modules and epidemiological statistics [12]. These datasets can potentially expedite proof-of-concept modelling and enable large-scale fairness analyses that would otherwise be infeasible under data-access restrictions. However, the use of synthetic data does not automatically guarantee fairness or fidelity. If the synthetic generator reflects biased priors or incomplete real-world distributions, it can propagate or even amplify inequities present in the source data [13]. Thus, models trained on synthetic records may learn spurious or proxy associations that misrepresent clinical reality, particularly for under-represented subgroups.

Methodologically, the intersection between synthetic data, XAI, and algorithmic fairness remains underexplored. Many structured clinical datasets contain highly correlated or redundant variables (in ML, variables tend to be called features). During training, unconstrained optimisation can encourage models to rely heavily on dominant predictors, which can overshadow the influence of weaker but clinically meaningful variables [2,14]. When these dominant predictors correlate with sensitive attributes or their proxies, models inadvertently learn biased decision boundaries. This issue can also persist in synthetic datasets, as generative processes are typically designed to reproduce the statistical dependencies present in the original data, which may include proxy relationships [15]. Post-hoc explanation methods can reveal such imbalances by attributing high importance to correlated variables [16], but understanding these feature reliance patterns alone does not determine the fairness of models, motivating the need to consider fairness alongside explainability.

Moreover, while fairness interventions can be introduced during model training, many existing approaches, such as adversarial debiasing or fairness regularisation, modify latent representations in ways that diminish interpretability for clinicians [17,18]. However, ideally, methods that aim to promote fairness in healthcare modelling should preserve explanatory clarity in the variables included in (or excluded from) the model in a way familiar to domain experts; the lack of explainability is recognised as a major barrier to clinicians' confidence and trust in AI/ML systems with XAI being identified as a step towards system trustworthiness [16].

This paper addresses these methodological and ethical concerns through the proposed SHIELD: a Shapley and Information-theory based framework for Equitable Learning via Dissimilar variable grouping. SHapley Additive Explanations (SHAP) [19] is a game-theoretic method that quantifies each variable's marginal contribution to a prediction. Grounded in Shapley values from cooperative game theory, SHAP satisfies key statistical properties, local accuracy, missingness, and consistency, making it an advanced, principled approach for evaluating model reasoning. SHIELD constructs a Conditional Mutual Information (CMI) matrix to identify and separate highly correlated or proxy variables, thereby weakening spurious associations that can amplify bias. Each group is encoded using a dedicated autoencoder whose decoder is retained to map latent representations back to the original variables, ensuring interpretability through SHAP attribution based on the decoder weights, where SHAP provides model-agnostic, statistically consistent estimate of variable influence beyond conventional performance metrics. That is, rather than merely an explainability aid, this paper deploys SHAP as a statistical evaluation tool (e.g., Figure 4) for fairness auditing across prediction and explanation levels. The aim is to redistribute model reliance more equitably across variables, thereby improving parity in both what and why the model predicts, while monitoring predictive performance to ensure that gains in fairness and explainability do not come at unacceptable cost.

Methodology

This section details the end-to-end methodology of SHIELD which shows how data collection, data preprocessing, variable grouping, model training, and evaluation interconnect (Figure 1). This study experiments with a tabular synthetic healthcare dataset, meaning the dataset consists of columns (variables) and rows (instances or individual patient records), which the SHAP analysis is focused at variable level. Once the dataset is collected, the SHIELD pipeline begins by preparing the input variables through imputation, encoding, and standardisation, ensuring that all variables are usable later. These preprocessed variables are then split into training/validation folds and a hold-out test set.

On the training side, SHIELD computes a CMI dissimilarity matrix to identify pairs of variables to group dissimilar variables together through different grouping methods (Greedy, Bicriterion, and K-plus). Each resulting group is encoded into a compact latent representation by a group-specific autoencoder. These grouped latent variables, as well as the original variables in the ungrouped baseline, are then used to train several ML models (Logistic Regression (LR) [20], Multi-Layer Perceptron (MLP) [21], and XGBoost (XGB) [22]), whose hyperparameters are tuned using five-fold stratified cross-validation. After the best model is selected, it is retrained on the full training data and evaluated on the test set.

The evaluation has three interconnected layers. Predictive performance metrics (accuracy, precision, recall, and F1-score) assess the correctness of the model's outputs. Fairness metrics (equal opportunity, equalised odds, predictive parity, N-Sigma error rate, distance to origin, and separability in bias quadrant) assess prediction-level and explanation-level parity across protected groups (i.e., race in this study). Finally, SHAP analysis is performed to attribute model predictions back to individual variables. In grouped configurations, SHAP is first computed at the latent level and then decomposed back to the original variables

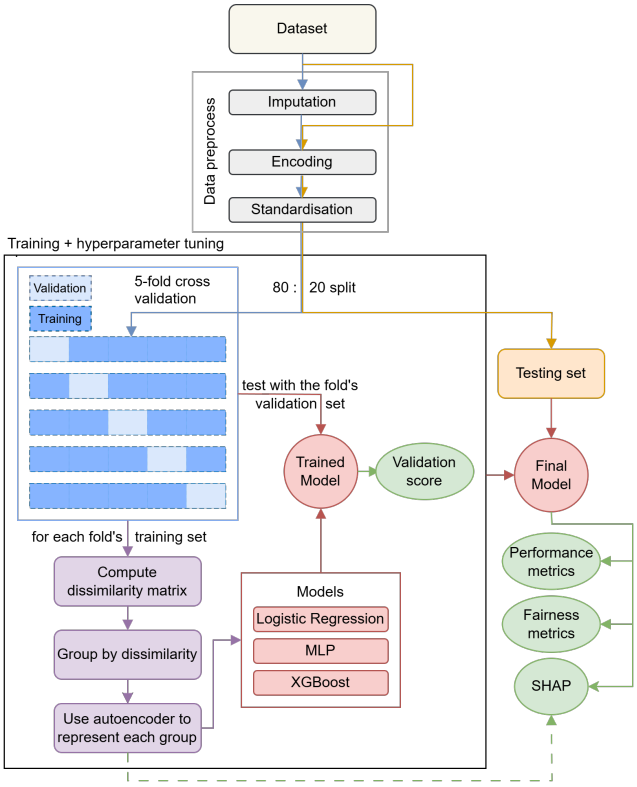


Figure 1. End-to-end workflow of SHIELD.

using decoder-based attribution redistribution, ensuring traceability to the original clinical variables.

Types of learning

SHIELD blends unsupervised structure discovery with supervised prediction to support equitable and interpretable learning. Unsupervised methods[23] learn structure directly from the input variables without using class labels, whereas supervised learning[24] requires labelled outcomes during model training to learn a predictive mapping from input variables to the target variable. SHIELD integrates both these types of learning in its end-to-end workflow (Figure 4) as follows:

The CMI computation, the construction of the dissimilarity matrix, and the dissimilarity-based grouping are all unsupervised, as they rely solely on relationships between input variables. The autoencoders used to derive latent representations for each variable group are also unsupervised, because they aim to reconstruct the input variables rather than predict labels.

In contrast, the classifiers are supervised models, trained on labelled data to predict disease outcomes. Furthermore, the evaluation pipeline, including performance metrics, fairness metrics, and SHAP analysis, results in from supervised models, since classification accuracy, group-parity measures, and attributional consistency must all be computed against known ground-truth outcomes.

Data collection and preprocessing

All data used in this study were pre-existing, synthetic and publicly available from Commonwealth Scientific and Industrial Research Organisation (CSIRO) with Creative Commons Attribution 4.0 International Licence[25]. The dataset, synthetically generated by Synthea[12], covered a range of clinical classification tasks (Table 1). Some diseases had an imbalanced prevalence distribution (e.g., chronic kidney disease and Alzheimer's) as a typical inherent property of medical datasets, while others had balanced prevalence (e.g., diabetes and hypertension). This reflects a conditional sampling of synthetic patients who are more likely to be diagnosed with diseases that prompt them to visit healthcare facilities than the general population.

Attribute	Value
D (# of variables)	45
N (# of instances)	89,419
C (# of Classes)	4
Diabetes prevalence	56%
Hypertension prevalence	43%
Chronic kidney disease prevalence	28%
Alzheimer's prevalence	3%

Table 1. Australian synthetic healthcare data with Synthea[25]. Note that each encounter (row) could have diagnosed with multiple diseases (hence the sum of prevalences is over 100%).

The following three core operations formed the backbone of the preprocessing pipeline: First, classifier-based imputation with XGB[22] replaced missing values so that each observation remained usable for model training. This imputation was most suitable given the context of the datasets, as it supports complex variable interactions and high missing rates without requiring full data encoding upfront. Second, categorical variables were encoded as numerical representations for compatibility with most algorithms. Specifically, this was done via one-hot encoding[26] for unordered predictors, label encoding for the target, and ordinal encoding of ranked predictors. Third, Z-score standardisation[27] mitigated the dominance of variables measured on larger numerical ranges.

Variable grouping by dissimilarity

In this study, variable grouping served a critical role in promoting fairness[28–30], beyond just enhancing dimensionality reduction[31]. The central idea was to separate variables that were highly correlated with each other, particularly those that might have acted as proxy variables for sensitive attributes, into distinct groups. Proxy variables, although not explicitly labelled as sensitive (e.g., socioeconomic status in place of race), could still lead to biased model behaviour if their collective influence remained unchecked[3,32]. Hence, the notion of grouping variables by dissimilarity, rather than by similarity, was a deliberate strategy to mitigate such risks[33]. If similar variables, including potential proxies, were grouped together, their joint effects might become more pronounced, leading to biased representations in latent space. Thus, grouping by similarity intended to consolidate correlated variables would require subsequent adjustments to explicitly

manage proxy variables. Conversely, spreading them across different groups through dissimilarity-based partitioning weakened their impact at the group level and provided a natural form of regularisation against undue influence.

All three grouping methods used in this study relied on an underlying dissimilarity matrix[31]. This matrix quantified the degree to which each pair of variables provides different information with respect to the target variable. In particular, dissimilarity was computed as the complement of CMI[34] between variables to identify variables that contribute unique, non-redundant signals to the prediction task.

Let X_i and X_j denote two input variables, and Y be the target variable. The dissimilarity between X_i and X_j was defined using their CMI given Y, which captured the shared information between the two variables conditional on the outcome variable[35]. Mathematically, CMI is expressed as

$$I(X_i; X_j \mid Y) = \sum_{x_i \in \mathcal{X}_i} \sum_{x_j \in \mathcal{X}_j} \sum_{y \in \mathcal{Y}} p(x_i, x_j, y) \log \left(\frac{p(x_i, x_j \mid y)}{p(x_i \mid y) p(x_j \mid y)} \right),$$

where $\mathcal{X}_i, \mathcal{X}_j$, and $\mathcal{Y}$ denote the sets of all possible values of the random variables X_i, X_j, and Y, respectively, and $p(\cdot)$ denotes the corresponding joint and conditional probability mass functions. To standardise the scale of CMI and obtain a bounded measure of dissimilarity, it was normalised through the sum of marginal entropies:

$$\text{Normalised CMI}_{i,j} = \frac{I(X_i; X_j \mid Y)}{H(X_i) + H(X_j) + \epsilon}, \quad (1)$$

where $H(X)$ is the Shannon entropy of feature X and ϵ is a small constant to prevent division by zero. The resulting dissimilarity was computed as

$$D_{i,j} = 1 - \text{Normalised CMI}_{i,j}.$$

The matrix D with entries $D_{i,j}$ is symmetric and encodes how dissimilar each pair of variables is, serving as the foundational input for all subsequent grouping strategies. Normalised CMI is bounded between 0 and 1, where larger values indicate stronger conditional dependence given Y[34,35]. Consequently, dissimilarity $D_{i,j}$ is also in the same interval $[0, 1]$, with larger values indicating more distinct variables.

Each grouping method, despite relying on different heuristics or optimisation strategies, fulfilled a common objective: maximising dissimilarity within groups to weaken the collective impact of correlated or proxy variables. Two primary metrics were used to assess the quality of variable grouping as follows.

First, diversity[36] quantified the average dissimilarity between variables that belonged to the same group. Formally, for each group G_k containing variable indices $i, j \in G_k$, the diversity was computed as

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

where K is the number of groups. A higher diversity indicated that variables within each group were more distinct from each other.

Second, dispersion[36] was a more conservative non-negative metric, focusing on the minimum pairwise dissimilarity between any two variables within a group.

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

Maximising dispersion ensured that even the most similar pair within each group was as dissimilar as possible, thus enforcing strong baseline for intra-group heterogeneity.

Finding a partition that had the highest possible value for both diversity and dispersion would be ideal. However, such a partition did not always exist as they innately conflicted with one another to a degree. For instance, one could simply merge variables into larger, more varied groups to raise diversity, but this would decrease dispersion, as some pairs within those groups could inevitably be more similar than others.

Greedy approach. The naïve variable grouping strategy adopted a greedy approach that relied on the dissimilarity matrix derived from CMI. The method proceeded as shown in Algorithm 1 given the dissimilarity matrix D and number of groups K: Initial seeds for the groups were selected based on the highest average dissimilarity scores across all variables, ensuring that each group begins with a representative variable that was maximally distinct from others. The algorithm then iteratively assigned the remaining variables to the group for which they exhibited the highest average dissimilarity with existing group members. This greedy assignment continued until all variables were allocated. The simplicity and intuitive heuristic of this method makes it a useful baseline for evaluating more sophisticated grouping approaches.

Bicriterion approach. The bicriterion anticlustering simultaneously maximised two complementary criteria, diversity and dispersion[36] (Algorithm 2). It aimed to avoid configurations where high overall diversity might still allow clusters of closely related (potential proxy) variables as it enforced relatively high dispersion at the same time. The algorithm attempted to approximate a Pareto-optimal set of groupings by using local search heuristics, adjusting the assignment of variables to groups to improve the following objective:

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

where α, β quantified the priorities of each criterion.

K-plus anticlustering approach. The K-plus anticlustering method extended traditional K-means anticlustering by addressing not only the similarity in group means but also discrepancies in high-order distribution moments, such as variance, skewness and kurtosis[31]. The objective was to form groups with maximum internal homogeneity (as opposed to conventional K-means objective), while being similar to each other across multiple statistical dimensions.

Algorithm 1 Greedy Variable Grouping

Input: Dissimilarity matrix D, number of groups K
Initialize K groups with variables having the highest row-wise sum in D
while there are unassigned variables **do**
  for each group g **do**
    for each unassigned variable f **do**
      Compute average dissimilarity between f and all variables in g
    end for
    Assign variable with maximum average dissimilarity to g
  end for
end while
Output: K dissimilar groups

Algorithm 2 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

Algorithm 3 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

Formally, this was achieved by constructing a set of augmented variables derived from the original attributes. These include squared deviations (for variance), cubic deviations (for skewness), and so forth. The combined objective function, known as the K-plus criterion, was a weighted sum of the standard K-means Error Sum of Squares (SSE) and additional SSE terms for each higher-order moment [31]. This formulation allowed for fine-tuned control over the statistical similarity of groups. Optimisation was performed through local search heuristics that iteratively swapped variables between groups to improve the composite objective (Algorithm 3).

Latent representations of groups. Once the variable groups have been identified, it was essential to develop appropriate latent representations for each group to enable downstream model training. The primary goal of this transformation was to condense the information within each group into a compact, yet informative vector that retained the group's key statistical and structural characteristics.

Unlike traditional variable grouping based on similarity, where dimensionality reduction techniques such as Principal Component Analysis (PCA) [37] can capture dominant correlated directions, the groups in SHIELD were intentionally constructed to consist of dissimilar variables. Consequently, summarisation strategies relying on correlation or redundancy were ineffective. Instead, a neural network-based approach using group-specific autoencoders [38] was adopted.

For each group, an autoencoder was trained to learn an efficient encoding of the group's variable set. Formally, let $X^{(k)} \in \mathbb{R}^{n_k \times d_k}$ denote the matrix of d_k variables in the k-th group across n_k samples. A group-specific autoencoder learnt an encoding function $f_k : \mathbb{R}^{d_k} \to \mathbb{R}^{m_k}$ and a decoding function $g_k : \mathbb{R}^{m_k} \to \mathbb{R}^{d_k}$ such that the reconstruction loss was minimised:

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

The latent vector $f_k(X_i^{(k)})$ thus became the representation of group k for sample i, encapsulating the non-redundant, informative essence of the original variables.

To ensure model interpretability, a mapping between each group's latent representation and its original variables was retained. This mapping was essential for decomposing SHAP values to approximate variable-level attributions by analysing decoder weights and sensitivity. It also allowed evaluating fairness metrics with respect to individual variables, ensuring that potential biases could be traced even after dimensionality reduction.

Concretely, under the setting of Formula (2), the decoder's linear layer was expressed as

$$g_k(z) = W_{\mathrm{dec}}^{(k)} z + b^{(k)}, \quad W_{\mathrm{dec}}^{(k)} \in \mathbb{R}^{d_k \times m_k},$$

where $z = f_k(X_i^{(k)}) \in \mathbb{R}^{m_k}$ denotes the latent representation. Then, each column of $W_{\mathrm{dec}}^{(k)}$ described how one latent coordinate contributed to all d_k original variables. Suppose a downstream classifier produces a vector of latent-space attributions

$$\phi^{(k)} = [\phi_1^{(k)}, \ldots, \phi_{m_k}^{(k)}]^T \tag{3}$$

for group k, where $\phi_j^{(k)}$ denotes the attribution assigned to the j-th latent coordinate. To distribute these back to the original variables, the element-wise absolute weight matrix was formed and each latent-to-variable mapping

was normalised so that the contributions summed to one:

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

The element-wise absolute operator was applied to the decoder weights to capture the magnitude of the relationship between each latent coordinate and the reconstructed variables. Decoder weights can take positive or negative values depending on the direction of reconstruction, whereas SHAP values already encode the directional influence of each latent coordinate on the model prediction. Taking the absolute value therefore isolates the strength of the latent–variable dependency while avoiding cancellation effects during attribution redistribution. The subsequent normalisation ensures that the contributions from each latent coordinate are proportionally distributed across the original variables. This operation is applied only during the attribution decomposition step and does not modify the training of the autoencoder. The final variable-level attribution vector for group k was then

$$\phi_{\text{original}}^{(k)} = \tilde{W}^{(k)}\phi^{(k)} \in \mathbb{R}^{d_k}, \tag{4}$$

so that each original variable i inherited

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

capturing both the model's sensitivity in latent space and the decoder's reconstruction-based mapping back to raw inputs.

Because the variable groups form a partition of the original feature space, each group-level attribution vector $\phi_{\text{original}}^{(k)} \in \mathbb{R}^{d_k}$ corresponds to a disjoint subset of the original variables. The global attribution vector over the full feature space is therefore obtained by concatenating these vectors in the original variable order:

$$\phi_{\text{original}} = \left[(\phi_{\text{original}}^{(1)})^T, \ldots, (\phi_{\text{original}}^{(K)})^T\right]^T \in \mathbb{R}^d,$$

where $d = \sum_{k=1}^{K} d_k$. This reconstruction preserves interpretability at the original variable level, enabling both SHAP-driven explanations and fairness assessments after benefiting from the grouping through dimensionality reduction. Once trained on the training data, the group-specific autoencoders were fixed and the same encoding function of f_k was reused to transform the test set, avoiding leakage and preserving the integrity of the learned transformations.

Training

This study evaluated the effectiveness of fairness-aware variable grouping strategies across a diverse set of ML models, including LR[20], MLP[21], and XGB[22] to explore linear, neural and tree-based approaches, respectively. This heterogeneity allowed for a robust assessment of the generalisability and fairness implications of the proposed methodology[14].

Symbol	Name	Range (Type)
Variable Grouping Stage		
K	Number of groups	[2, 10] (integer)
α, β	Bicriterion weights	[0, 1] (float)
w_2, w_3, w_4	Moment weights	[0, 1] (float)
ϵ	Smoothing constant	$[10^{-10}, 10^{-6}]$ (float)
Model Training Stage		
C	Regularisation (LR, XGB)	[0.01, 100] (float)
n_estimators	Number of trees	[50, 500] (integer)
max_depth	Maximum tree depth	[3, 15] (integer)
learning_rate	Learning rate (XGB, MLP)	[0.001, 0.3] (float)
hidden_layers	Hidden layer sizes (MLP)	Varies (tuple)
alpha	L_2 penalty (MLP)	[0.0001, 0.1] (float)
k_{cv}	number of folds in k-fold CV	[3, 10] (integer)

Table 2. Hyperparameters used in variable grouping and model training.

Bayesian optimisation was used to tune hyperparameters[39] as defined in Table 2. The next step was then to perform five-fold stratified cross-validation[40] on the training set, optimising the mean Receiver Operating Characteristic Area Under the Curve (ROC-AUC) over its validation set; ROC-AUC is a widely used classification performance metric that evaluates a model's ability to distinguish between classes across all possible decision thresholds by integrating the true positive rate and false positive rate into a single scalar measure[41]. Each Bayesian search was iterated 30 times, which empirical studies have shown to be sufficient for convergence in comparable settings[42].

Evaluation

Performance metrics. In this study, four widely used metrics, accuracy, precision, recall and F1-score, were employed to provide complementary perspectives on classifier behaviour, especially in datasets exhibiting class imbalance. Although ROC-AUC is theoretically superior to the four metrics as a threshold-independent measure of discriminative performance (and was therefore used during cross-validation), it was not included among the primary evaluation metrics because accuracy, precision, recall and F1-score directly reflect the false-positive and false-negative behaviour at the operating threshold in a more clinically intuitive way. Since ROC-AUC already had ensured strong discriminative capacity during training, the final evaluation emphasised metrics that offer easily interpretable and more efficient performance insights. Each metric was reported for baseline, raw, and

dissimilarity-grouped configurations to highlight trade-offs introduced by the grouping strategies. While accuracy captured overall correctness, it could have obscured poor performance on the minority class for imbalanced dataset[43]. To address this, the F1-score was used as a harmonic mean of precision and recall.

SHAP. Explainability is an indispensable requirement for any fairness-enhancing pipeline deployed in high-stakes domains such as healthcare. SHIELD considers this aspect through the use of SHAP, which quantifies the contribution of each input variable to a model prediction. Given a model f and an input x, the SHAP value for variable j is formally defined as:

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

where F denotes the full set of variables, S denotes a subset not containing j, and $f_S(x)$ denotes the model output when only the variables in subset S are observed. This formulation captures the marginal contribution of variable j, averaged across all possible coalitions, thereby ensuring the axiomatic properties of local accuracy, missingness, and consistency[19]. This makes SHAP fundamentally different from, and complementary to traditional performance metrics. While performance metrics evaluate 'how well' a model predicts, SHAP provides a statistical and functional decomposition of 'why' the model predicts what it does, and is therefore an analytical tool rather than a performance measure[19,44].

In the ungrouped configuration, SHAP values were computed directly for each original variable. In the grouped configuration, the explanation process is mediated by the latent representation z_k, where SHAP values are first approximated at the latent level via Equation (3). These latent attributions were then decomposed to variable-level importances using the decoder weight matrix $W_{\text{dec}}^{(k)}$ via Equation (4). This proportional redistribution ensures that the contribution of each latent factor is faithfully allocated across its constituent variables.

The integration of SHAP into SHIELD thus has a dual role. First, it provides faithful explanations of model predictions at both latent and variable levels, enabling clinicians to scrutinise individual decisions. Second, it serves as a diagnostic tool for evaluating the fairness effects of grouping, since the distribution of SHAP values directly reflects whether predictive power is concentrated in a few variables or more equitably shared. This duality moves SHAP beyond its conventional use as a post-hoc explainability method, positioning it as an integral component of the fairness pipeline.

Fairness metrics. A core starting point in fairness research was group fairness, which is to ensure model performance metrics were comparable across subgroups defined by sensitive attributes such as gender or ethnicity. This research focused on equal opportunity and equalised odds[45].

Equal opportunity required True Positive Rates (TPR) be equal across groups:

$$p(\hat{Y}=1 \mid Y=1, A=1) = p(\hat{Y}=1 \mid Y=1, A=0),$$

where A denotes a binary protected variable. This ensured that individuals in all groups had an equal chance of a beneficial outcome when they genuinely qualified for it, which is a key concern when model decisions could influence healthcare delivery or treatment prioritisation.

Equalised odds strengthened this by also requiring equal False Positive Rates (FPR):

$$p(\hat{Y}=1 \mid Y=0, A=1) = p(\hat{Y}=1 \mid Y=0, A=0).$$

However, equalised odds could sometimes conflict with clinical realities if the underlying base rates genuinely differed due to biological or demographic variation. These datasets, consisting of objective clinical records rather than subjective human ratings, were less likely to reflect historical biases encoded through human judgement. Consequently, equal opportunity was particularly appropriate here since it corrected for unfair treatment without forcing artificial equality where medical evidence supported different base rates[46]. This design choice demonstrated a balance between fairness and respecting the clinical integrity of ground truth labels.

While group fairness metrics exposed mean differences between groups, they did not account for uncertainty due to small sample sizes or high variance in subgroup distributions. Fairness improvements that appear large in percentage terms might be statistically insignificant when sample sizes are small[47]. N-Sigma index was computed to safeguard against over-interpreting noisy fairness estimates[47]:

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

where ϵ_i and σ_i^2 denote the mean and variance of the error rates for group i. This normalised gap ensured that any apparent fairness gains were robust to sampling variation. This step was important when working with health records, where minority group sizes could be limited in real-world hospital datasets.

Prediction parity alone did not guarantee the fairness of a model's internal reasoning. Models that appeared fair at the prediction level could still produce biased explanations, which could undermine trust in contexts where interpretability was critical, such as patient-specific risk scores or variable-driven diagnostic rules[48].

Let $f : \mathbb{R}^d \to \mathbb{R}$ denote the trained prediction model and let $x \in \mathbb{R}^d$ denote a specific input instance. Let $X = (X_1, \ldots, X_d)$ denote the random variable vector representing the data distribution. The SHAP framework decomposes the prediction into additive feature attributions

$$f(x) = \mathbb{E}[f(X)] + \sum_{j=1}^{d} \phi_j(f,x),$$

where $\phi_j(f, x)$ denotes the Shapley attribution assigned to the j-th input variable.

To capture this, explanation bias was audited by measuring differences in local SHAP attributions for protected groups:

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

where ϕ_j denotes the Shapley value for j-th variable. This was then plotted in the bias quadrant alongside prediction-level disparities to reveal how local explanations align (or conflict) with model outcomes. This visualisation allowed interpretation of the four distinct bias regimes:

1. **High prediction bias, High explanation bias:** Both the model's outcomes and its explanations unfairly favoured one group. For instance, if a diabetes readmission model shows higher TPR for males and the SHAP attribution for 'sex' is consistently higher for males, this suggests the model both behaves unfairly and justifies it unfairly, which is perhaps the most concerning scenario.
2. **Low prediction bias, High explanation bias:** Predictions appear fair on average, but explanations revealed that the protected variable still influenced individual decisions in a biased manner. For example, the TPR may be equal for genders, but local attributions for 'sex' are higher for males, suggesting hidden proxy effects.
3. **Low prediction bias, Low explanation bias:** The ideal region, since predictions were equitable and explanations confirmed no undue reliance on sensitive variables. For example, 'sex' contributes negligibly and equally across groups.
4. **High prediction bias, Low explanation bias:** Predictions showed disparities, but explanations did not attribute this to the protected variable itself, indicating the bias likely came from other correlated variables. For instance, the model's TPR is higher for males but 'sex' SHAP values are balanced, suggesting a proxy like 'employment status' might be driving hidden structural bias.

Two metrics were computed to quantify the models' performance in the bias quadrant. First, the average distance to the origin quantified bias magnitude at the instance level by taking the Euclidean norm of each point's coordinates and the origin. Formally, with points $p_i = (\phi_i^{(A)},\ \hat{p}_i - \mu_{A_i})$ after base-rate centring, the metric was computed as $\frac{1}{n}\sum_i \|p_i\|_2$, where smaller values indicated closer proximity to the ideal $(0, 0)$ and therefore lower combined prediction-explanation bias. Second, separability measured how far the privileged and unprivileged groups diverged in this joint space using the Bhattacharyya distance between their empirical Gaussian approximates as follows:

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

where $\bar{\Sigma} = \frac{1}{2}(\Sigma_1 + \Sigma_2)$ and larger D_{B} indicating greater distributional divergence of prediction-explanation behaviour across groups[49]. Hence, the distance to the origin (bias magnitude) and Bhattacharyya separability (group divergence) provided complementary views of fairness in the bias quadrant.

Results

Performance metrics

All configurations of grouping methods and datasets achieved high predictive scores across all metrics ranging from 0.814 to 0.999. On average, the grouped configurations showed a marginal yet systematic reduction of of 1.38% in predictive performance compared with the ungrouped baseline. This minor decline corroborates the fairness and explainability gains from SHIELD were not achieved at the significant expense of predictive capability. Among the grouping strategies, K-plus consistently yielded the lowest performance, with an average decrease of 4.37%. As shown in Figure 2, the colour intensity is largely uniform within each column. This suggests the variation between datasets, rather than the choice of grouping method, primarily accounted for the observed differences in absolute scores.

The relatively weaker performance of K-plus can be attributed to its algorithmic emphasis on higher-order moment matching, such as variance, skewness, and kurtosis, during the anticlustering process. While this constraint enhances statistical balance among groups, it can distort the discriminative structure of the input space by distributing predictive variables across multiple groups. This diffusion of signal likely reduced the effective separability of the latent representations learnt by downstream models, leading to a measurable decline in accuracy, precision, and recall. In contrast, the bicriterion and greedy methods explicitly optimise pairwise dissimilarity directly from the matrix D, thereby preserving dominant variable relationships while still mitigating potential proxy bias and redundancy, which explains their near-parity with the ungrouped baseline.

Performance patterns also varied systematically across disease datasets. The chronic kidney disease (CKD) task exhibited the lowest precision and recall, which directly impacted its F1-score. This result aligns with the dataset's pronounced class imbalance, where the minority class instances were underrepresented and harder to detect. The model thus tended to favour specificity over sensitivity, yielding higher false-negative rates. Conversely, the Alzheimer's classification task recorded the highest scores across all metrics, despite its extreme imbalance. This counterintuitive outcome arises because Alzheimer's cases in the synthetic dataset were simulated with strong, distinctive variable patterns that sharply differentiate positive from negative cases. The distinguishing variable patterns that enabled this strong predictive separability in Alzheimer's also resulted in distinct explanation-level fairness outcomes, as further discussed in the next subsection.

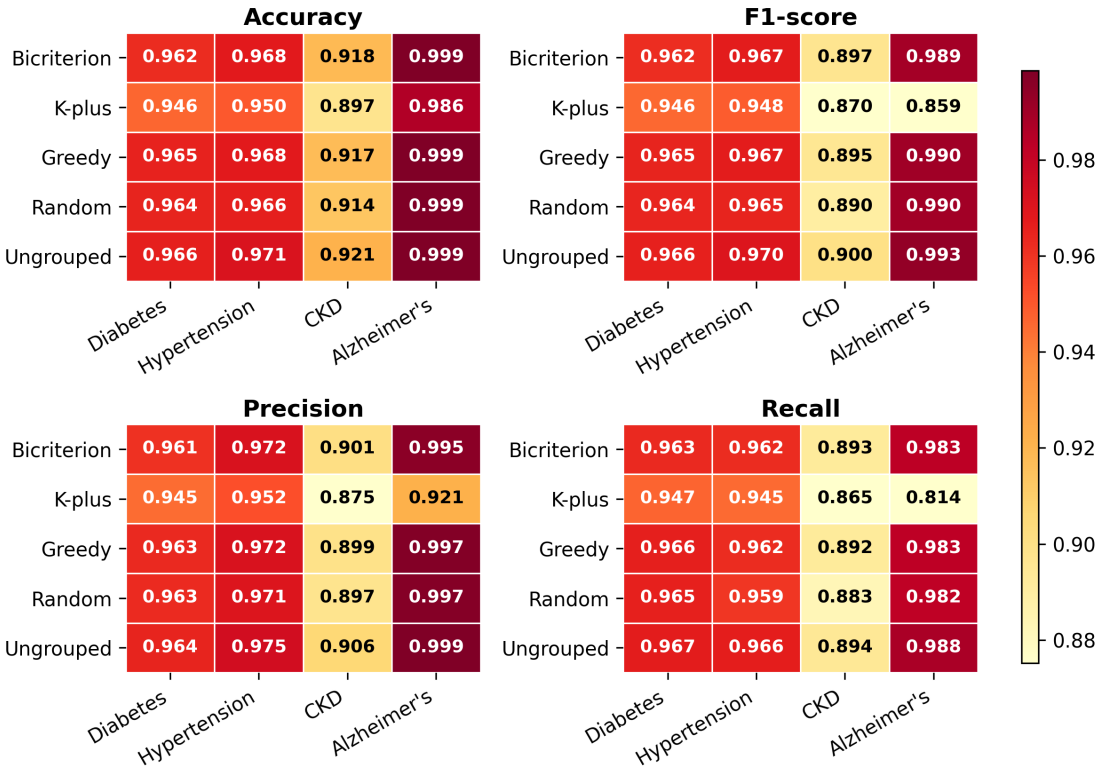


Figure 2. Mean accuracy, F1-score, precision, and recall of models across datasets and grouping methods.

SHAP and fairness metrics

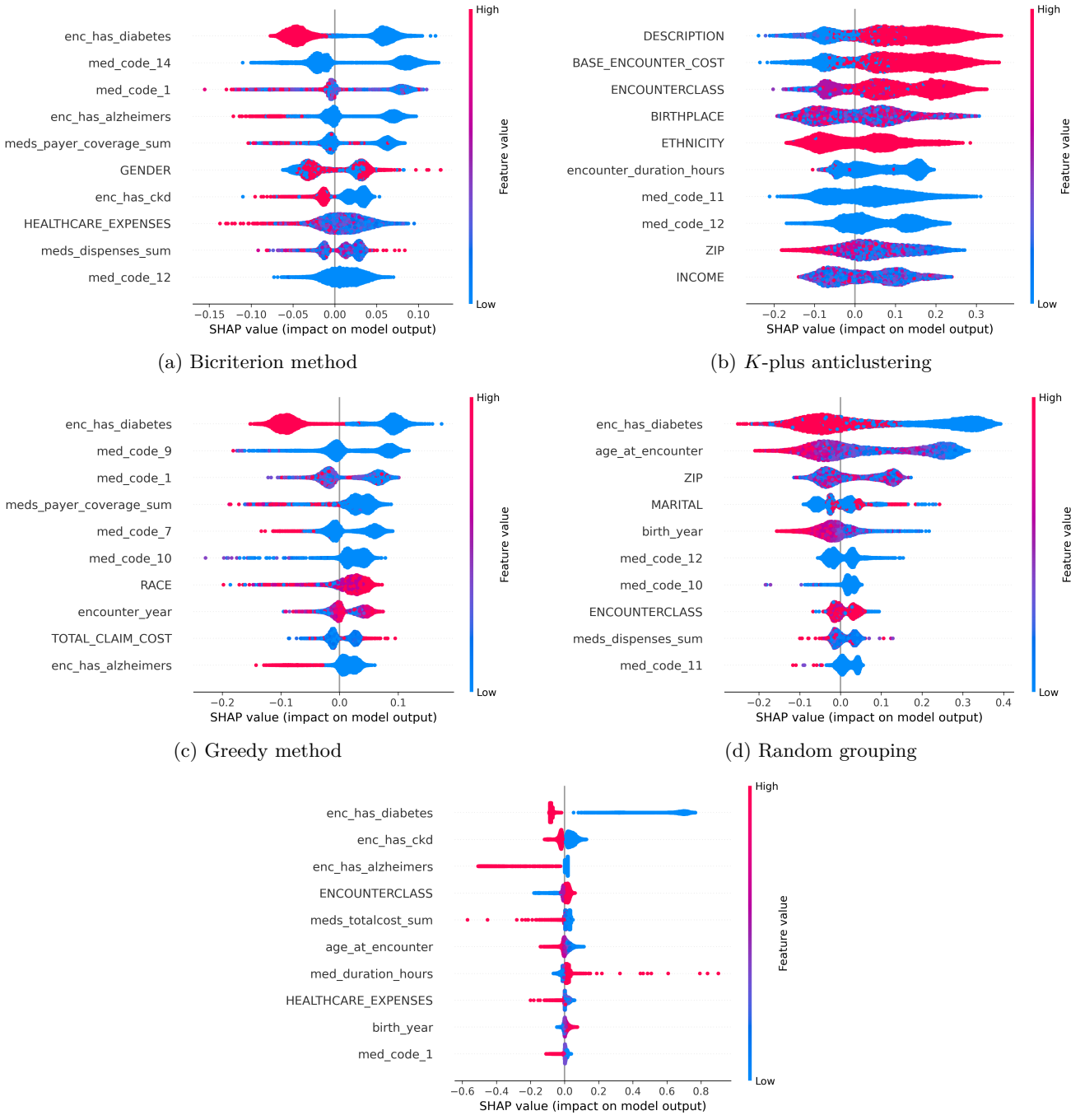
Grouped versus ungrouped. A consistent pattern across datasets was that grouped representations led to more equitable use of variables and instances compared to the ungrouped baseline. As illustrated by Figure 3, the ungrouped case was dominated by a small subset of variables, producing steep drop-offs in importance and leaving many variables with near-zero contribution. This concentration implied that large portions of the variable space were underused, and in some cases entire variables contributed nothing to the model's decision-making. Grouping counteracted this effect by flattening the SHAP distribution: more variables were assigned moderate levels of importance and fewer instances were associated with zero SHAP values. In practice, this meant that grouped models make more use of the available data, essentially reducing 'waste'.

The degree of equitable distribution of the variable attribution was further analysed as follows: A few outlier variables were within a dominating SHAP range (e.g., 0.3 and 0.6) while the entire box and whisker plots sat near 0.0 (Figure 4). All grouping methods resolved this concentration to a different degree, as their boxes were visibly wider with fewer outliers. The number of completely unused variables (dots on the 0.0 axis) decreased from five to one as a result of grouping via dissimilarity. K-plus anticlustering widened the SHAP ranges the most (0.18 median and 0.32 range), followed by random (0.05 median and 0.13 range) and bicriterion (0.04 median and 0.14 range).

Comparison between grouping methods. While grouping generally improved fairness relative to the ungrouped baseline, the extent of improvement differed markedly between methods. Random grouping was the most variable, as it achieved results comparable to more principled methods in some cases, but it also performed the worst in others. For example, in Diabetes, Random achieved the N-Sigma error rate of 0.018, very close to Bicriterion at 0.017, and better than K-plus at 0.024. However, in the Alzheimer's task, Random resulted in relatively poor separability of 0.020 compared to Bicriterion's 0.005 and K-plus' at 0.002.

Greedy grouping, despite its higher computational cost, did not consistently outperform Random. It often produced weaker fairness outcomes than Bicriterion and sometimes even worse than the ungrouped case. For instance, in CKD, Greedy resulted a N-Sigma error rate of 0.019 compared to 0.014 and 0.017 with Bicriterion and ungrouped, respectively. A similar trend was observed in average distance, where Greedy recorded 0.364 in CKD, which was the worst among all methods and worse than the ungrouped baseline at 0.345. Nevertheless, it still achieved the second highest fairness improvement of 19% on average. This indicates that the heuristic strategy of locally maximising dissimilarity did not guarantee globally fairer or more balanced group structures, but still diminished the unjust influence of sensitive attribute to some extent.

Bicriterion consistently produced the strongest and most stable fairness results in both outcome and explanation parity. By explicitly optimising both diversity (high average dissimilarity between groups) and dispersion (ensuring no group is too similar), it led to the highest improvement of fairness metrics on average by 25%. These results confirm that



(a) Bicriterion method

(b) K-plus anticlustering

(c) Greedy method

(d) Random grouping

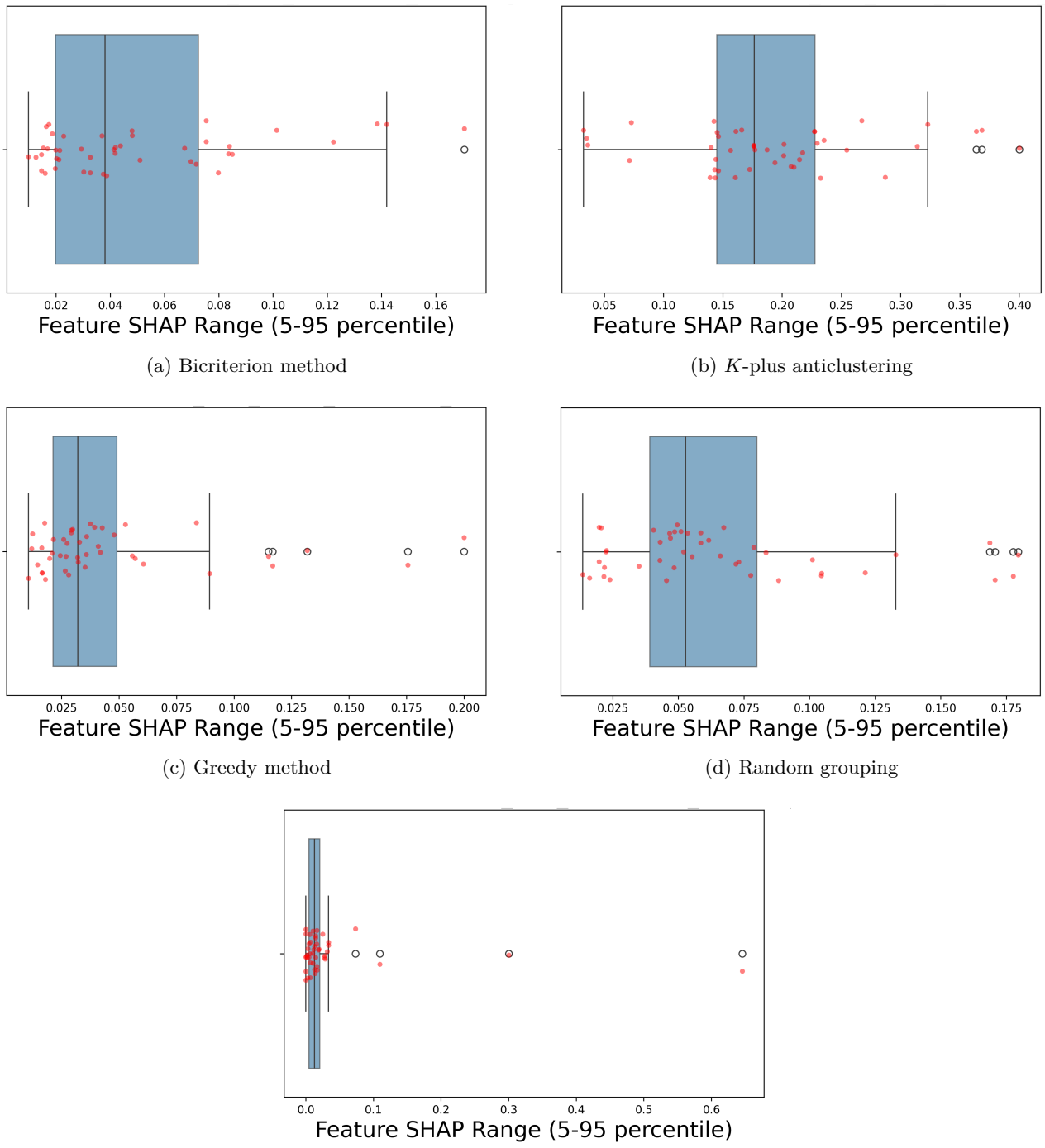
(e) Ungrouped

Figure 3. Comparison between SHAP beeswarm plots of CKD classification via LR across different grouping methods. The term 'Feature' in this figure refers to a variable in the ML context.

an explicit optimisation of diversity and dispersion provided a systematic advantage over heuristic or random strategies.

K-plus excelled at mitigating explanation-level disparities, consistently achieving the lowest bias magnitude and group separability across all datasets. It was the only method to attain near-zero average distance (< 0.01) in every task and almost halved both distance and separability compared with other approaches in Alzheimer's. This indicates that K-plus effectively reduced attributional divergence between privileged and unprivileged groups, yielding nearly indistinguishable local SHAP patterns. However, this strength in explanation parity did not translate to group outcome parity, which was by far the weakest among all methods. It resulted in 111% decline of average equal opportunity, equalised odds and predictive parity compared to ungrouped cases. This highlights that geometric balance in latent representations did not necessarily ensure equitable predictive behaviour.

Comparison between classifiers. The model-wise SHAP summaries reinforce that the grouping effect was not model-specific. Flatter importance spectra were observed than their ungrouped counterparts across all classifiers (Figure 5), suggesting less reliance on a few dominant variables. The most contributing variables



(a) Bicriterion method

(b) K-plus anticlustering

(c) Greedy method

(d) Random grouping

(e) Ungrouped

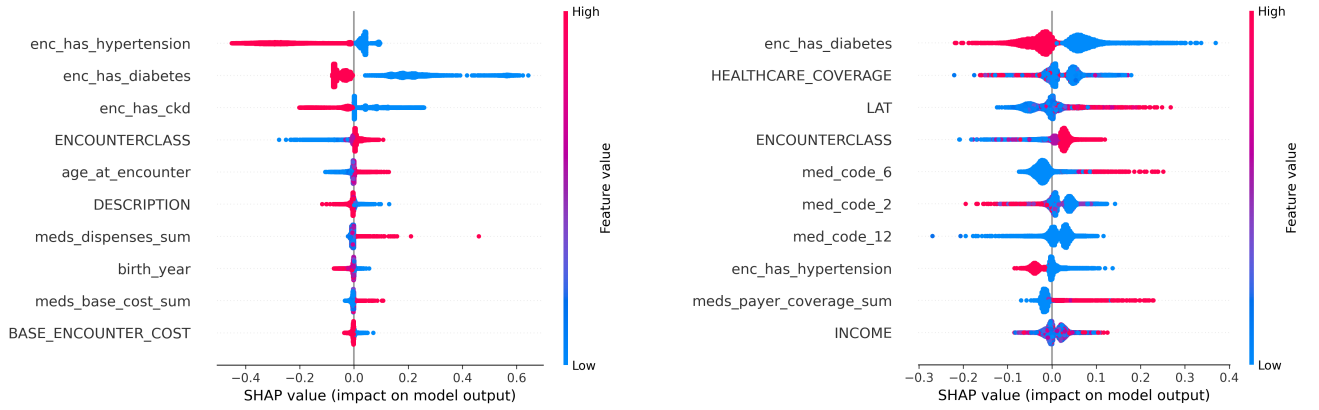
Figure 4. Comparison between box and whisker plots of the SHAP ranges for classifying Diabetes via MLP across different grouping methods. The term 'Feature' in this figure refers to a variable in the ML context.

(e.g., `enc_has_hypertension`, `enc_has_diabetes`, and `enc_has_ckd`) remained unchanged across different models for ungrouped cases, while they were not fixed for grouped cases. This implies that grouping adjusted the use of variables appropriately for a given model as opposed to the ungrouped case with more emphasis on the inherent structure of the dataset that outweighed the choice of model when making predictions.

Discussion

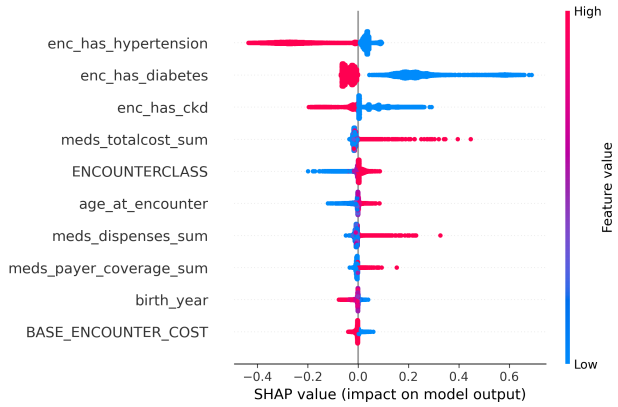
Principal findings and the significance of the study

This study investigated whether grouping variables by conditional dissimilarity in synthetic healthcare datasets and mapping grouped representations back to the original space via a decoder could make ML models both more explainable and equitable to support clinical decision. SHIELD integrated three core components: a dissimilarity-driven grouping stage, a decoder that localises latent effects to original

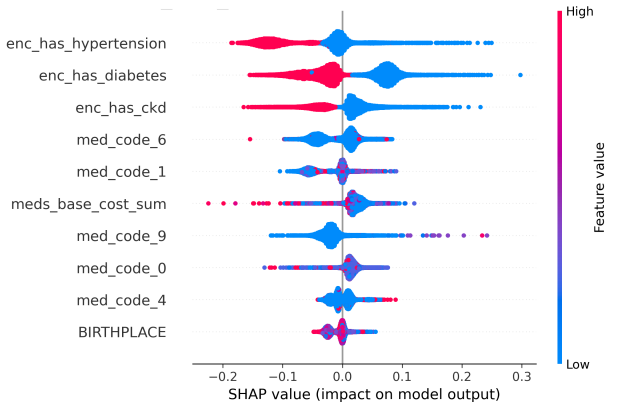


(a) LR (ungrouped)

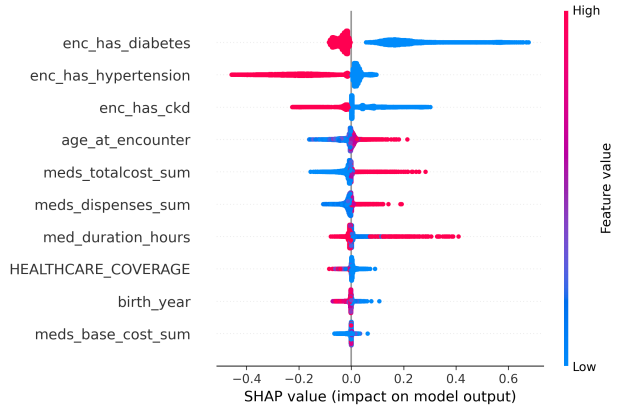
(b) LR (grouped)



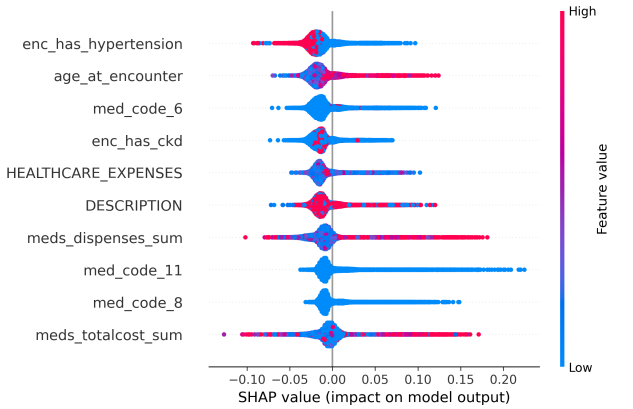
(c) MLP (ungrouped)



(d) MLP (grouped)



(e) XGB (ungrouped)



(f) XGB (grouped)

Figure 5. Comparison between SHAP beeswarm plots of Alzheimer's classification with bicriterion grouping across different models. The term 'Feature' in this figure refers to a variable in the ML context.

variables, and an audit suite that couples attributional analyses with group-parity metrics. SHIELD was evaluated across various clinical classification tasks via multiple classifiers, including linear, neural, and tree ensembles, to test for model-agnostic effects. The use of synthetic data enabled comprehensive fairness and explainability evaluation under reproducible, bias-controlled conditions without risking patient privacy.

Empirically, the approach consistently reduced the concentration of importance in a few dominant variables. Global SHAP summaries showed flatter spectra under grouped representations relative to ungrouped baselines, indicating broader participation of variables in the decision process. These findings corroborated the hypothesis that the grouping can mitigate bias patterns inherited from synthetic data generation processes, where latent dependencies between demographic and clinical variables are often preserved from source distributions. Fairness analyses complemented the explainability findings. Using six standard metrics, including equal opportunity, equalised odds, predictive parity, N-Sigma error rate, average distance from origin in the bias quadrant, and Bhattacharyya distance based separability, ungrouped cases often exhibited substantial disparities, which were improved to a different degree across grouping methods.

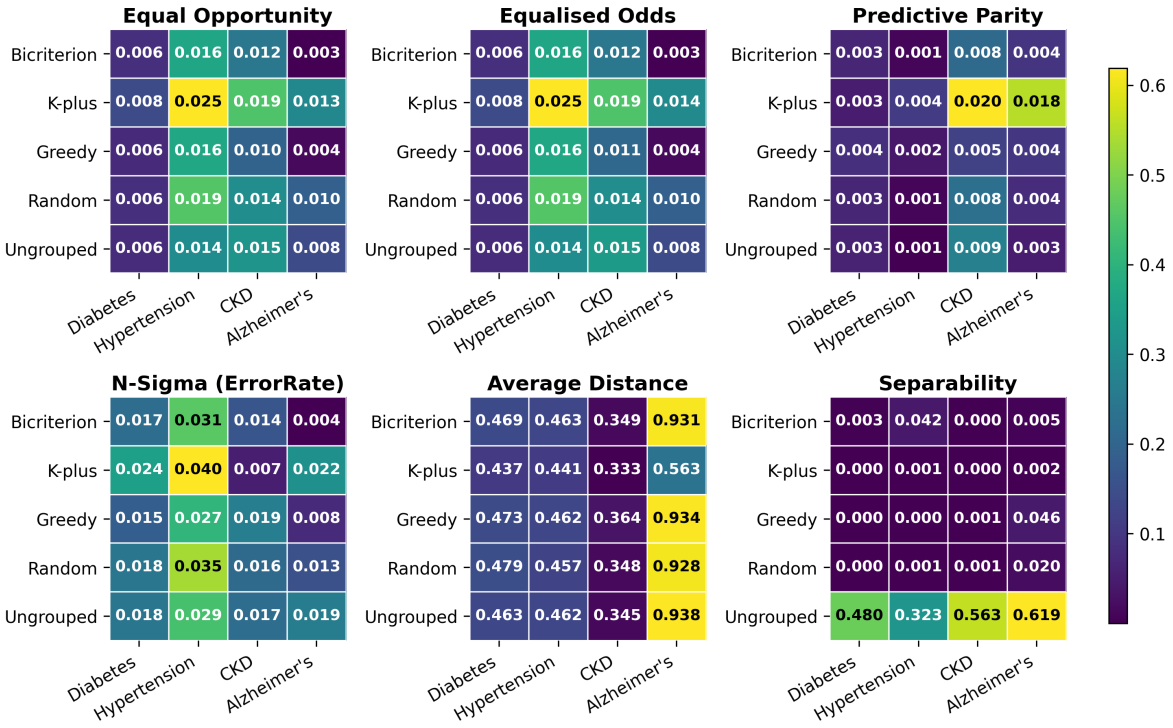


Figure 6. Equal opportunity, equalised odds, predictive parity, N-Sigma error rate, the average distance and separability in the bias quadrant across datasets and grouping methods. Note that lower is better for all fairness metrics above.

In particular, the bias-quadrant view provided an integrated perspective by plotting prediction parity against explanation parity on protected attributes. Grouping reliably contracted points toward the origin along the attribution axis, signalling reductions in explanation disparity, even when prediction disparities persisted across the observed range. This underscored a key insight: improving how decisions are justified does not automatically equalise outcomes, so methods need to be assessed on both axes. In this view, K-plus most consistently contracted toward the low-bias region, though this geometric stability did not always translate to the best scores across all fairness metrics. In this respect, Bicriterion was the most reliable method, delivering consistent fairness gains across datasets and metrics. Therefore, Bicriterion should be the preferred grouping strategy in fairness-critical applications, while K-plus might serve as a complementary method in contexts where geometric balance is prioritised.

Different grouping methods exhibited different trends. Bicriterion, which balances diversity and dispersion during grouping, was the most reliable across tasks. K-plus often excelled at reducing geometric bias magnitude but was less consistent on parity metrics. Random grouping occasionally matched stronger methods in certain configurations but lacked robustness. Greedy strategies were computationally heavier without commensurate gains. These patterns, together with the attributional evidence above, suggested that explicitly optimising both diversity and dispersion is a pragmatic choice if equitable learning needs to be promoted.

From a translational perspective, the compatibility of SHIELD with respect to synthetic data offers a governance-compliant environment for evaluating fairness interventions before real-world deployment. In practical terms, the results suggested that dissimilarity-based grouping with decoder mapping could deliver benefits beyond metric gains. By drawing signal from a broader share of variables and instances, grouped models are more sample-efficient and can reduce participant burden in prospective studies, since acceptable behaviour of models may be achievable without continuously enlarging the cohort. A more equitable reliance on variables also supports parsimonious testing designs and data collection protocols, saving acquisition and processing costs while keeping explanations traceable to native clinical variables through decoder-mapped SHAP.

Limitations and future work

Nevertheless, the study's scope and design placed boundaries on external validity. All experiments used structured tabular data with supervised classification endpoints and cross-validation. However, the study did not evaluate time-to-event outcomes, free-text modalities, nor conduct prospective clinical validation. Hyperparameter search spaces were finite, and the number of groups was not systematically optimised. Finally, the proposed claims were empirical and aligned with the domain knowledge but were not yet corroborated by formal guarantees on risk or attributional faithfulness after decoding.

Future work should extend SHIELD beyond variable-level explainability to incorporate instance-level attribution. While this study focused on how conditional

dissimilarity reshapes variable importance and fairness parity between sensitive attributes, the framework treats all training instances uniformly. Recent approaches, such as Data Shapley[50] and Residual Shapley decomposition (RSHAP)[51] offer principled ways to quantify the influence of individual training points on predictive behaviour. Integrating SHAP methods at instance level with SHIELD's decoder-based attribution could further reveal whether unfair decision patterns originate from specific influential patient records rather than general variable structure as a whole. This combined analysis would strengthen SHIELD's capacity to diagnose bias source within synthetic datasets and support more granular, clinician-aligned auditing in the bias quadrant space.

Future work should also consider applying SHIELD to semi-synthetic datasets, where records are incrementally anchored to real-world distributions[52]. This middle ground between real and synthetic datasets would allow fairness interventions to be stress-tested under controlled yet clinically realistic conditions, enhancing SHIELD's translational relevance for digital health governance and model auditing workflows.

Comparison with prior work

Several preprocessing approaches have been proposed to enhance fairness in ML by modifying the training data prior to model learning. Representative examples include reweighing, which adjusts the importance of training samples across protected groups[18], and Disparate Impact Remover (DIR), which transforms feature distributions to reduce statistical dependence between predictors and protected attributes[53]. These methods primarily target prediction-level fairness and do not explicitly evaluate whether explanatory importance is equitably distributed across variables.

Figure 7 compares these prior preprocessing techniques with SHIELD grouping strategies across predictive performance, outcome-level fairness, and explanation-level fairness metrics. Performance metrics (Accuracy and F1-score) remain largely comparable across methods, indicating that fairness interventions do not substantially degrade predictive capability for this task. However, differences emerge when examining fairness metrics. Traditional preprocessing methods such as DIR and Reweighing reduce disparities in outcome-level metrics, primarily the equal opportunity, but they exhibit substantially larger values in the explanation-level metrics, particularly the separability. This indicates that even when prediction outcomes appear relatively balanced across groups, the internal attribution structure of the model may still diverge substantially.

Additionally, the results of this study are consistent with prior evaluations of SHIELD on real-world clinical datasets[54]. In both settings, dissimilarity-based grouping reduced the concentration of SHAP importance in a small subset of variables and produced a more balanced attribution distribution across the feature space. This suggests that the fairness-oriented regularisation effect of dissimilarity grouping is not specific to a particular dataset type. However, the synthetic experiments in this study reveal additional distinctions among grouping strategies: bicriterion grouping consistently achieved balanced improvements across fairness metrics, whereas K-plus more strongly reduced geometric bias magnitude but was less consistent in outcome-level parity.

These results highlight an important distinction between SHIELD and prior preprocessing methods. While techniques such as Reweighing and DIR operate by adjusting sample distributions or feature representations to reduce statistical dependence on protected attributes, SHIELD instead modifies the structural organisation of variables through dissimilarity-based grouping. This mechanism targets proxy bias arising from correlations among predictors and encourages a more distributed pattern of explanatory importance across the feature space. Because SHIELD retains decoder mappings from latent representations back to the original variables, it also preserves interpretability, enabling fairness evaluation at the level of clinically meaningful variables.

Broader impact statement

The findings of this study highlight how synthetic healthcare data can serve as a powerful environment for developing and auditing equitable ML systems before they are introduced into real clinical workflows. By uncovering how attributional imbalance at the variable-level can emerge even in privacy-reserving synthetic datasets, SHIELD illustrates the importance of evaluating not only predictive performance but also the fairness and explainability of the decision-making process. This contributes to the wider effort of building digital health systems that are trustworthy, explainable, and safe for diverse patient populations.

More broadly, grouping by dissimilarity provides a lightweight intervention that is model-agnostic, data-efficient, and immediately applicable within existing health informatics pipelines. It supports synthetic data's role as a safe tool for early fairness evaluation by offering an interpretable mechanism to weaken spurious dependencies before real-world deployment at variable-level. These properties align with the broader goals of digital health research, which are to promote equitable, reproducible, and trustworthy AI systems that can be scrutinised by clinicians, regulators, and patients. This work contributes to a scalable pathway towards responsible adoption of AI in healthcare by demonstrating the potential to improve fairness and explainability without sacrificing practicality and predictive performance.

Conclusion

In conclusion, the results showed that grouping dissimilar variables via CMI, auditing both outcomes and explanations, and preserving traceability back to clinical variables form a coherent path toward transparent and equitable decision support. While there remains work to do on theory, instance-level attributions, and broader task coverage, the

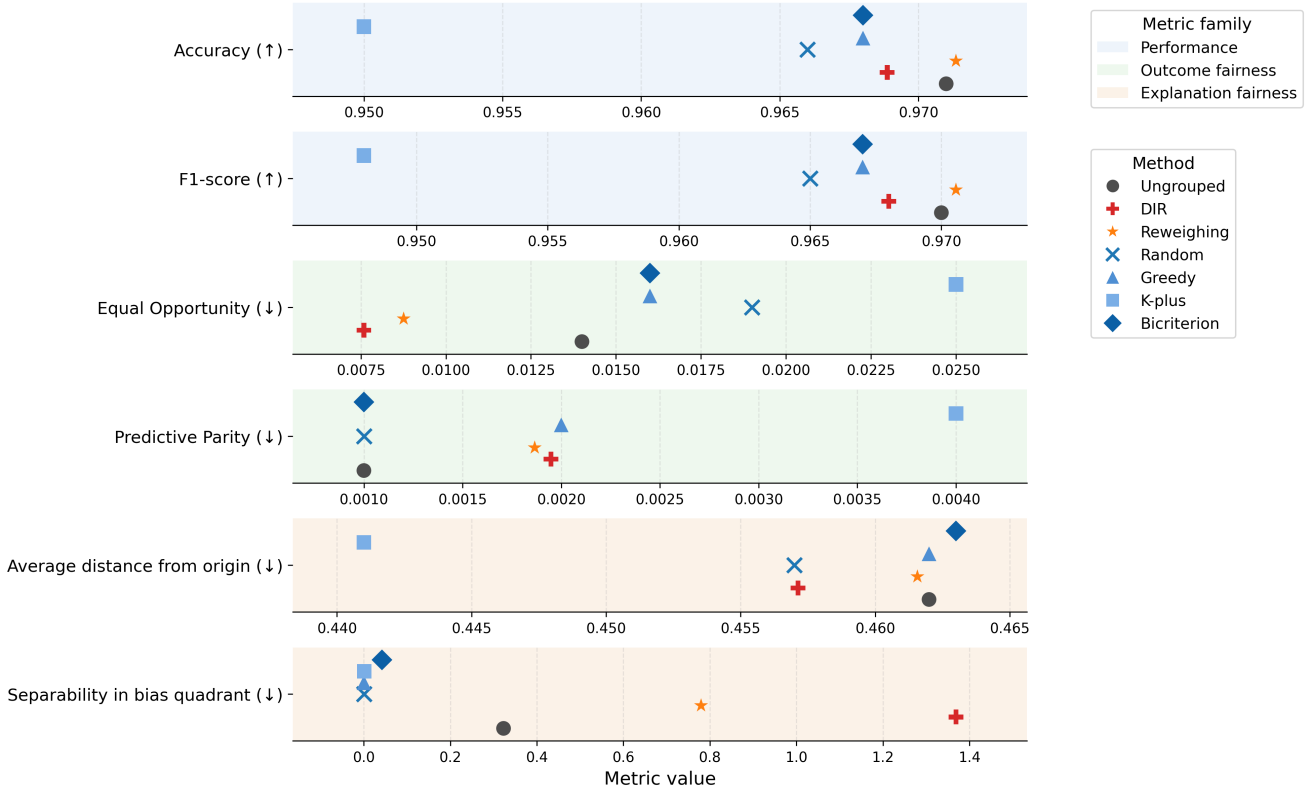


Figure 7. Comparison of performance, outcome fairness, and explanation fairness across SHIELD and prior methods on Hypertension classification.

contributions of this study provide a concrete step from concept to practice. By leveraging synthetic healthcare data as a validation substrate, SHIELD demonstrates how equitable and transparent AI can be developed responsibly within privacy-preserving ecosystems.

Statements and declarations

Ethical considerations

All data used in this study were existing, synthetic and available to public by CSIRO with Creative Commons Attribution 4.0 International Licence[25].

Consent to participate

Not applicable.

Consent for publication

All co-authors gave their consent for publication. No other consent was applicable.

Declaration of conflicting interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding statement

This study was supported by the Medical Research Future Fund, Australia under award number GA187319 (HS and AB).

Data availability

The dataset can be accessed from the CSIRO Data Access Portal[25]. All the code and artefacts can be found on this GitHub repository[55].

References

1. Rashidi HH, Pantanowitz J, Hanna MG et al. Introduction to Artificial Intelligence and Machine Learning in Pathology and Medicine: Generative and Nongenerative Artificial Intelligence Basics. *Modern Pathology* 2025; 38(4): 100688. DOI:https://doi.org/10.1016/j.modpat.2024.100688.
2. Obermeyer Z, Powers B, Vogeli C et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; 366(6464): 447–453. DOI: https://doi.org/10.1126/science.aax2342.
3. Barocas S and Selbst AD. Big Data's Disparate Impact. *California Law Review* 2016; 104(3): 671–732. DOI: https://dx.doi.org/10.15779/Z38BG31.
4. Rajkomar A, Dean J and Kohane I. Machine Learning in Medicine. *The New England Journal of Medicine* 2019; 380(14): 1347–1358. DOI:https://doi.org/10.1056/NEJMra1814259.
5. Huang W, Suominen H, Liu T et al. Explainable discovery of disease biomarkers: The case of ovarian cancer to illustrate the best practice in machine learning and shapley analysis. *Journal of Biomedical Informatics* 2023; 141: 104365. DOI:https://doi.org/10.1016/j.jbi.2023.104365.

6. Sadeghi Z, Alizadehsani R, CIFCI MA et al. A review of Explainable Artificial Intelligence in healthcare. *Computers and Electrical Engineering* 2024; 118: 109370. DOI:https://doi.org/10.1016/j.compeleceng.2024.109370.
7. Huang W, Suominen H, Liu T et al. A three-step machine learning strategy for reproducible biomarker screening in ovarian cancer. *Machine Learning Health* 2025; 1: 015011. DOI:https://doi.org/10.1088/3049-477X/ae06aa.
8. Bauer JM and Michalowski M. Human-centered explainability evaluation in clinical decision-making: a critical review of the literature. *Journal of the American Medical Informatics Association* 2025; 32(9): 1477–1484. DOI:https://doi.org/10.1093/jamia/ocaf110.
9. Wiens J, Saria S, Sendak M et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine* 2019; 25(9): 1337–1340. DOI:https://doi.org/10.1038/s41591-019-0548-6.
10. Pezoulas VC, Zaridis DI, Mylona E et al. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and Structural Biotechnology Journal* 2024; 23: 2892–2910. DOI:https://doi.org/10.1016/j.csbj.2024.07.005.
11. Giuffrè M and Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *npj Digital Medicine* 2023; 6. DOI:https://doi.org/10.1038/s41746-023-00927-3.
12. Walonoski J, Hall D, Gregorowicz A et al. Synthetic patient population simulator, 2025. URL https://github.com/synthetichealth/synthea.
13. Bhanot K, Qi M, Erickson JS et al. The Problem of Fairness in Synthetic Healthcare Data. *Entropy* 2021; 23(9): 1165. DOI:https://doi.org/10.3390/e23091165.
14. Friedler SA, Scheidegger C, Venkatasubramanian S et al. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, pp. 329–338. DOI:https://doi.org/10.1145/3287560.3287589.
15. Goncalves A, Ray P, Soper B et al. Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology* 2020; 20(108). DOI:https://doi.org/10.1186/s12874-020-00977-1.
16. Markus AF, Kors JA and Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics* 2021; 113: 103655. DOI:https://doi.org/10.1016/j.jbi.2020.103655.
17. Zhang BH, Lemoine B and Mitchell M. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. pp. 335–340. DOI:https://doi.org/10.1145/3278721.3278779.
18. Kamiran F and Calders T. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 2012; 33(1): 1–33. DOI:https://doi.org/10.1007/s10115-011-0463-8.
19. Lundberg SM and Lee SI. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17, Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964, p. 4768–4777. DOI:https://doi.org/10.48550/arXiv.1705.07874.
20. Hosmer DW, Lemeshow S and Sturdivant RX. *Applied Logistic Regression*. John Wiley & Sons, 2013. ISBN 9781118548387. DOI:https://doi.org/10.1002/9781118548387.
21. Rumelhart DE, Hinton GE and Williams RJ. Learning representations by back-propagating errors. *Nature* 1986; 323(6088): 533–536. DOI:https://doi.org/10.1038/323533a0.
22. Chen T and Guestrin C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 785–794. DOI:https://doi.org/10.1145/2939672.2939785.
23. Watson DS. On the Philosophy of Unsupervised Learning. *Philosophy & Technology* 2023; 36(2). DOI:https://doi.org/10.1007/s13347-023-00635-6.
24. El Mrabet MA, El Makkaoui K and Faize A. Supervised Machine Learning: A Survey. In *2021 4th International Conference on Advanced Communication Technologies and Networking (CommNet)*. pp. 1–10. DOI:https://doi.org/10.1109/CommNet52204.2021.9641998.
25. Diouf I, O'Brien M, Hassanzadeh H et al. Australian synthetic healthcare data with synthea, 2024. DOI:https://doi.org/10.25919/efcw-bm49.
26. Potdar K, Pardawala T and Pai C. A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. *International Journal of Computer Applications* 2017; 175(4): 7–9. DOI:https://doi.org/10.5120/ijca2017915495.
27. Patro SGK and Sahu KK. Normalization: A Preprocessing Stage. *International Advanced Research Journal in Science, Engineering and Technology* 2015; 2(3). DOI:https://doi.org/10.17148/IARJSET.2015.2305.
28. Zemel R, Wu Y, Swersky K et al. Learning Fair Representations. In Dasgupta S and McAllester D (eds.) *Proceedings of the 30th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, volume 28. Atlanta, Georgia, USA: PMLR, pp. 325–333. URL https://proceedings.mlr.press/v28/zemel13.html.
29. Madras D, Creager E, Pitassi T et al. Learning Adversarially Fair and Transferable Representations. In Dy J and Krause A (eds.) *Proceedings of the 35th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, volume 80. PMLR, pp. 3384–3393. URL https://proceedings.mlr.press/v80/madras18a.html.
30. Kearns M and Roth A. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, 2019. ISBN 978-0-19-094820-7.
31. Papenberg M. K-Plus anticlustering: An improved k-means criterion for maximizing between-group similarity. *British Journal of Mathematical and Statistical Psychology* 2024; 77(1): 80–102. DOI:https://doi.org/10.1111/bmsp.12315.

32. Datta A, Fredrikson M, Ko G et al. Proxy Non-Discrimination in Data-Driven Systems, 2017. DOI: https://doi.org/10.48550/arXiv.1707.08120.
33. Brown G, Pocock A, Zhao MJ et al. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The Journal of Machine Learning Research* 2012; 13(1): 27–66. DOI: https://dl.acm.org/doi/10.5555/2503308.2188387.
34. Vergara JR and Estévez PA. A review of feature selection methods based on mutual information. *Neural Computing and Applications* 2014; 24: 175–186. DOI: https://doi.org/10.1007/s00521-013-1368-0.
35. Cover TM and Thomas JA. *Elements of Information Theory*. John Wiley & Sons, 2005. ISBN 9780471748823. DOI:https://doi.org/10.1002/047174882X.
36. Brusco MJ, Cradit DJ and Steinley D. Combining diversity and dispersion criteria for anticlustering: A bicriterion approach. *British Journal of Mathematical and Statistical Psychology* 2020; 73(3): 375–396. DOI: https://doi.org/10.1111/bmsp.12186.
37. Wold S, Esbensen K and Geladi P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* 1987; 2(1): 37–52. DOI:https://doi.org/10.1016/0169-7439(87)80084-9.
38. Kramer M. Autoassociative neural networks. *Computers & Chemical Engineering* 1992; 16(4): 313–328. DOI: https://doi.org/10.1016/0098-1354(92)80051-A.
39. Snoek J, Larochelle H and Adams RP. Practical Bayesian optimization of machine learning algorithms. In *NIPS'12: Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 2. pp. 2951–2959. DOI:https://dl.acm.org/doi/10.5555/2999325.2999464.
40. Browne MW. Cross-Validation Methods. *Journal of Mathematical Psychology* 2000; 44(1): 108–132. DOI: https://doi.org/10.1006/jmps.1999.1279.
41. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 1997; 30(7): 1145–1159. DOI:https://doi.org/10.1016/S0031-3203(96)00142-2.
42. Bergstra J, Yamins D and Cox D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In Dasgupta S and McAllester D (eds.) *Proceedings of the 30th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, volume 28. Atlanta, Georgia, USA: PMLR, pp. 115–123. URL https://proceedings.mlr.press/v28/bergstra13.html.
43. Ghanem M, Ghaith AK, El-Hajj VG et al. Limitations in Evaluating Machine Learning Models for Imbalanced Binary Outcome Classification in Spine Surgery: A Systematic Review. *Brain Sciences* 2023; 13(12). DOI: https://doi.org/10.3390/brainsci13121723.
44. Aas K, Jullum M and Løland A. Explaining individual predictions when features are dependent: More accurate approximateions to Shapley values. *Artificial Intelligence* 2021; 298. DOI:https://doi.org/10.1016/j.artint.2021.103502.
45. Hardt M, Price E and Srebro N. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16, Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510838819, p. 3323–3331. DOI:https://dl.acm.org/doi/10.5555/3157382.3157469.
46. Barr CJS, Erdelyi O, Docherty PD et al. A Review of Fairness and A Practical Guide to Selecting Context-Appropriate Fairness Metrics in Machine Learning, 2026. DOI:https://doi.org/10.48550/arXiv.2411.06624.
47. DeAlcala D, Serna I, Morales A et al. Measuring Bias in AI Models: An Statistical Approach Introducing N-Sigma. *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)* 2023; : 1167–1172DOI:https://doi.org/10.1109/COMPSAC57700.2023.00176.
48. Jain A, Ravula M and Ghosh J. Biased Models Have Biased Explanations, 2020. DOI:https://doi.org/10.48550/arXiv.2012.10986.
49. Kailath T. The Divergence and Bhattacharyya Distance Measures in Signal Selection. *IEEE Transactions on Communication Technology* 1967; 15(1): 52–60. DOI: https://doi.org/10.1109/TCOM.1967.1089532.
50. Ghorbani A and Zou J. Data Shapley: Equitable Valuation of Data for Machine Learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, *Proceedings of Machine Learning Research*, volume 97. PMLR, pp. 2242–2251. URL https://proceedings.mlr.press/v97/ghorbani19c.html.
51. Liu T and Barnard AS. Shapley Based Residual Decomposition for Instance Analysis. In *Proceedings of the 40th International Conference on Machine Learning*. ICML'23, PMLR, pp. 21375–21387. URL https://proceedings.mlr.press/v202/liu23b.html.
52. Lyu Y, Dai S, Wu P et al. A semi-synthetic dataset generation framework for causal inference in recommender systems, 2022. DOI:https://doi.org/10.48550/arXiv.2202.11351.
53. Feldman M, Friedler SA, Moeller J et al. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15, New York, NY, USA: Association for Computing Machinery. ISBN 9781450336642, p. 259–268. DOI:https://doi.org/10.1145/2783258.2783311.
54. Yun H, Suominen H and Barnard AS. *SHIELD: A SHapley and Information-theory based framework for Equitable Learning via Dissimilar feature grouping*. Honours thesis, Australian National University, 2025.
55. Yun G. Shield, 2025. URL https://github.com/geun-yun/SHIELD.