The Australian National University
2600 ACT | Canberra | Australia

Australian National University

School of Computing

College of Systems and Society

SHIELD: A SHapley and Information-theory based framework for Equitable Learning via Dissimilar feature grouping

— Honours thesis (S1/S2 2025)

A thesis submitted for the degree
Bachelor of Advanced Computing (Research and Development) (Honours)

By:
Hyeonggeun Yun

Supervisors:
Prof. Hanna Suominen
Prof. Amanda Barnard

October 2025

Declaration:

I declare that this work:

- upholds the principles of academic integrity, as defined in the Academic Integrity Rule;
- is original, except where collaboration (for example group work) has been authorised in writing by the course convener in the class summary and/or LMS course site;
- is produced for the purposes of this assessment task and has not been submitted for assessment in any other context, except where authorised in writing by the course convener;
- gives appropriate acknowledgement of the ideas, scholarship and intellectual property of others insofar as these have been used;
- in no part involves copying, cheating, collusion, fabrication, plagiarism or recycling.

I acknowledge that I am expected to have undertaken Academic Integrity training through the Epigeum Academic Integrity modules prior to submitting an assessment, and so acknowledge that ignorance of the rules around academic integrity cannot be an excuse for any breach.

October (2025), Hyeonggeun Yun

Abstract

Machine learning models are increasingly applied in clinical and biomedical settings, due to their ability to capture an intricate underlying structure, yet their complexity can obscure critical insights and risk propagating biases that disproportionately affect vulnerable populations. Hence, this thesis introduces SHIELD: A SHapley and Information-theory based framework for Equitable Learning via Dissimilar feature grouping, which combines dissimilarity-driven feature grouping with interpretable latent representations to mitigate proxy bias and enhance equitable learning of the resulting model. By constructing a dissimilarity matrix based on conditional mutual information (CMI), features are grouped to weaken correlations that might encode sensitive attributes, reducing redundant signals that may contribute to unfairness. This automated approach is more efficient than clustering similar features and fixing problematic groups post-hoc. Group-specific autoencoders learn latent representations that summarise each group's unique information while preserving a decoder-weight mapping back to the original features. This enables precise SHapley Additive exPlanations (SHAP) value decomposition, leading to interpretable feature-level attributions despite dimensionality reduction. Experiments on four benchmark clinical datasets demonstrated that the proposed grouping approaches, greedy, Bicriterion, and K-plus anticlustering, achieved notable improvements in fairness metrics and produced more evenly distributed feature importance compared to raw features and traditional baselines. This was evident as grouping on average led to 9.47% improvement in the distance from origin of bias quadrant, which accounts for both explanation and prediction bias. In addition, the fairness overview score, which considers other typical fairness metrics as well, was improved by 2.42% when grouped by dissimilarity. While a modest reduction of 3.43% in accuracy and 5.16% in F1-score was observed, it remained within acceptable limits for clinical applications, demonstrating the feasibility of this fairness-performance trade-off. Overall, SHIELD provides a principled framework that integrates dissimilarity-based grouping, latent representation learning, and explanation-level auditing to promote equitable and explainable machine learning for health informatics.

Keywords: Machine learning, Feature grouping, Conditional mutual information, SHAP, Equitable learning, Health informatics, Explainable AI, Evaluation study

Acknowledgements

I would like to begin this thesis by acknowledging and honouring the First Scientists and Researchers of this country. I have lived, worked, studied and written this thesis on the beautiful lands of the Ngunnawal and Ngambri peoples over the past four years. I pay my deepest respects to their elders past and present, and acknowledge their ongoing custodianship of country, recognising that sovereignty was never ceded.

This Honours thesis marks one of the most memorable journeys of my academic life so far. It has taught me not only about research, machine learning, and explainable and equitable learning, but also about perseverance, curiosity, and gratitude. None of this would have been possible without the support and guidance of the people around me.

First and foremost, I would like to sincerely thank my supervisors, Professor Hanna Suominen and Professor Amanda Barnard. Your wisdom, guidance, and unwavering belief in me have shaped both thesis and the way I approach research. Your mentorship and kindness have been a constant source of inspiration, and I am truly grateful for the opportunities and trust you have given me. Your insightful feedback, encouragement, and resilience have motivated me to think critically and pursue excellence.

I would also like to acknowledge the Australian National University for providing such an intellectually stimulating and supportive environment throughout my four years of study, where I was able to explore my passions freely.

To my partner Nayoon, thank you for your endless support, understanding, and love throughout this journey. Your encouragement during late nights and stressful weeks kept me grounded and motivated.

I am also thankful to my friends Alex and Michael for sharing countless discussions, laughs, and moments of support that made the long research days more bearable.

Thank you to my family and friends for your unconditional love, sacrifices, and belief in me from the very beginning. Your support made it possible for me to pursue this journey with confidence and joy.

Finally, I would like to thank all of you, the readers of this thesis. I hope that you find this work both insightful and thought-provoking, and that it contributes in some way to your understanding of equitable and explainable machine learning. This research was written with the belief that transparency and fairness are shared responsibilities across our community. I sincerely thank you for taking the time to read my work.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

Artificial Intelligence (AI) and Machine Learning (ML) systems are increasingly deployed in healthcare to assist with diagnosis, triage, and treatment planning. These settings are high-stakes, since model outputs can directly affect patient outcomes and resource allocation. 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 have already been shown to exacerbate disparities. For example, some models were found to allocate fewer resources to patients with minority demographics or some risk prediction tools were systematically underestimating risk for disadvantaged groups [94, 10]. Recent editorials in Journal of the American Medical Informatics Association (JAMIA) further emphasised that explainability in clinical ML should be judged not only by interpretability and fidelity but also by clinical value [14, 118]. This means explanations must not only be understandable by human and faithful to the model, but also genuinely support safe and effective care. Even interpretable and faithful explanations can undermine clinical decisions, if the quality of the model systematically differs across patient groups in ways unrelated to the clinical condition. Hence, equity must be promoted along with transparency to make explanations consistently reliable and trustworthy.

Beyond these concerns, the nature of clinical tabular datasets further compounds the problem. Many outcomes of clinical interest are genuinely rare in the population, and minority subpopulations can be underrepresented due to access barriers, resulting in highly imbalanced class distributions [102, 56]. Data are also often incomplete because tests are only ordered when clinically indicated, and electronic health records are prone to missingness from documentation or system fragmentation. Moreover, structured records combine heterogeneous variables, such as lab tests, demographics, comorbidities, that differ in scale, prevalence, and reliability [45]. This combination of imbalance, sparsity, and heterogeneity makes clinical ML more error-prone, and disparities in data coverage or quality can directly lead to disparities in care. These challenges corroborate the significance of developing methods that not only maintain predictive accuracy but also yield equitable and trustworthy explanations, so that models remain reliable despite the non-trivial nature of datasets.

Methodologically, several challenges limit current practice of applying ML explainability and fairness methods in clinical setting. Firstly, many structured clinical datasets contain correlated or proxy features [82]. A correlated feature is one that shares statistical dependence with others, while a proxy feature indirectly encodes sensitive information (e.g. postcode as a proxy for socioeconomic status). A supervised training without fairness constraints often drives models to rely heavily on a few dominant predictors [94, 47]. This can mask the contribution of weaker but clinically meaningful features. When sensitive attributes or their proxies are correlated with these dominant predictors, the model may effectively learn to use the sensitive attribute indirectly. In such cases, differences in outcomes arise from non-clinical factors rather than genuine variation in disease risk. Moreover, post-hoc explanations attribute high importance to these dominant features, making inequity appear justified and thereby amplifying it [36, 60].

Secondly, explanation methods (hereafter, 'explainers'), such as Local Interpretable Model-Agnostic Explanations (LIME) [95] and SHapley Additive exPlanations (SHAP) [74], are widely used to attribute predictions to features. LIME perturbs inputs to learn a simple local surrogate, and SHAP quantifies each feature's contribution to a prediction using principles from cooperative game theory. However, these popular explainers typically assume approximate feature independence, which refers to the idea that predictor variables (columns) can be treated as if they were statistically independent. This is not to be confused with the Independent and Identically Distributed (i.i.d) assumption about independent instances across patients (rows). This assumption rarely holds in clinical data, where lab measures, comorbidities, and demographic factors can be interdependent. Violations can misallocate credit across correlated variables, producing inconsistent or misleading explanation attributions [126, 60]. In practice, this makes explanations difficult to trust, especially when they inform fairness audits [1, 126, 60].

Lastly, fairer outcomes should not be achieved at the cost of interpretability and explainability. While it is possible to intervene the training process to recover a fairer outcome in cases where models do not achieve parity across groups, many of the approaches modify the learning objective or incorporate adversarial components in latent representations that clinicians cannot understand [124, 63]. Yet, it is the variables like lab results, vital signs, and demographic characteristics that clinicians must see to validate and act on a model's behaviour. Hence, a modular equitable approach should preserves explanations in the native feature space to make it clinically usable.

In summary, these clinical and methodological concerns form a virtuous circle. Better methods can make explanations more trustworthy and equitable, which in turn builds confidence in applying the model to clinical settings, which then motivates further methodological refinements. Figure 1.1 illustrates how this thesis sits at the intersection of Information theory, feature grouping, SHAP-based explainability, and equitable learning to foster this circle. Building on this intersection, the thesis proposes **SHIELD: A SHapley and Information-theoretic framework for Equitable Learning via Dissimilar feature grouping**. The aim is to redistribute model reliance more equitably across features, and yield more reliable explanations under feature dependence,

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.

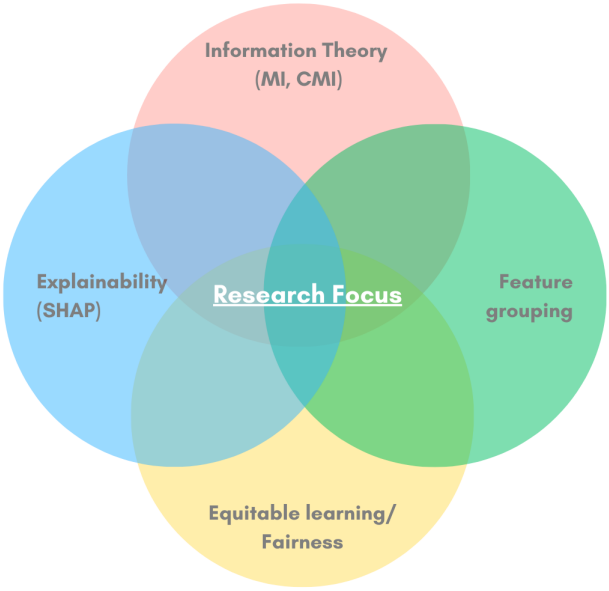


Figure 1.1: Venn diagram of research areas to identify the gap that is aimed to be filled by this thesis.

1.2 Research questions and expectations

The motivation above naturally led the author to ponder the following questions. First, does grouping redistribute SHAP attributions, which quantify how much each feature contributes to a model prediction, towards a more balanced profile? Does this redistribution lead to fairer outcomes across protected groups (i.e. cohorts defined by sensitive attributes such as sex, ethnicity, or age) [3, 56]? Second, how consistent are these effects across different tabular datasets, feature grouping methods, and model classes [20, 23, 29]? Third, does grouping remain effective under the challenging conditions of clinical data, such as limited sample sizes, missingness, heterogeneous variables, and imbalanced outcomes, where common metrics like accuracy can be misleading [102, 94]?

Finally, all the motivations and questions above can converge to a single research question that this thesis aims to address: "Can grouping mutually dissimilar features improve equitable learning in clinical prediction without unacceptably degrading predictive performance?" This question intertwines four distinct concepts, which are feature grouping, dissimilarity, equitable learning, and performance, where each must be further unpacked.

It should also be noted that the theoretical and methodological details of the descriptions below are captured in other chapters of the thesis, as summarised at the end of this section.

Feature grouping, in the context of this research, refers to partitioning the raw predictor set into non-overlapping subsets that serve as primitive units for subsequent representation learning. Unlike conventional cluster analysis, which aggregates highly correlated variables to exploit redundancy, the proposed anticlustering framework deliberately assembles variables that share little conditional mutual information [104]. The intent is to prevent proxy variables for protected attributes from reinforcing one another once they are projected into latent space. By distributing near-duplicates and/or socio-demographic surrogates across different partitions, the learnt embeddings are encouraged to equitably weight features which a model's decision boundary might otherwise be dominated by a few features.

Dissimilar features are defined through the complement of conditional mutual information given the outcome label. Two predictors are therefore deemed dissimilar if they convey largely independent information, after conditioning on the clinical endpoint or target label, measured in terms of entropy. This definition is more strict than mere use of marginal correlation. It isolates redundancy that is predictably relevant, ensuring that each partition captures a unique axis of clinical signal rather than residual noise. The underlying hypothesis is that such controlled heterogeneity acts as an implicit regularisation, suppressing spurious pathways that have historically channelled bias into automated decision systems [10, 36].

Equitable learning is the primary constraint this thesis aims to satisfy. Hence, it is operationalised in Section 2.2, where the author formalise outcomes and explanation parity and the metrics used throughout.

Performance refers to how well a model predicts in practice. This thesis uses standard classification metrics that clinicians and ML researchers use too, which are accuracy, precision, recall and F1 [93, 102, 94], with F1 being especially insightful for imbalanced datasets. In addition, robustness is assessed by repeating cross-validation with different random seeds and monitoring the spread of results. Throughout the study, any gains in fairness or explainability are acceptable only if they do not cause a clinically significant reduction in performance. In this trade-off, sensitivity of the model must be checked with extra caution in settings where false negatives (missed diagnosis) carry substantial clinical risk [56, 94].

The expectation from this investigation is that structured use of feature information can redistribute feature contribution away from a few dominant ones, yield more balanced SHAP profiles, and reduce disparities between protected groups, while preserving competitive predictive accuracy. A secondary expectation is that these benefits will depend on the combination of dataset characteristics, grouping strategy and learner type. They

are anticipated to be most visible when data are scarce or imbalanced, and broadly consistent across model classes, provided the explanation pipeline remains faithful to the original clinical variables.

1.3 Contributions

This thesis advances equitable and explainable machine learning by introducing a grouping framework that redistributes model reliance across features while preserving interpretability in the original feature space, and by coupling attribution analysis with fairness auditing. The main contributions are as follows.

First, a dissimilarity-based grouping pipeline is proposed in which input features are partitioned by conditional dissimilarity and encoded to a compact latent representation, then mapped back to the original feature space through decoder weights for explanation. This decoder-mapped step allows the computation of SHAP attributions on the latent units and rigorously decompose them into per-feature contributions, so explanations remain expressed in the clinical variables used by practitioners. The design is agnostic to the downstream learner and accommodates multiple grouping strategies, including bicriterion and K-plus anticlustering.

Second, a joint audit of outcomes and explanations is developed through the bias-quadrant visualisation and its summary metrics. Prediction bias is plotted on the vertical axis and explanation bias (differences in SHAP by protected attribute) on the horizontal axis, enabling a single-view diagnosis of "what" the model predicts and "why". This is complemented with the average distance from the origin as a bias-magnitude indicator and an aggregated Fairness Overview score that blends outcome and explanation parity, providing a compact, generalisable comparator.

Third, a cross-dataset, cross-model empirical study is conducted to isolate the effect of grouping. Five widely used learners, Logistic Regression, Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Random Forest and XGBoost (see Section 4.5.1), are evaluated on four healthcare datasets plus an imbalanced obesity variant. For each configuration, this thesis reports global and per-instance SHAP summaries, fairness metrics, and bias-quadrants, using a consistent background and fixed-index protocol so grouped and ungrouped analyses are directly comparable.

Fourth, instance-level explainability is extended with diagnostics that summarise how much of a single prediction is carried by the most influential features. This visualisation is paired with a stacked-bar summary of the total absolute attribution and its decomposition into the top nine contributors and the remainder, offering a concise measure of dominance versus dispersion that mirrors the global SHAP effects.

Fifth, evidence-based guidance is provided on when and how to group. Across tasks, grouping flattens attribution spectra and often improves fairness, with the strongest

gains in smaller datasets and under class imbalance. Bicriterion emerges as the most reliable method overall, K-plus is competitive when reducing bias magnitude, Random is variable, and Greedy rarely justifies its extra cost. These regularities give practitioners actionable defaults and caveats for fairness-critical applications.

Finally, the thesis assembles a modular, reproducible workflow that integrates grouping, decoding, SHAP computation, fairness assessment, and visual reporting. This end-to-end design allows future work to swap encoders, grouping objectives, or learners without altering the explanation or auditing stages. This sets the stage for the theoretical and instance-level extensions outlined in the Future works (Chapter 6).

The rest of this thesis is organised as follows. Chapter 2 provides the background, introducing the general machine learning pipeline, the distinction between equity and fairness, explainable AI methods with emphasis on SHAP, relevant Information theory principles, and domain knowledge for the datasets. Chapter 3 reviews related work across three intersections: SHAP with Information theory, grouping-based feature selection, and fairness metrics. Chapter 4 presents the methodology, including data collection, preprocessing, feature grouping by conditional dissimilarity, model training, and evaluation protocols. Chapter 5 reports results and discussion, covering preprocessing, grouping analysis, tuned hyperparameters, model performance, SHAP and fairness metrics, and clinical implications, followed by limitations. As mentioned, Chapter 6 outlines directions for future work, such as theoretical validation, instance-level extensions, and regression tasks. Finally, Chapter 7 concludes the thesis with a summary of contributions and implications for equitable and explainable clinical ML.

Chapter 2

Background

2.1 General machine learning pipeline

A ML study, particularly one applied in critical fields such as healthcare, must proceed through a disciplined sequence of stages. Each stage carries distinct assumptions, potential failure modes, and best-practice safeguards that, if neglected, can undermine the reliability of the entire process and lead to serious consequences. Broadly, an ML pipeline can be conceptualised in five core stages: data collection and preprocessing, feature engineering and selection, model training, model evaluation, and model interpretation. Methodological errors at any stage may propagate downstream, leading to biased, misleading, or clinically unsafe outcomes [41, 66, 117].

The quality of an ML model is intrinsically linked to the quality of its data. Preprocessing transforms raw data into a trainable and analysable form, which typically involves the handling of missing values, normalisation of continuous variables, encoding of categorical features, and mitigation of class imbalance [69]. In healthcare, where data are often heterogeneous and noisy, ranging from clinical notes to sensor measurements, robust preprocessing is particularly important. Failure to do so can introduce systematic biases or amplify artefacts, thereby compromising downstream predictive accuracy [79].

Feature engineering creates informative representations of raw variables, while feature selection identifies the most relevant predictors for the task at hand. These steps reduce dimensionality, mitigate overfitting, and enhance computational efficiency [53]. In medical datasets, which may include thousands of attributes spanning laboratory tests, imaging, and demographic information, careful feature selection is indispensable for isolating clinically meaningful predictors while excluding redundant or spurious variables [45]. Failures to perform appropriate feature engineering and selection can lead to consequences like that of the previous stage [79]. This reinforces the broader point that model quality depends not only on the quality of the raw data itself but also on the rigour with which it is curated, processed, and transformed into features.

Once features are curated, the model training stage involves fitting algorithms to optimise predictive objectives. The choice of model, whether it is linear, non-linear, neural-based or tree-based, depends on dataset size, problem complexity, and interpretability requirements [20]. Hyperparameter tuning, often conducted via grid search, random search, or Bayesian optimisation along with cross-validation, ensures models strike an

appropriate balance between bias and variance [17]. In clinical contexts, these choices must be justified not only statistically but also align with domain knowledge through the inspections of their explanations.

Evaluation determines whether a model generalises beyond its training data, which makes use of the testing set that should not have been touched until this stage. Performance is quantified through metrics such as accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC) [93]. In healthcare, however, metrics like sensitivity and specificity often take precedence, since the consequences of false negatives may be more severe than false positives [94]. Beyond these metrics, robust statistical analysis should also be performed, typically focusing on clinically meaningful effect sizes rather than solely reporting statistical significance [5, 59].

Examining interpretability and explainability is the final and perhaps the most critical stage when ML is applied in healthcare. While highly complex models may offer strong predictive performance, they are often opaque to end-users. Post-hoc, model-agnostic explanation techniques such as LIME and SHAP allow predictions to be decomposed into contributions from individual features [74, 95]. This stage promotes transparency, strengthens clinician trust, and can generate novel clinical insights by highlighting unexpected relationships within the data [27].

2.2 Equitable and Fair learning

Equity and fairness play complementary but distinct roles in clinical ML. Equity concerns the just distribution of benefit relative to need [96], while fairness tests for unjustified disparities in error or outcome across protected groups [84]. They are often spoken of together and treated as if "more is always better" for both [94, 10, 47]. In practice, they can pull in different directions, so maximising one may reduce the other. When risk or clinical need truly differs between protected groups, equity favours allocating more resources or sensitivity where need is higher. A strict fairness constraint that pursues identical error rates or predictive values across groups can then reduce benefit to those with greater need. Conversely, a policy that equalises downstream benefit by tailoring actions to risk can yield different true or false positive rates, failing a fairness test. A well-documented case is in the prediction of cardiovascular risk, where women and men can differ in the incidence of adverse outcomes at baseline. Adjusting thresholds by sex can improve equity of treatment allocation but can break fairness constraints that require identical performance metrics [44].

Consequently, this thesis distinguishes equity from fairness, then specialises both to the modelling choices and evaluation protocols proposed later. Equity refers to proportional allocation of benefit and burden. In clinical decision support, this means that patients who differ in clinically relevant need or risk may justifiably receive different actions so

that downstream outcomes are comparable. Fairness refers to the absence of unjustified discrimination. Groups that differ only in protected attributes, namely sex or ethnicity, should not receive systematically worse outcomes nor processes, unless they are indeed among the valid deciding factors.

In ML, fairness is commonly operationalised as parity of outcome or error rates across protected groups. This thesis adopts three standard criteria. Equality of Opportunity requires similar true positive rates across groups [54]. Equalised Odds requires similarity of both true and false positive rates [11]. Predictive Parity requires similar positive predictive values [47]. This study also reports an error-rate disparity derived from the N-sigma idea as a compact statistical summary [37], together with a two dimensional bias quadrant summary based on average distance from the origin [60]. These metrics are evaluated on held-out data and are reported alongside conventional predictive metrics throughout Chapter 5.

In SHIELD, equity is involved at the representation and explanation level. Beyond outcome parity, this work investigates whether a model distributes explanatory credit across available features and instances rather than concentrating reliance on a narrow, potentially proxy-laden subset. This aspect is known as equitable learning in explanations [11]. Nevertheless, it is important to acknowledge that a flatter distribution of explanatory credit (see Figure 2.1) is not always desirable, nor valid. In some clinical problems, a small set of biomarkers may legitimately carry most of the signal [82] (e.g. sex would be a genuine deciding factor of one's ability to be pregnant). In such cases, forcing attributions to spread evenly can dilute true signal and harm utility [80]. Therefore, this thesis treats explanation parity as conditional on two checks. Firstly, grouped models should preserve predictive performance within acceptable margins. Secondly, the induced feature rankings remain plausible in light of domain knowledge. Where these diverge, this thesis prioritises clinical validity and report the tension explicitly rather than claiming a fairness gain from dispersion alone [3, 56, 80].

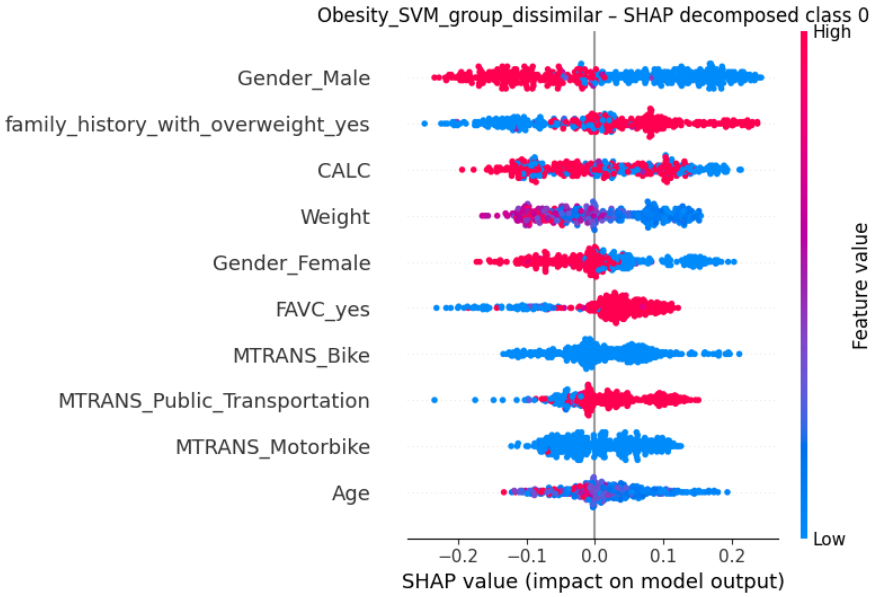


Figure 2.1: Example of SHAP beeswarm plot, where flatter distribution of explnatory credit is not necessarily desirable. The plot is for predicting Insufficient Weight via SVM and greedy approach of feature grouping. Note how `Gender_Male` contributes the most to the decision, even more than `Weight` (although only by a small margin), since the features are 'flat'. Also note the SHAP values with respect to feature value for `Gender_Male` and `Gender_Female` are not opposite, which is questionable or uninformative since it means not being male positively influences the prediction towards insufficient weight, but so does not being female.

2.3 Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) has emerged in response to a persistent trade-off in machine learning between model accuracy and model interpretability [12, 3]. As shown in Figure 2.2, highly intricate models such as deep learning often achieve state-of-the-art predictive performance but are opaque to human users, whereas simpler models such as logistic regression or decision trees are more interpretable but may sacrifice accuracy. This trade-off is particularly problematic in healthcare, where predictive accuracy alone is insufficient [56]. Clinical adoption requires that model behaviour can be scrutinised, justified, and aligned with medical reasoning [3, 94]. The goal of XAI is to bridge this gap by designing models that are inherently more interpretable or by applying post-hoc explanation techniques that clarify the predictions of otherwise black-box models. In this way, XAI provides the methodological foundation for balancing performance with transparency, accountability, and trust in high-stakes domains.

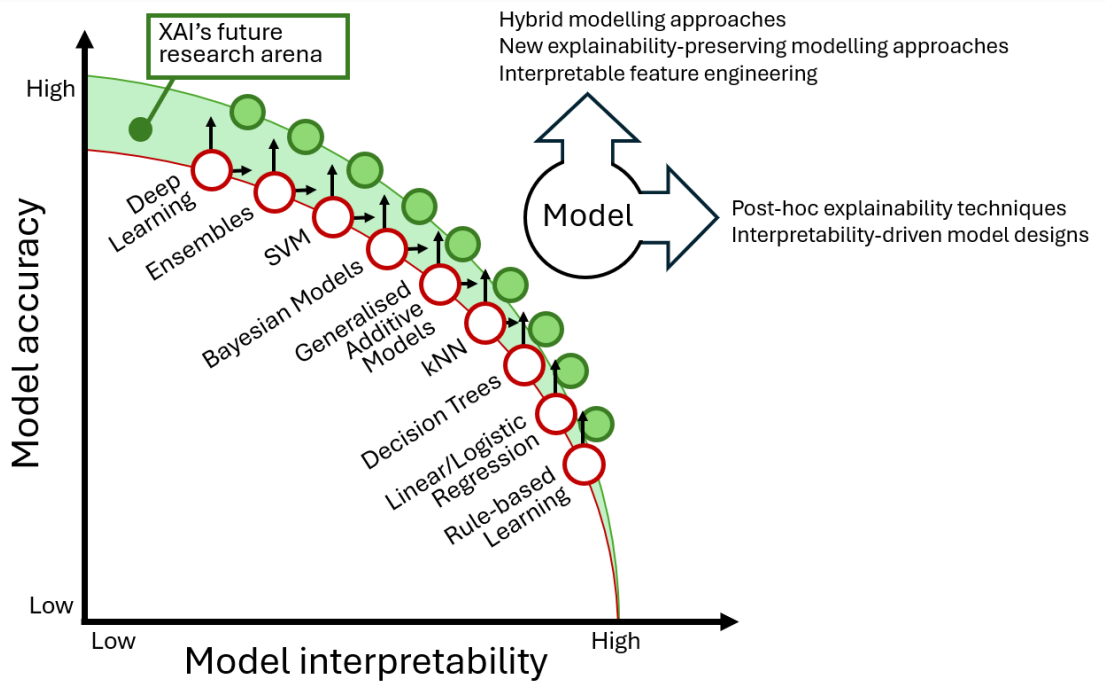


Figure 2.2: Illustration of trade-off between model accuracy and interpretability, adapted from Barredo et al. [12].

Throughout this thesis, interpretability refers to a model's inherent ability to be understood by humans, and explainability focuses on providing human-understandable reasons for a model's specific predictions or outputs, often through post-hoc techniques that break down complex processes [43]. This thesis focuses on improving explainability over constraining models to be intrinsically interpretable, since SHIELD is designed to be model-agnostic. One of the components in SHIELD that supports explainability is a decoder, which transforms latent representations of grouping to human-readable results. This focus allows for comparison across diverse model classes under a common explanatory lens.

Among the most widely used explainability tools are SHAP and LIME. SHAP provides an additive attribution satisfying local accuracy and consistency by connecting to Shapley values from cooperative game theory, while LIME fits a locally faithful surrogate around the instance of interest. SHAP is chosen over LIME in this study, because its attributions satisfy clear axioms (local accuracy, consistency, missingness) and are additive, so contributions sum to the prediction difference. These properties make it straightforward to aggregate instance-level attributions into global summaries, compare groups for fairness auditing, and map contributions across representation layers. In particular, the additivity allows attributions to be decoded from latent groups back to the original clinical features without changing their semantics. SHAP also has efficient implementations for common model families and supports both model-agnostic and model-specific

explainers, which helps to maintain methodological consistency across the experiments [74, 74]. By contrast, LIME relies on sampling and local surrogate fitting, which can be sensitive to kernel choice and perturbation design, leading to higher variance in explanations and less direct aggregation to corpus-level fairness analyses [95]. For these reasons, SHAP provides a better match to the goals of this study.

Nevertheless, not all explanation needs are satisfied by additive attributions. In decision support, counterfactual explanations specify minimal, actionable changes to flip an outcome (e.g. "to obtain a positive prediction, raise x_1 by ..."), which complements attribution by articulating feasible recourse [115]. In clinical settings, pairing attribution ("why") with counterfactual recourse ("how to change") and fairness auditing ("for whom") provides a multi-view account of model behaviour. Therefore, the following perspective is adopted throughout this thesis. SHAP is used for instance and corpus level attribution, fairness metrics for group-wise disparities, and decoded mappings to connect latent structure back to clinically meaningful variables.

2.4 SHAP

SHAP was introduced by Lundberg and Lee to unify a growing set of feature attribution methods under a single additive framework with clear theoretical guarantees [74]. Their motivation was practical and theoretical, since many widely used post hoc explainers produced incompatible scores and sometimes contradictory rankings, making it difficult to compare explanations across models or datasets. By linking attributions to Shapley values from cooperative game theory, SHAP selects a unique additive attribution that satisfies local accuracy, missingness, and consistency. This can also be computed or closely approximated for common model classes.

Formally, for a trained predictor $f : \mathbb{R}^p \to \mathbb{R}$ and an instance x, SHAP defines a baseline $\phi_0 = \mathbb{E}[f(X)]$ and feature contributions $\{\phi_j(f, x)\}_{j=1}^{p}$ that satisfy local additivity,

$$f(x) = \phi_0 + \sum_{j=1}^{p} \phi_j(f, x). \tag{2.1}$$

The value of a coalition S of present features is given by a conditional expectation $v_x(S) = \mathbb{E}[f(X) \mid X_S = x_S]$, and each ϕ_j averages the marginal contribution of feature j over all subsets $S \subseteq \{1, \ldots, p\} \setminus \{j\}$ using Shapley weights. This gives a common numerical scale to compare how grouped and ungrouped representations use information.

The choice of background distribution determines both the baseline ϕ_0 and how "missing" features are simulated when forming $v_x(S)$. Marginal maskers approximate X_{-S} by independent draws from their empirical distribution, which is computationally light but can be inaccurate when predictors are dependent. Conditional maskers preserve correlation by sampling from $X_{-S} \mid X_S = x_S$ at increased computational cost. In healthcare

data, where strong dependencies are common, this design choice affects attribution stability. To make fair comparisons between grouped and ungrouped models, the same background cohort and masking scheme are used across configurations.

In this thesis, a range of commonly used model families were selected to test whether grouping effects generalise across different learning paradigms. Random Forest [33] and XGBoost [29] represent ensemble tree methods that are widely applied for their robustness and ability to capture non-linear feature interactions. Logistic regression [57] serves as a classical linear baseline that is both interpretable and statistically familiar in clinical contexts. SVM [32] and MLP [20] capture non-linear decision boundaries through margin maximisation and neural representations, respectively.

Exact Shapley computation is exponential in p, so practical explainers provide tractable estimators that retain local additivity. Hence, appropriate SHAP variants were chosen to make explanations consistent and feasible across these algorithms. Random Forest and XGBoost are explained with TreeExplainer [73], which exploits the structure of decision trees to compute SHAP values in polynomial time. Logistic regression is explained with LinearExplainer [74] or, when embedded in a pipeline, with a model-agnostic variant. SVM and MLP are explained with KernelExplainer [74], which approximates SHAP values by fitting locally weighted surrogates, using the same background masker across grouped and ungrouped runs to ensure comparability.

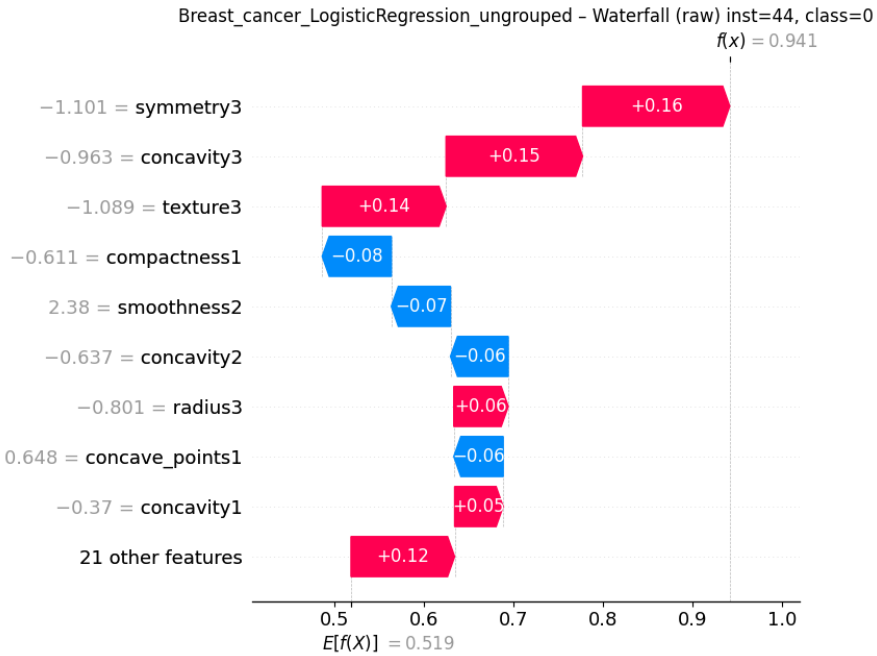


Figure 2.3: Example figure of SHAP waterfall plot, illustrating the case for prediction of Breast cancer via Logistic Regression without feature grouping.

Local explanations are visualised with waterfall plots (see 2.3) that show how the contributions move the prediction from the baseline to $f(x)$. Aggregating absolute contributions over a test set yields global importance profiles $\mathbb{E}_x[|\phi_j(f,x)|]$, which is displayed as bar or beeswarm plots (e.g. Figure 2.1). This thesis also tracks the prevalence of near-zero attributions across instances, since widespread zeros indicate under-used variables.

Correlated features present a particular challenge because attribution can become sensitive to the masking scheme and model structure. In tree ensembles, correlation can inflate or deflate importance in ways that are not purely causal [52, 129], and more broadly, Shapley implementations can exhibit observation or structural biases when background assumptions are misspecified [127, 126]. The methodology addresses this by grouping conditionally dissimilar features before learning, then mapping latent contributions back to the original variables with a decoder.

Explanations must also be judged for faithfulness to the trained model and for alignment with fairness goals. Biased models can produce biased explanations [60], so SHAP is interpreted alongside outcome-level fairness metrics and the bias-quadrant visualisations. Sensitivity to the background cohort and masker is made explicit, and attributions are compared under a constant background across grouped and ungrouped runs to isolate the effect of representation. These practices ensure that shifts observed in attribution distributions can be credibly linked to grouping rather than to the explainer itself.

It should now be clear SHAP supports two complementary roles within this thesis. At instance level, waterfalls diagnose whether a prediction relies on a narrow, correlated subset or on a broader base of features, which is appropriate for auditing individual patient's behaviour. At feature level, aggregated attributions quantify equity of reasoning by measuring how widely the model spreads importance across variables.

2.5 Information theory

Information theory offers a principled language for quantifying uncertainty and statistical dependence. Originating with Shannon's work on communication and coding [104], it provides tools that are directly relevant to ML on clinical tabular data. Throughout this thesis, the following core quantities are used to capture relevance, redundancy, and potential proxy relationships.

Let X be a discrete random variable with probability function $p(x)$. The entropy

$$H(X) \;=\; -\sum_{x} p(x)\,\log p(x) \tag{2.2}$$

measures the average uncertainty of X. For jointly distributed variables (X, Y), the joint and conditional entropies are

$$H(X,Y) \;=\; -\sum_{x,y} p(x,y)\,\log p(x,y), \qquad H(Y \mid X) \;=\; -\sum_{x,y} p(x,y)\,\log p(y \mid x). \quad (2.3)$$

Logarithms may be taken with base 2 for bits or base e. The choice only rescales values, and thus the study uses base 2.

2.5.1 Mutual information

Mutual information (MI) quantifies the total statistical dependence between two variables. It can be written as a Kullback-Leibler (KL) divergence between the joint and the product of marginals,

$$I(X;Y) \;=\; \sum_{x,y} p(x,y)\,\log \frac{p(x,y)}{p(x)\,p(y)}, \quad (2.4)$$

and equivalently as an expected reduction in entropy,

$$I(X;Y) \;=\; H(X) - H(X \mid Y) \;=\; H(Y) - H(Y \mid X). \quad (2.5)$$

MI is non-negative, symmetric, and equals zero if and only if X and Y are independent [34]. It obeys a chain rule,

$$I(X;Y,Z) \;=\; I(X;Y) + I(X;Z \mid Y), \quad (2.6)$$

and a data processing inequality, which states that if $X \to Z \to Y$ forms a Markov chain then $I(X;Y) \le I(X;Z)$. Two interpretations are especially useful for tabular healthcare features. First, $I(X_j;Y)$ serves as a relevance score for feature X_j with respect to the clinical label Y. Second, pairwise $I(X_j;X_k)$ reflects redundancy among features, including correlations arising from measurement practices or physiology.

These ideas motivate classic filter methods for feature selection that seek high relevance with low redundancy. This balance can be maintained by favouring sets whose members have large $I(X_j;Y)$ while keeping average $I(X_j;X_k)$ small [90, 114]. Many information-theoretic views capture these criteria as trade-offs among relevance, redundancy, and complementarity, expressed through MI decompositions and approximations [24]. Although estimation from finite samples requires care, these properties explain why MI is widely used to reason about which variables carry unique predictive signal and which largely duplicate others.

2.5.2 Conditional mutual information

Conditional mutual information (CMI) isolates the dependence between two variables that remains after conditioning on a third. It is defined by

$$I(X;Y \mid Z) = \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y \mid z)}{p(x \mid z)\, p(y \mid z)} = H(X \mid Z) - H(X \mid Y,Z). \quad (2.7)$$

CMI is non-negative and equals zero precisely when X and Y are conditionally independent given Z. The chain rule above shows that $I(X;Y \mid Z)$ measures the incremental information about X obtained from Y once Z is known. This viewpoint underpins conditional criteria for feature selection, for example selecting the next feature X_j that maximises $I(X_j;Y \mid S)$ where S is the set of already chosen features, thereby preferring features that add information beyond what S explains [24, 85, 114].

CMI also provides language for reasoning about proxies and confounding. Let A denote a protected attribute and Y be a true label. The quantity $I(X_j;A \mid Y)$ captures how much information about A leaks through feature X_j after accounting for clinical state. Small values suggest that any association between X_j and A is explained by Y rather than by spurious pathways, while large values indicate residual linkage that could support proxy discrimination [36, 123]. Because CMI conditions on context, it is well suited to distinguish clinically necessary dependence from unwanted encoding of sensitive information.

2.6 Domain knowledge for datasets

The datasets analysed in this thesis were selected not for a specific clinical agenda but for their methodological diversity and relevance to evaluating equitable and explainable ML in healthcare domain. Each dataset represents a distinct domain of healthcare but, more importantly, spans a range of data properties that make predictive modelling and explanation challenging: from small, curated cohorts (e.g. breast cancer) to large, administrative datasets with repeated encounters (e.g. diabetes), from relatively balanced outcomes to highly imbalanced ones, and from homogeneous numerical features to heterogeneous mixes of categorical and continuous variables. Together, these datasets allow the experiment to test whether SHIELD generalises across varying conditions of clinical tabular data. Their inclusion also reflects considerations of governance and responsible use, as discussed in Section 4.1, ensuring that methodological insights are developed in line with ethical standards of clinical data research.

2.6.1 Obesity

Obesity is a chronic, multifactorial disease characterised by excessive fat accumulation that poses a major risk to health. It has reached epidemic proportions globally, with

the World Health Organisation (WHO) estimating that worldwide obesity has nearly tripled since 1975 [121]. Obesity is associated with increased risk of cardiovascular diseases, type 2 diabetes, musculoskeletal disorders, and certain types of cancer [58]. The aetiology of obesity is complex, involving the interplay between genetic, behavioural, and environmental factors. While body mass index (BMI) remains the most commonly used diagnostic indicator, it is an imperfect measure that fails to capture underlying heterogeneity in fat distribution and metabolic health [26]. Lifestyle factors such as dietary patterns, physical activity, and sedentary behaviour play a particularly critical role in the development and persistence of obesity [77, 81].

In addition to its clinical implications, obesity imposes a considerable economic burden. In the United States alone, obesity-related healthcare costs are estimated to exceed $170 billion annually [116]. Similar trends are observed worldwide, with obesity increasingly affecting low and middle income countries due to rapid urbanisation, changes in food environments, and reduced levels of physical activity [81].

To study this significant disease through a machine learning lens, the University of California, Irvine (UCI) repository provides the *Estimation of Obesity Levels Based on Eating Habits and Physical Condition* dataset [87]. This dataset was constructed from a population sample of individuals in Mexico, Peru, and Colombia, encompassing diverse socio-demographic and lifestyle characteristics. It contains 17 attributes capturing eating habits (e.g. frequency of high-calorie food consumption, vegetable intake, alcohol consumption), physical condition (e.g. frequency of physical activity, use of transportation), and demographic information (e.g. gender, age, family history of obesity). The target variable classifies individuals into one of seven categories ranging from "Insufficient Weight" to "Obesity Type III".

The dataset's inclusion of behavioural and lifestyle factors offers unique value beyond purely clinical measurements such as BMI. In particular, it reflects the multi-dimensional drivers of obesity and enables the evaluation of predictive models that capture not only anthropometric risk but also modifiable lifestyle determinants. This aligns with contemporary medical understanding that obesity is not merely a condition of excess weight but a product of behavioural, social, and environmental interactions. When combined with model explainability techniques such as SHAP, this dataset provides an opportunity to assess whether models identify clinically plausible risk factors in line with epidemiological evidence.

2.6.2 Breast cancer

Breast cancer is the most common cancer among women worldwide, accounting for approximately one in four cancer diagnoses in women [22]. It is a heterogeneous disease with multiple subtypes, broadly classified as invasive or non-invasive, that differ in prognosis, treatment response, and underlying molecular mechanisms [91]. Risk factors are multifactorial, including genetic predispositions, reproductive history, hormonal exposure,

environmental influences and lifestyle factors such as alcohol consumption and physical inactivity [83]. Early detection remains crucial for reducing mortality, with mammography and other imaging techniques forming the cornerstone of screening programs in many countries. In Australia, the National BreastScreen program has significantly contributed to earlier diagnoses and improved survival outcomes, with five-year survival rates now exceeding 90% [7].

Despite these advances, disparities persist. Women in rural and remote areas, as well as Aboriginal and Torres Strait Islanders, experience later diagnoses and poorer outcomes compared to the general population [7]. These disparities highlight the importance of not only technological innovation in diagnostics, but also equitable access to healthcare resources. From a clinical perspective, pathological assessment of tumour biopsies remains the gold standard for diagnosis, with histopathological features such as nuclear size, shape, and chromatin texture being particularly informative in distinguishing benign from malignant tumours.

To facilitate computational research into breast cancer detection, the UCI ML Repository hosts the *Breast Cancer Wisconsin (Diagnostic)* dataset [120]. This dataset originates from digitised images of fine needle aspirates (FNAs) of breast masses, collected and curated by the University of Wisconsin Hospitals. Each sample is described by 30 real-valued features computed from cell nuclei present in the aspirates, such as radius, texture, smoothness, concavity, and symmetry. The features are derived from fundamental morphological and textural properties, capturing clinically salient aspects of nuclear atypia that pathologists use when differentiating between benign and malignant tissue. The dataset consists of 569 instances, of which 357 are benign and 212 are malignant.

The dataset has been used in many ML research as a benchmark due to its balance of medical relevance and computational tractability [105, 42, 107]. When applied in conjunction with explainability techniques such as SHAP, models trained on this dataset can be assessed not only for their predictive accuracy but also for their capacity to highlight biologically plausible markers that align with established diagnostic criteria. As such, the dataset provides a powerful foundation for bridging statistical learning with domain knowledge, reinforcing the importance of model explanations that mirror established medical understanding.

2.6.3 Heart disease

Coronary heart disease is a major contributor to morbidity and mortality in Australia and worldwide. In 2022, cardiovascular disease accounted for about one quarter of all deaths in Australia, and an estimated 600,000 adults had experienced coronary heart disease at some point in their lives [6]. Coronary events are common and costly, with about 57,300 acute coronary events estimated in 2021 [6]. The pathophysiology is driven by atherosclerosis and thrombosis that reduce myocardial perfusion [61]. Clinical presentation ranges from stable angina to acute coronary syndromes. Its risk is shaped by well

established factors, including blood pressure, lipids, smoking, diabetes and electrocardiographic evidence of left ventricular hypertrophy [64, 119]. Contemporary prevention guidelines operationalise these variables through multivariate equations to estimate absolute risk over 10 years [51].

The UCI Heart Disease repository was used in this research, focusing on the Cleveland subset that contains 303 patients with 14 attributes and a target indicating the presence of heart disease [61]. The attributes capture demographics, symptoms, physiology and test results, including sex, chest pain type, resting blood pressure, maximum heart rate and the number of major vessels coloured by fluoroscopy [61]. In the original formulation, the target takes values of 0 to 4 to indicate disease severity, and many studies binarise this to presence versus absence of disease for classification. The Cleveland subset is widely used because it contains the fewest missing values among the sites collected and was used in the seminal validation work by Detrano and colleagues on probability algorithms for coronary artery disease [39]. These variables align closely with established clinical risk constructs, which makes the dataset suitable for evaluating both predictive performance and model explanations in this study.

2.6.4 Diabetes

Diabetes is a chronic metabolic disorder characterised by persistent hyperglycaemia resulting from impaired insulin secretion and/or action. Its global burden is substantial and rising. The International Diabetes Federation estimates that more than 530 million adults live with diabetes, with projections exceeding 780 million by 2045 [109]. Type 2 diabetes accounts for the vast majority of cases and is driven by insulin resistance and progressive beta cell dysfunction, with contributions from genetic susceptibility, adiposity, diet quality, physical inactivity, and social determinants of health. The condition is associated with microvascular complications such as retinopathy, nephropathy, and neuropathy, and macrovascular disease including coronary artery disease, stroke, and peripheral arterial disease. Traditional clinical guidance emphasises comprehensive risk factor management and individualised glycaemic index targets to balance benefits and harms [4].

Unplanned hospital readmission is a salient quality and cost outcome in diabetes care. Readmissions often follow acute metabolic decompensation, infection, or cardiovascular events, and they are associated with higher mortality risk and system costs. Suboptimal inpatient glycaemic control, complexity of comorbidity, polypharmacy, and fragmented transitions of care have all been linked to increased readmission risk [99]. These features make diabetes a clinically meaningful dataset for predictive modelling and for examining how model explanations align with known drivers of adverse outcomes.

This study employed the *Diabetes 130-US Hospitals for years 1999 to 2008* dataset from the UCI ML Repository [31]. The dataset aggregates 101,766 inpatient encounters from 130 hospitals and integrated delivery networks over a 10 year period and was

introduced and analysed by Strack and colleagues in a study of HbA1c measurement and readmission [108]. Each encounter contains demographics and administrative fields (age band, gender, race, admission type, admission source, discharge disposition, time in hospital, payer code, medical specialty), diagnostic information (primary, secondary, and tertiary ICD-9 codes), utilisation history of hospitals (number of inpatient, outpatient, and emergency visits in the prior year), and treatment proxies (number of laboratory procedures, number of medications, diabetes specific medications such as metformin or insulin with indicators of dose change or stability). Laboratory result indicators include categorical summaries for HbA1c (e.g. $> 8\%$, $> 7\%$, normal, none) and serum glucose. The target variable records readmission as `<30`, `>30`, or `NO`. Many studies, including this thesis, binarise this outcome to focus on 30 day readmission.

Several characteristics of this dataset influence modelling and interpretation. First, the outcome is strongly imbalanced, since most encounters are not followed by readmission within 30 days. Second, many predictors are high cardinality categorical variables that require careful encoding to avoid leakage and to maintain clinical plausibility. Third, the dataset contains repeated encounters per patient identifier, which can inflate apparent performance if train and test partitions are not constructed at patient level. Finally, some fields are absent for a large fraction of encounters, including race and payer code. Despite these challenges, the feature set closely mirrors factors known to influence readmission risk, which allows model explanations to be compared against established clinical knowledge about comorbidity burden, intensity of inpatient care, and treatment changes at discharge [108, 99].

Chapter 3

Related Work

This chapter introduces existing literature, which explored the intersections between a pair of or at most three circles in Figure 1.1. Hence, the following works serve as inspiring stepping stones for this thesis to address all of these circles.

3.1 Synergy between SHAP and information theory

Recent studies have explored the synergy between SHAP and MI to balance model explainability and statistical relevance. For example, Kim *et al.* proposed a multi-agent reinforcement learning (RL) framework for clinical feature selection, where each agent's reward combined SHAP attribution with MI scores ($R_i = \alpha \cdot \text{SHAP}_i + (1-\alpha) \cdot \text{MI}_i$) [67]. Their empirical results on a dataset of patients with end-stage renal disease indicated that the method outperformed traditional selection techniques, including PCA, mutual RL and SHAP RL [98, 15], in terms of F1-score and recall, particularly for minority class predictions. This highlights how SHAP can be grounded in statistical dependence measures to prioritise features that are both explanatory and relevant to the outcome.

Another feature selection approach that involves two aspects is introduced by Palanichamy and Ramasamy [85]. Their methods incorporate both MI and CMI to evaluate feature relevance and redundancy within a class-sensitive context. The Improved Mutual Information Feature Selection (IMIFS) algorithm introduces a class-aware scoring mechanism that iteratively selects features maximising relevance (MI) while minimising redundancy (CMI) with respect to already selected features:

$$\text{IMIFS}(f_i) = \frac{2 \cdot I(f_i; C)}{H(C) + H(f_i)} - \frac{1}{|S|} \sum_{f_s \in S} \frac{2 \cdot I(f_i; f_s \mid C)}{H(f_i \mid C) + H(f_s \mid C)}, \tag{3.1}$$

where C is the class label and S is the set of already selected features. Evaluated on UCI datasets, IMIFS achieved higher accuracy and more compact feature sets compared to standard MI-based methods. The results underline the importance of identifying correlation or causal inference between features and classes, which a sole use of SHAP inherently lacks of.

Expanding the information-theoretic perspective, Manikandan and Abirami proposed a two-stage selection process based on MI and Monte Carlo Tree Search (MCTS) for filtering both redundant and irrelevant features [78]. Initially, approximate Markov blankets

are used to eliminate redundancy, followed by Monte Carlo exploration of feature subsets to refine relevance assessment. This approach demonstrated statistically significant improvement in classification accuracy across multiple microarray datasets, while dynamically adapting the selected feature size without fixed thresholds. This suggests a natural extension where SHAP explanations are anchored by information-theoretic criteria to enhance both fairness and interpretability.

3.2 Feature grouping literature review

While many approaches, including those mentioned in previous section, treat feature selection or elimination distinct from grouping, Kuzudsli *et al.* conducted extensive review on existing grouping-based feature selection [70] . In supervised learning settings, feature selection (FS) through grouping has become an increasingly used strategy, particularly in high-dimensional domains such as genomics and image analysis [101]. The central idea is to cluster features into groups based on some similar metric and then select representative features from each group. This not only reduces dimensionality but can also enhances interpretability and model performance.

Clustering-based grouping is by far the most common strategy. Several representative works illustrate its variations:

- One of the notable methods in this category is the ensemble-based clustering and ranking technique by Yu *et al.* (2020), who used K-means clustering to form feature groups followed by three independent ranking strategies (t-test, signal-to-noise ratio, and SAM) [122]. Their method culminates in an ensemble feature selection where a feature must appear across all subsets to be retained. This redundancy check aims to reinforce feature relevance but overlooks inter-feature correlations, a critical aspect in domains like genomics where gene co-expression patterns are common.
- Shang *et al.* (2007) implemented hierarchical clustering using an information compression index to group features [103]. Within each cluster, they applied the Fisher criterion to rank features, selecting the top-ranked feature as representative. This method uniquely focuses on maximising class separability within clusters, offering a nuanced balance between information compression and discriminatory power.
- Zhang *et al.* (2018) modified the affinity propagation algorithm to generate feature clusters and subsequently applied a sequential selection strategy to each cluster [125]. This two-step approach integrates unsupervised grouping with a wrapper-style FS, aiming to retain contextual interdependencies within clusters during selection.

Beyond clustering, alternative formulations are discussed to emphasise stability or direct integration with learning algorithms:

- He and Yu (2010) deviated from traditional clustering via kernel density estimation, paired with mean shift clustering [55]. They then applied an F-statistic-based evaluation to select representative features. This approach emphasises feature stability across different data samples, a crucial attribute often neglected in conventional methods. Their follow-up work extended this to an ensemble framework to further enhance feature robustness.
- Regularisation technique is another application of this topic. Fahrmeir *et al.* (2009) introduced a method that leverages regularised regression models to determine feature groupings and representatives, inherently embedding the grouping process within the model training phase [46]. This integration is particularly beneficial for managing the bias-variance trade-off in high-dimensional datasets.

Each of these methods contributes distinct perspectives on how to effectively harness feature grouping in various settings. While some prioritise computational efficiency and simplicity, others delve into complex models aiming for higher accuracy and stability. This suggests how the choice of method often hinges on the specific requirements of the application domain, such as the need for model interpretability, computational resources, or dataset structure. In the case of this thesis, grouping is pursued not to reduce dimensionality, but rather to redistribute explanatory credit and improve equity in model explanations under clinical conditions.

3.3 Fairness metrics from the perspectives of outcome, statistics and explanation

Beyond the technical robustness and interpretability achieved through feature selection and attribution tools like SHAP, MI and CMI, fairness remains a critical dimension of trustworthy machine learning. Algorithmic decisions must be transparent, equitable and fair across demographic groups, especially in high-stakes applications. Several works have examined the intersection of explainability and fairness, highlighting both theoretical tensions and practical solutions.

One of the foundational critiques of existing fairness metrics is presented in the work by Hardt *et al.*, who argue that demographic parity, which requires outcomes to be independent of protected attributes, may paradoxically promote unfair treatment by admitting unqualified individuals or denying qualified ones solely for the sake of statistical parity [54]. As a remedy, they introduce two alternative criteria: Equalised Odds and Equal Opportunity, both rooted in the joint distribution of predictions, outcomes, and protected attributes. Their framework also proposes a post-processing correction, applied after model training, which adjusts decision thresholds to satisfy the desired fairness constraints without altering the internal model parameters. Notably, their method preserves

predictive accuracy better than adjustment based on demographic parity while improving fairness across subgroups. This indicates the metric is more suitable for deployment in sensitive domains such as credit scoring or criminal justice.

While Hardt *et al.* focused on post hoc fairness correction, Jain *et al.* propose a novel statistical framework called N-Sigma to measure algorithmic bias in AI models, particularly in face recognition systems [60] . Inspired by the 5-sigma threshold used in hypothesis testing in physics, this metric quantifies performance disparities between demographic groups as a continuous, interpretable value:

$$N = \frac{\mu_{G1} - \mu_{G2}}{\sigma_{G1}}, \tag{3.2}$$

where μ_{G1} and μ_{G2} are the means of the two populations being compared and σ_{G1} is the standard deviation of the population used as reference. Unlike binary hypothesis tests (e.g. t-tests), N-Sigma facilitates risk-tiered decisions. Models with sigma differences exceeding certain thresholds can be classified as moderate or high risk. Their evaluation on the Racial Faces in the Wild (RFW) dataset revealed that even models trained to be demographically neutral still exhibited considerable disparities, underscoring the necessity for distribution-aware fairness metrics. Importantly, this method provides a regulatory-aligned lens for AI risk assessment, in line with emerging policy frameworks such as the EU's AI Act.

Another complementary perspective is offered by Zhao *et al.*, who delve into the biases intrinsic to SHAP-based explanations [127]. Their error analysis framework distinguishes between observation bias (due to data sparsity) and structural bias (due to simplifying assumptions such as feature independence). These biases result in over-informative or under-informative explanations. The concepts are formalised and empirically evaluated using datasets like Bike Sharing and Census Income. Using a novel OOD (Out-of-Distribution) detection-based total variation distance (TVD) metric, the authors show that structural bias is particularly severe under assumption-based removal functions (e.g. marginal, uniform), where distributional drift can exceed 80-90%. This work highlights the trade-off between tractability and fidelity in XAI, pointing toward the need for hybrid or adaptive methods that balance data availability with distributional accuracy.

Further bridging the gap between explainability and fairness, Jain *et al.* also demonstrated that biased models inherently produce biased explanations, corroborating the hypothesis that explanation tools must be scrutinised for fairness, not just accuracy [60]. Their work proposes a SHAP-based post-processing algorithm that detects and mitigates bias using quadrant-based intervention strategies. This method plots instances on a SHAP value versus prediction deviation plane, and adjusts outcomes based on their quadrant and distance from a fairness-neutral baseline (i.e. origin of the plane). This approach differs from traditional post-processors (e.g. random flipping) by using individual SHAP contributions as justification for correction, thereby improving individual-level fairness without sacrificing group-level metrics. Tested on the COMPAS dataset,

the method successfully reduced racial bias in recidivism predictions, while maintaining classification performance.

Guided by these findings, the evaluation of this thesis incorporates both outcome and explanation level fairness. Specifically, SHIELD combines parity metrics, N-sigma and the bias-quadrant view that jointly highlight disparities in predictions and explanations. Unlike post-hoc thresholding or label-flipping approaches, this work targets representation-level interventions, aiming to spread model reliance more evenly across features. This design choice directly addresses mechanisms highlighted by prior work, which showed that biased models yield biased explanations and that explainer assumptions can introduce structural errors.

Together, these works illustrate a multifaceted approach to fairness, ranging from theoretical definitions and statistical audits to explanation-based corrections. They highlight that transparency and fairness are mutually reinforcing components of responsible ML deployment. As SHAP becomes more widely adopted, its applications must be understood not only in terms of attribution explainability but also in how it interacts with underlying model biases. This integration of fairness into the ML pipeline with consideration of XAI is essential for advancing models that are not only technically robust but also socially aligned.

Chapter 4

Methodology

This chapter details the end-to-end methodology underpinning SHIELD, the fairness-aware clinical ML pipeline, as visually summarised in Figure 4.1. It begins by motivating the choice of four publicly available, de-identified UCI medical datasets and recording the ethical safeguards that govern their secondary use. A rigorous preprocessing pipeline is then described: missing values were imputed with an XGBoost-based classifier; categorical variables were encoded with a tiered one-hot, label, or ordinal strategy; and numerical features were standardised where required. A principled train-test split ensured the test set remained untouched by imputation, averting information leakage.

The core innovation of SHIELD was the dissimilarity-based feature-grouping framework driven by CMI. Three anticlustering algorithms, which are greedy, bicriterion, and K-plus, were introduced, each optimising diversity and dispersion in complementary ways. A data-driven procedure with graphical heuristics determined the optimal number of partitions, balancing intra-group heterogeneity against downstream predictive and fairness performance. Group-specific autoencoders then produced compact latent embeddings while preserving feature-level interpretability via decoder-weight reconstructions.

Five representative classifiers, Logistic Regression, SVM, MLP, Random Forest, and XGBoost, were trained on these embeddings. Hyperparameters were tuned jointly through Gaussian-process Bayesian optimisation to maximise cross-validated Receiver Operating Characteristic - Area Under Curve (ROC-AUC) subject to fairness constraints [106]. Finally, a multi-layered evaluation protocol assessed predictive performance (accuracy, precision, recall, F1), explanation fidelity (SHAP decomposition), and algorithmic fairness (Equal Opportunity, Equalised Odds, N-Sigma index, and bias-quadrant analysis). This comprehensive framework ensures that any gains in equity and transparency are achieved without clinically unacceptable losses in performance, thereby laying a rigorous foundation for the empirical results presented in the next chapter.

Note that the code and artefacts for this thesis can be found at this repository.

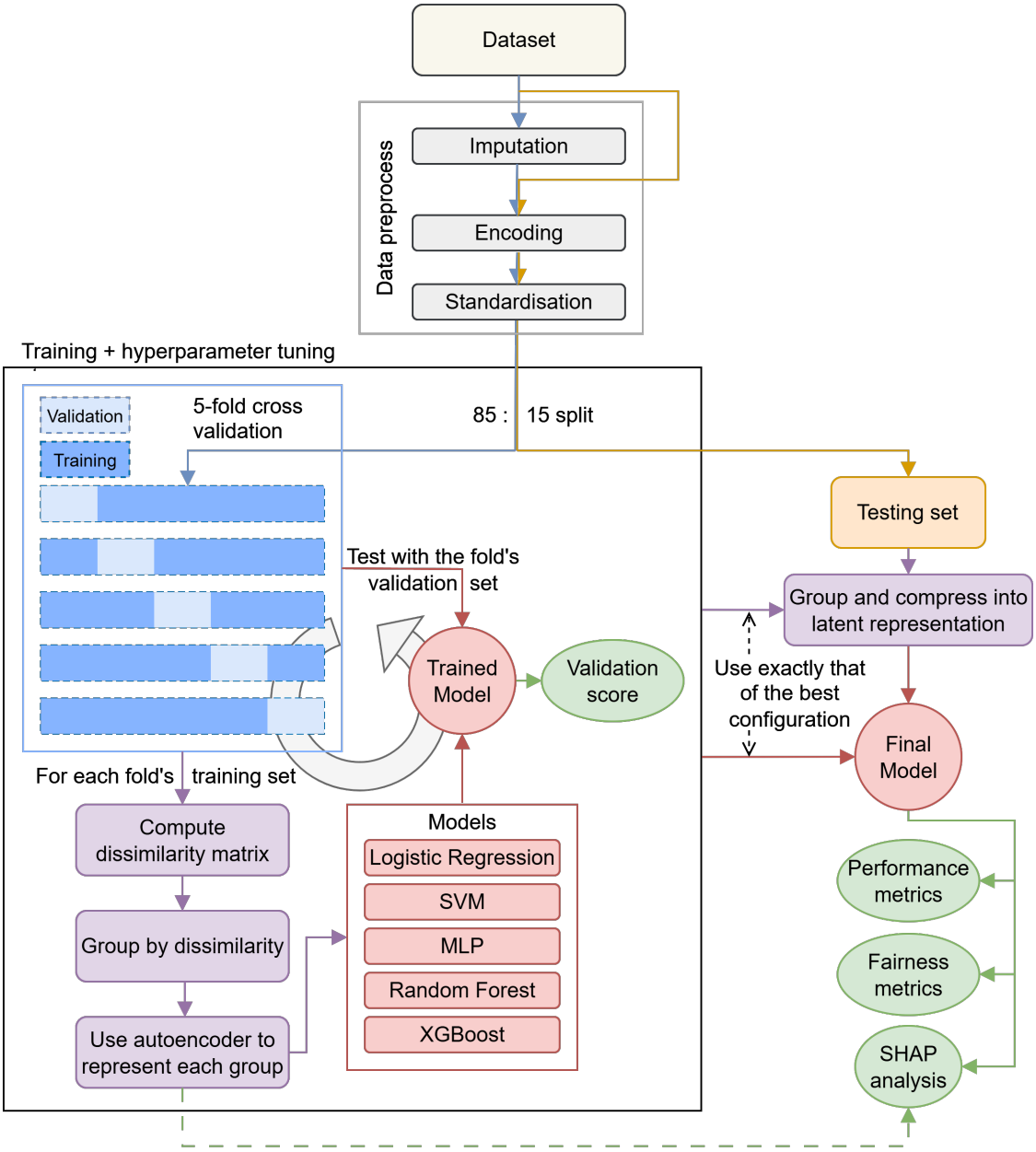


Figure 4.1: End-to-end workflow of SHIELD.

4.1 Data collection

All data used in this study were existing, de-identified clinical and biomedical dataset obtained from the UCI Machine Learning Repository [113]. Thus, no primary data

collection from human participants was performed.

Ethics regarding the data collection should never be treated lightly, despite its easiness to neglect it. It is not just a matter of legal compliance but also a fundamental aspect of responsible and trustworthy data practices. To properly uphold this expectation, 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, "Explainable Health Informatics"), under the low-risk review pathway on 13th of June, 2025 [112]. As a result, all raw data files were imported directly from the repository via `ucimlrepo` package, and accessed only by the project team. No identifiable information was present in the datasets, and all analyses adhered to the ANU Ethics Office's data management and confidentiality requirements.

As can be seen in Tables 4.1 and 4.2 below, the four benchmark datasets were selected to cover a range of clinical classification tasks, each licensed under Creative Commons Attribution 4.0 and widely used in the literature. Evidently, there was a great deal of research that used these datasets to propose, verify and test their models [105, 42, 40, 130]. Furthermore, it was intentional that most datasets had an imbalanced class distribution, because it is reflective of the inherent property of medical datasets. A balanced Obesity dataset was still chosen to expose the influence of class distribution. Nevertheless, this Obesity dataset was resampled using SMOTE to create Obesity_imbalanced with the following target class proportions: Insufficient weight - 50%, Normal weight - 20%, Overweight I - 10%, Overweight II - 8%, Obesity I - 6%, Obesity II - 4%, Obesity III - 2%. This adjustment was to directly examine the resilience of grouping methods under skewed label distributions, by making use of the same dataset in two different class distributions. Consequently, the datasets were wisely chosen for their clinical relevance, diversity of feature types, and public availability, ensuring reproducibility and comparability.

Attribute	**Obesity**	**Diabetes**
D (# of features)	16	47
N (# of instances)	2111	101766
C (# of Classes)	7 (Balanced)	3 (Imbalanced)
Class Distribution (%)	14, 14, 14, 17, 13, 14, 15	54, 35, 11
Context of missing values	MAR	MAR
Views (k)	142.06	84.43
Benchmark Exists?	No	No
Notes	$D \ll N$	$D \ll N$

Table 4.1: Datasets without benchmark [87, 31].

Attribute	Heart Disease	Breast Cancer
D (# of features)	13	30
N (# of instances)	303	569
C (# of Classes)	5 (Imbalanced)	2 (Imbalanced)
Class Distribution (%)	54, 18, 12, 12, 4	63, 37
Context of missing values	MAR	MAR
Views (k)	699.04	426.15
Benchmark Exists?	Yes	Yes
Notes	$D < N$	$D < N$

Table 4.2: Datasets with benchmark [61, 120].

4.2 Data preprocessing

Robust and interpretable machine learning results depend just as much on the quality of data preparation as on the choice of algorithm. Accordingly, the following subsections explain the step-by-step approach to preserving clinically meaningful extremes, converting categorical information into numerical form, placing heterogeneous features onto comparable scales, and reconstructing plausible values for missing entries. By clarifying why each decision is made and which models truly require it, this section lays the foundation for the fairness, performance, and explainability analysis presented in later chapters.

Effective data preprocessing is fundamental to building robust and accurate machine learning models [68]. Without properly curated and prepared data, even the most sophisticated algorithms may underperform or perform unfairly [69]. Thus, this study treated preprocessing not as a perfunctory step but as an integral component of methodological rigour.

Three core operations formed the backbone of the preprocessing pipeline. First, imputation replaced missing values so that each observation remained usable for model training. Secondly, encoding translated categorical variables common in medical records into numerical representations compatible with most algorithms. Lastly, scaling (normalisation or standardisation) mitigated the dominance of features measured on larger numerical ranges. In sum, these procedures transformed raw data into inputs that satisfied the assumptions of each model.

Table 4.3 summarises how the five representative models used in this thesis interact with those preprocessing stages. As corroborated by the table, most linear, kernel, and neural models require all three components, while tree-based ensembles are more flexible.

Notably, XGBoost can ingest unscaled, non-encoded inputs and natively handles missing values, making it attractive when preservation of a pristine test set is paramount. This heterogeneity in requirements motivates the granular discussion of each step provided in the following subsections.

Model	Requires encoding	Requires standardisation	Requires to handle missing values
Logistic Regression	Yes	Optional/beneficial	Yes
SVM (RBF kernel)	Yes	Yes (highly recommended)	Yes
Neural Network (MLP)	Yes	Yes (highly recommended)	Yes
Random Forest	Yes	No	Yes
XGBoost	No	No	No

Table 4.3: Preprocessing Requirements for Selected Models [69, 76].

4.2.1 Outlier removal

Outlier removal is commonly used in ML pipelines to reduce the influence of extreme values that may distort model training and degrade generalisability [2]. Retaining unjustified extreme observations carry risk. They can exert disproportionate leverage on fitted parameters, distort decision thresholds, and inflate variance, especially for margin or distance based learners and models trained with non-robust losses [69, 131, 2]. However, its appropriateness is highly context-dependent, especially in clinical research, where numerically atypical values may represent valid and diagnostically significant cases rather than errors [131].

In this study, no outlier removal was performed for clear reasons. First, the original introduction of the Diabetes dataset by Strack et al. [31] explicitly described the inclusion criteria for 70,000 inpatient diabetes encounters but did not report statistical outlier filtering. Instead, they highlighted the importance of preserving real-world variability to examine historical patterns of care, aligning with best practice for large observational datasets. Similarly, the Heart disease dataset by Detrano et al. [39] demonstrated that valid extreme values were critical for probability modelling and must not be removed without clear evidence of error.

Secondly, subsequent applications of comparable clinical datasets show that synthetic balancing or imputation may be used for missing data (e.g. SMOTE balancing for obesity levels in Mendoza et al. [86]), but valid extreme observations are retained unless domain knowledge confirms they are artefactual. Removing legitimate edge cases could bias results, especially in medical research where severe cases often carry critical information.

In the context of real-world hospital data, extreme lab results or unusually long hospital stays can reflect severe disease trajectories or rare complications, not noise. Therefore,

consistent with the established precedent of the original datasets, no outlier removal was performed to preserve the integrity and representativeness of the cohort.

4.2.2 Encoding

Many medical datasets contain categorical variables that must be expressed numerically before most learning algorithms can process them. Broadly, three encoding strategies are available, namely one-hot, label, and ordinal, and the choice depends on (i) whether the category has an inherent order and (ii) whether the variable is an input feature or the prediction target. Selecting an inappropriate strategy can inject spurious structure into the data and bias downstream analyses, so the rationale for each decision is stated explicitly below.

Nominal features such as race or medication type lack intrinsic ordering. To ensure that no artificial hierarchy is imposed, these variables were encoded with one-hot encoding, which expands each category into a binary indicator column [92]. Although effective and model-agnostic, one-hot encoding inflates dimensionality and can produce sparse matrices when categories are numerous. This trade-off was deemed acceptable given the modest cardinality of the nominal variables in the datasets used in this thesis.

For the outcome variable, splitting a single column into multiple one-hot columns would complicate performance metrics and probabilistic calibration. Instead, label encoding was applied, assigning an integer identifier to each outcome class [92]. Because tree-based models treat these integers as mere labels, no ordinal bias is introduced. For linear or kernel methods, potential ordering artefacts were mitigated by using one-versus-rest decision functions during training.

Variables such as symptom severity or age bracket convey a natural rank. These were transformed with ordinal encoding so that the numeric representation preserves monotonic relationships (e.g. $Never \rightarrow 0$, $Sometimes \rightarrow 1$, $Often \rightarrow 2$, $Always \rightarrow 3$). Ordinal encoding maintains the information content of the original scale while avoiding the dimensionality explosion associated with one-hot encoding.

Taken together, this tiered approach of one-hot for unordered predictors, label encoding for the target, and ordinal encoding for ranked predictors provided a principled mapping from categorical data to numeric space while minimising information loss and modelling bias.

4.2.3 Normalisation and standardisation

It is common to see real-world datasets to contain features with varying units and scales since each feature inherently has its unique statistics (e.g. Age will typically be in order of 10^1, while salary can be in order of 10^4, 10^5 or even higher). For some models, such disparity can adversely impact model convergence and performance [68]. Normalising

or standardising these features ensure all variables contribute comparably to distance-based models and accelerates learning in gradient-based algorithms. Common methods are as follows [89]:

$$\text{Min-Max Normalisation: } X_{\text{new}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}, \tag{4.1}$$

$$\text{Z-score Standardisation: } X_{new} = \frac{X - \mu}{\sigma}, \tag{4.2}$$

where μ, σ are the mean and standard deviation for X, respectively. The Z-score standardisation maps each feature to a normal distribution with zero mean and unit variance, so values are unbounded but typically lie in $[-3, 3]$. Min-max normalisation maps to $[0, 1]$, where larger values correspond to higher raw magnitudes.

The normalisation method is preferred when the model assumes bounded input, since it is sensitive to outliers and unseen data containing values outside the original range can lead to distortion. On the other hand, the standardisation method handles outliers more robustly through the use of mean and standard deviation, instead of range. It is also suitable for algorithms that assume normality, such as logistic regression and SVM, which are used in this thesis. Hence, SHIELD utilised the standardisation where necessary. It should also be noted that this was applied to ordinal encoded features, but not nominal ones as they do not have such mean and standard deviation.

4.2.4 Imputation

Missing data is a pervasive issue in clinical datasets due to recording inconsistencies, omitted tests due to confidential or privacy constraints [28]. In this study, some models did not accept any missing value. A potential bias towards data instances with missing values should be considered even if a model tolerated them. The simplest approach was to drop either instances (row) or features (columns) with missing data [71]. While it may have been effective under MCAR (Missing Completely at Random) assumptions, it risked substantial data loss if the missing rate was high.

More sophisticated methods aim to estimate values based on observed data as follows:

- Classifier-based imputation [62]: Treats imputations as a supervised learning task, training a classifier to predict missing nominal values using other observed features. This method accommodates missingness in other columns and performs well under MAR (Missing at Random) assumptions.
- Label spreading/propagation [128]: Semi-supervised learning techniques using graph-based smoothing. They require encoded inputs and assume strong inter-feature relations but do not support missingness in features beyond the target.

- Iterative imputation [8]: Models each feature with missing values as a function of other features in a round-robin fashion. It balances accuracy and complexity, although it is sensitive to model bias.
- KNN imputation [111]: Fills in missing values using the average (or mode) of k nearest neighbours. It is effective under MCAR but computationally expensive and sensitive to scaling.
- Simple imputation [71, Sect. 4]: Replaces missing entries with mean, median or mode value. It is fast and robust under MCAR, but naive under MAR or MNAR assumptions.

Table 4.4 depicts suitability in different contexts and requirement for each imputation method. As can be seen, classifier-based imputation supports complex feature interactions and high missing rates without requiring full data encoding upfront. In addition, all of these benefits come with a relatively efficient time complexity (assuming the conventional case, where $n < d$). Consequently, classifier-based imputation using XGBoost was adopted for key nominal features with substantial missingness.

Method	**Context of missing data**	**Handles high missing rate**	**Requires encoding**	**Accepts missing values in other features**	**Time complexity**
Classifier-based	MAR	Yes	No	Yes	$O(n \cdot d)$
Label spreading	MAR	Yes	Yes	No	$O(n^2)$
Label propagation	MAR	Yes	Yes	No	$O(n^2)$
Iterative Imputer	MAR, MCAR	Moderately	Yes	Yes/No	$O(k \cdot n \cdot d^2)$
KNN Imputer	MCAR	Moderately	Yes	No	$O(n^2 \cdot d)$
Simple Imputer	MCAR	No	Yes/No	Yes	$O(n)$

Table 4.4: Comparison between imputation methods.

4.2.5 Order of preprocessing

While the preceding subsections have detailed each preprocessing component individually, the sequence in which these operations are applied is equally critical for maintaining data integrity and ensuring model compatibility. The models considered in this study fell into three categories based on their preprocessing requirements: (i) requiring all the preprocessing components, (ii) requiring encoding and imputation only, and (iii) requiring none.

For models that required all components, the order was carefully structured as follows:

1. Imputation: Missing values should be handled prior to any transformation of categorical variables. This preserved the original data semantics and avoided introducing artificial patterns that may have biased the imputation model, especially if encoding added dimensionality or implied ordinal relationships where none existed.

2. Encoding: Once missing values were imputed, categorical features were converted into a numerical format suited for the learning algorithm.
3. Standardisation: This step was last since missing value or ordinal categories could not be standardised.

As illustrated in Figure 4.2, the order was preserved for models that required partial preprocessing, as they skipped any unnecessary steps (e.g. RandomForest : imputation → encoding. By adhering to this systematic order, the preprocessing pipeline maintained consistency and mitigated the risk of information leakage or distortion during transformation.

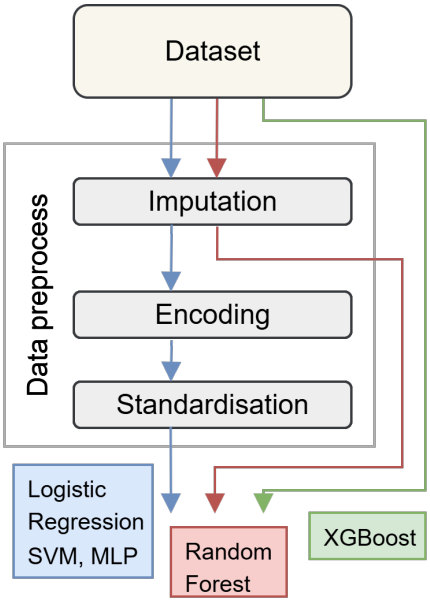


Figure 4.2: Data preprocess flowchart for each model.

4.3 Train-test split

A principled train-test split was essential to reliably assess the generalisability of ML models [69]. The testing set was to be treated truly unseen and thus remained as untouched and independent as possible throughout the preprocessing and model training stages. Any information leakage from the training process into the test set could lead to overly optimistic evaluations and undermine the validity of model comparisons.

In this study, particular care was taken to ensure the testing set was constructed exclusively from rows that were originally complete, as illustrated in Figure 4.1. This eliminated the need to apply imputation methods to the test set, thereby avoiding the

risk of incorporating learned patterns or distributional biases from the training data into the imputation model.

However, other steps, encoding and standardisation, still needed to be applied to the test set to ensure compatibility with the model. For instance, categorical features and numerical features had to be encoded and standardised, respectively, using the same mappings that derived the training data. This ensured the feature representations in the test set were consistent with those seen during model training, without introducing information leakage.

4.4 Feature grouping by dissimilarity

In this study, feature grouping played a critical role in promoting fairness [123, 75, 65], not just enhancing dimensionality reduction [88]. The central idea was to separate features 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), can still lead to biased model behaviour if their collective influence remained unchecked [10, 36]. Hence, the notion of grouping features by dissimilarity, rather than similarity, was a deliberate strategy to mitigate such risks [24]. If similar features, including potential proxies, were grouped together, their joint effects may have become more pronounced, leading to biased representations in latent space. Thus, grouping by similarity would have required subsequent adjustments to explicitly manage proxy variables, as it intends to consolidate correlated features. 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 [88]. This matrix quantified the degree to which each pair of features provides different information with respect to the target variable. In particular, dissimilarity is computed as the complement of CMI between features, aimed to identify features that contribute unique, non-redundant signals to the prediction task [114].

Let X_i and X_j denote two input features, and Y be the target variable. The dissimilarity between X_i and X_j is defined using their CMI given Y, which captured the shared information between the two features conditional on the outcome variable [34]. Mathematically, CMI was expressed as

$$I(X_i; X_j \mid Y) = \sum_{x_i, x_j, y} p(x_i, x_j, y) \log \left(\frac{p(x_i, x_j \mid y)}{p(x_i \mid y) p(x_j \mid y)} \right). \tag{4.3}$$

To standardise the scale of CMI and obtain a bounded measure of dissimilarity, it was normalised through the sum of marginal entropies:

$$\text{Normalised CMI}_{i,j} = \frac{I(X_i; X_j \mid Y)}{H(X_i) + H(X_j) + \epsilon}, \tag{4.4}$$

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

$$D_{i,j} = 1 - \text{Normalised CMI}_{i,j}. \tag{4.5}$$

The matrix D with entries $D_{i,j}$ is symmetric and encodes how dissimilar each pair of features 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, 114]. Consequently, dissimilarity $D_{i,j}$ is also in the same interval $[0, 1]$, with larger values indicating more distinct features.

4.4.1 Evaluation metrics for grouping

Each grouping method, despite relying on different heuristics or optimisation strategies, sought to fulfil a common objective: maximising dissimilarity within groups to weaken the collective impact of correlated or proxy features. Two primary metrics were employed to assess the quality of feature grouping:

- Diversity: This non-negative metric quantified the average dissimilarity between features that belonged to the same group. Formally, for each group G_k containing feature indices $i, j \in G_k$, the diversity was computed as

$$\text{Diversity} = \sum_{\forall k} \sum_{(i<j)\in G_k} \frac{D_{i,j}}{K|G_k|}. \tag{4.6}$$

  A higher diversity indicated that features within each group were more distinct from each other.

- Dispersion: This was a more conservative non-negative metric, focusing on the minimum pairwise dissimilarity between any two features within a group.

$$\text{Dispersion} = \min_{\forall k}\{\min_{(i<j)\in G_k}\{D_{i,j}\}\}. \tag{4.7}$$

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

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

4.4.2 Naive approach

The naive feature grouping strategy adopted a straightforward greedy algorithm 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 features, ensuring that each group begins with a representative feature that was maximally distinct from others. Then, the algorithm iteratively assigned the remaining features to the group for which they exhibited the highest average dissimilarity with existing group members. This greedy assignment continued until all features were allocated.

The simplicity of this method allowed for efficient computation, serving it as a useful baseline for evaluating more sophisticated grouping approaches. Also, the number of groups is denoted as K throughout this thesis instead of the conventional k (hence K-plus, not k-plus), as the lower case is reserved for number of folds in the cross validation.

Algorithm 1 Naive Feature Grouping via CMI-based Dissimilarity

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

4.4.3 Bicriterion approach

The bicriterion approach to anticlustering simultaneously maximised two complementary criteria, diversity and dispersion [25] (see Algorithm 2). It aimed to avoid configurations where high overall diversity might still allow clusters of closely related (potential proxy) features 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 features to groups to improve the following objective:

$$\text{obj} = \alpha \cdot \text{Diversity} + \beta \cdot \text{Dispersion}, \tag{4.8}$$

where α, β quantified the priorities of each criterion.

Algorithm 2 Bicriterion Anticlustering

Input: Dissimilarity matrix D, number of groups K, weights α, β
Randomly initialise groups $G_1, ..., G_K$
repeat
  Compute Diversity and Dispersion for current partition p
  for each pair of features (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

4.4.4 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 [88]. 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.

Formally, this was achieved by constructing a set of augmented features 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. This formulation allowed for fine-tuned control over the statistical similarity of groups. Optimisation was performed through local search heuristics that iteratively swapped features between groups to improve the composite objective.

Algorithm 3 K-plus Anticlustering

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

4.4.5 Choosing the most optimal number of partitions

The hyperparameter K, the number of feature partitions, determined how well dissimilarity-based grouping fulfils its dual goals of dimensionality reduction and fairness. When K is set too small, correlated or proxy features are inevitably forced into the same group, amplifying their combined effect and undermining the intended fairness protection. Conversely, if K approaches the total number of raw features, grouping degenerates into an identity mapping that offers neither interpretability nor computational benefit. Selecting an optimal K therefore involves balancing intra-group heterogeneity against model-level performance.

Initially, intrinsic anticlustering quality was inspected as a function of K by plotting diversity and dispersion, over $K \in \{2, \ldots, 10\}$. Analogous to the elbow and silhouette diagnostics in conventional clustering [97], the intersection of diversity and dispersion curves often revealed a knee point beyond which additional partitions yielded diminishing returns. This graphical heuristic offered a quick sanity check before more computationally intensive searches.

The following systematic search was unfortunately not executed due to time and computational constraints, which is further discussed in Section 5.5. Because K interacts with downstream learning objectives, it was planned to be treated as another hyperparameter in the Bayesian optimisation loop described in Section 4.5.2. The surrogate model would have jointly explored K and other grouping weights (α, β), proposing settings that maximised five-fold cross-validated ROC-AUC. This end-to-end search would have ensured that the chosen K aligned with both performance and fairness criteria rather than solely with internal dispersion statistics.

4.4.6 Latent representation of groups

Once the feature groups have been identified, it became 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 feature grouping based on similarity, where dimensionality reduction techniques such as PCA can capture dominant correlated directions, the groups in SHIELD were intentionally constructed to consist of dissimilar features. Consequently, summarisation strategies relying on correlation or redundancy were ineffective. Instead, a neural network-based approach using group-specific autoencoders was adopted.

For each group, an autoencoder was trained to learn an efficient encoding of the group's feature set. Formally, let $X^{(k)} \in \mathbb{R}^{n \times d_k}$ denote the matrix of d_k features in k-th group across n samples. A group-specific autoencoder learns an encoding function $f_k : \mathbb{R}^{d_k} \rightarrow$

$\mathbb{R}^m$ and a decoding function $g_k : \mathbb{R}^m \to \mathbb{R}^{d_k}$ such that the reconstruction loss was minimised:

$$\min_{f_k, g_k} \left(\frac{1}{n} \sum_{i=1}^{n} \left|\left| X_i^{(k)} - g_k(f_k(X_i^{(k)})) \right|\right|^2 \right). \tag{4.9}$$

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 features.

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

Concretely, under the setting of Equation (4.9), the decoder's linear layer was expressed as

$$g_k(z) = W_{\text{dec}}^{(k)} z + b^{(k)}, \quad W_{\text{dec}}^{(k)} \in \mathbb{R}^{d_k \times m}. \tag{4.10}$$

Then, each column of $W_{\text{dec}}^{(k)}$ described how one latent coordinate contributed to all d_k original features. Suppose a downstream classifier produces a vector of latent-space attributions

$$\phi^{(k)} = [\phi_1, \ldots, \phi_m]^T \in \mathbb{R}^m \tag{4.11}$$

for group k. To distribute these back to the original features, the elementwise absolute weight matrix was formed

$$\left| W_{\text{dec}}^{(k)} \right|, \left| W_{\text{dec}}^{(k)} \right|_{ij} = \left| W_{\text{dec}}^{(k)}[i, j] \right|. \tag{4.12}$$

and normalised each latent-to-feature mapping so that the contributions summed to one:

$$\tilde{W}_{ij} = \frac{\left| W_{\text{dec}}^{(k)} \right|_{ij}}{\sum_{i'=1}^{d_k} \left| W_{\text{dec}}^{(k)} \right|_{i'j}}, \text{for } \; (i = 1, \ldots, d_k; j = 1, \ldots, m). \tag{4.13}$$

The final feature-level attribution vector for group k was then

$$\phi_{\text{original}}^{(k)} = \tilde{W}^{(k)} \phi^{(k)} \in \mathbb{R}^{d_k}, \tag{4.14}$$

so that each original feature i inherited

$$\left[\phi_{\text{original}}^{(k)} \right]_i = \sum_{j=1}^{m} \tilde{W}_{ij} \phi_j, \tag{4.15}$$

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

By concatenating these per-group decompositions across all K groups, an attribution vector was recovered over the full original feature space. This principled usage of decoder-weight-based mapping preserved the interpretability of individual variable even after dimensionality reduction, enabling both SHAP-driven explanations and fairness assessments at the original feature level.

Once trained on the training data, the group-specific autoencoders were fixed and reused to transform the test set. This ensured consistent latent representations across both training and test phases, avoiding leakage and preserving the integrity of the learned transformations. By applying the same encoding functions f_k to the test data, the model guaranteed comparability of latent vectors and prevented retraining-induced drift that could have biased evaluation metrics.

In sum, autoencoders provided a principled and flexible framework for deriving latent group embeddings, well-suited for the fairness-aware, dissimilarity-based grouping paradigm employed in this study.

4.5 Training

This section details how models were fit and tuned on both raw and grouped feature representations to isolate the effect of dissimilarity-based grouping on performance and fairness. It begins by motivating the choice of five representative classifiers, namely Logistic Regression, SVM, MLP, Random Forest, and XGBoost, which span linear, kernel, neural, and tree-ensemble families (see Subsection 4.5.1). It then describes the hyperparameter optimisation procedure that standardises comparison across models and representations, using cross-validated Bayesian optimisation with consistent preprocessing, data splits, and evaluation protocols (see Subsection 4.5.2). These design choices ensure that any observed patterns arise from the representation strategy rather than eccentricity of a particular learner or tuning regime.

4.5.1 Base models

This study evaluated the effectiveness of fairness-aware feature grouping strategies across a diverse set of machine learning models, each representing a distinct family of learning paradigms. The chosen models included both linear and non-linear learners, interpretable and complex architectures, as well as tree-based and neural approaches. This heterogeneity allowed for a robust assessment of the generalisability and fairness implications of the proposed methodology [47]:

- **Logistic Regression** is a widely used linear model suitable for binary classification tasks. It models the probability of the positive class using the logistic function and is optimised via maximum likelihood estimation [57]. Due to its simplicity and interpretability, it serves as a strong baseline for performance and fairness assessments. However, it assumes linear relationships between features and the log-odds of the target, which may limit its capacity on complex datasets. Its coefficients offer direct interpretability, making it a favoured model in clinical and policy-related applications where transparency is paramount.
- **Support Vector Machine (SVM)** with a radial basis function (RBF) kernel is employed to capture non-linear decision boundaries. It aims to maximise the margin between classes while using kernel tricks to implicitly project data into higher-dimensional spaces [32, 21]. SVMs are particularly effective in high-dimensional settings and provide robust decision boundaries in the presenece of outliers . However, they require extensive hyperparameter tuning and are sensitive to feature scaling. Moreover, their-black-box nature makes interpretation challenging, particularly when fairness explanations are required [16].
- **Multi-Layer Perceptron (MLP)** is a feedforward neural network consisting of multiple fully connected layers with non-linear activation function [100]. It is capable of learning complex patterns through backpropagation and gradient-based optimisation. MLPs offer significant representational flexibility but require careful tuning of architecture and regularisation to prevent overfitting [48], especially in small-to-medium sized datasets. Their non-linearity and depth enable the modelling of intricate feature interactions, but they also obscure the individual contribution of each feature, complicating fairness attribution and interpretability unless supported by post-hoc explainability tools [80].
- **Random Forest** is an ensemble method based on decision trees trained on bootstrapped subsets of the data with random feature selection at each split [23]. It offers strong performance and robustness to noise and overfitting, especially when dealing with unstructured or heterogeneous data. Due to its ensemble nature, feature importances can be aggregated to provide some interpretability, although interactions between trees can make explanations less transparent than with simpler models [19].
- **XGBoost** is a gradient boosting algorithm that sequentially trains shallow decision trees to minimise a specified loss function [29]. Known for its predictive power and computational efficiency, XGBoost includes regularisation mechanisms that prevent overfitting and supports sparsity-aware learning. Unlike traditional models, XGBoost handles categorical data internally and supports missing values during training. It offers nuanced control over model complexity through hyperparameters such as rate, tree depth, and regularisation weights, making it highly adaptable but also complex to tune. Feature attribution methods such as SHAP are particularly effective with XGBoost due to its additive structure [74].

4.5.2 Hyperparameter tuning

Hyperparameter tuning played a crucial role in obtaining reliable and high-performing classifiers, particularly when comparing models across different feature representations [106]. Traditional grid or random search could have been prohibitively expensive and prone to missing promising regions of the parameter space [17]. Instead, SHIELD employed a Bayesian optimisation framework, specifically a Gaussian process-based surrogate model with an acquisition function of expected-improvement, to efficiently explore each learner's hyperparameters [106]. This approach iteratively proposed new configurations that balanced exploration of uncertain regions against exploitation of areas known to yield high cross-validated performance.

The search space for each learner is defined in Table 4.5, with priors chosen to reflect orders of magnitude (e.g. log-uniform for regularisation and learning rate parameters) or categorical choices for architectural decisions (e.g. MLP hidden layer sizes). The next step was then to perform stratified five-fold cross-validation on the training set, optimising the mean ROC-AUC over its validation set. Each Bayesian search was iterated 30 times, which empirical studies have shown to be sufficient for convergence in comparable settings [18].

Once the surrogate model identified the best hyperparameter combination, the corresponding pipeline was refitted on the entirety of the training data. The tuned models were then evaluated on an independent test set, untouched during both training and tuning, to yield unbiased estimates of generalised performance.

Symbol	Name	Range (Type)	Description
		Feature Grouping Stage	
K	Number of groups	[2, 10] (integer)	Number of feature groups to form
α, β	Bicriterion weights	[0, 1] (float)	Trade-off between diversity and dispersion
w_2, w_3, w_4	Moment weights	[0, 1] (float)	Weights for variance, skewness, kurtosis
ϵ	Smoothing constant	$[10^{-10}, 10^{-6}]$ (float)	Prevents division by zero in CMI normalisation
		Model Training Stage	
C	Regularisation (LR, SVM)	[0.01, 100] (float)	Inverse of regularisation strength
γ	Kernel coefficient (SVM)	[0.001, 1] (float)	Affects RBF kernel spread
n_estimators	Number of trees	[50, 500] (integer)	Used in Random Forest and XGBoost
max_depth	Maximum tree depth	[3, 15] (integer)	Controls model complexity in tree models
learning_rate	Learning rate (XGB, MLP)	[0.001, 0.3] (float)	Step size shrinkage
hidden_layers	Hidden layer sizes (MLP)	Varies (tuple)	Defines structure of MLP
alpha	L2 penalty (MLP)	[0.0001, 0.1] (float)	Regularisation term for MLP
k	k in k-fold CV	[3, 10] (integer)	Number of folds in cross-validation

Table 4.5: Hyperparameters used in feature grouping and model training.

4.6 Evaluation

The evaluation strategy in this thesis was designed to provide a holistic understanding of how dissimilarity-based grouping affects ML models. It extends beyond predictive performance to also encompass explainability and fairness, acknowledging that robust accuracy alone is insufficient in high-stakes settings. First, Subsection 4.6.1 evaluates predictive capacity using accuracy, precision, recall and F1-score, with particular attention to imbalanced clinical datasets. Second, Subsection 4.6.2 assesses model SHAP-based explanations, ensuring that feature-level attributions remain equitable and transparent through dissimilar grouping and latent-space transformations. Finally, Subsection 4.6.3 audits fairness from multiple perspectives, including group parity, statistical normalisation, and explanation-level analyses. Together, these layers of evaluation provide a principled framework to balance predictive performance, explainability, and fairness, ensuring that predictive improvements are not achieved at the expense of ethical and clinical integrity.

4.6.1 Performance metrics

Robust evaluation of model performance was essential to ensure that fairness-enhancing transformations did not come at the cost of unacceptable degradation in predictive accuracy. 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. Each metric was reported for baseline, raw, and dissimilarity-grouped configurations to highlight trade-offs introduced by the grouping strategies.

Formally, let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Then, accuracy was defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4.16}$$

While accuracy captured overall correctness, it could have obscured poor performance on the minority class for imbalanced dataset [49]. To address this, the F1-score was used as a harmonic mean of precision and recall:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall (sensitivity)} = \frac{TP}{TP + FN}, \tag{4.17}$$

$$\text{Specificity} = \frac{TN}{TN + FP}, \quad \text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4.18}$$

Precision is relevant in medical contexts where false positives can trigger unnecessary follow-up tests, procedures, costs and anxiety [56, 94]. Recall (sensitivity) ensures that true cases are not overlooked, which is often the primary concern in clinical screening and triage [56, 102]. Specificity was another relevant metric that quantifies the model's ability to correctly rule out false cases, but was not explicilty used in this thesis, since it can be directly derived from the other metrics [56]. That is, it can be computed with accuracy and recall given the class prevalence π (proportion of positive cases) [93], which is available as can be seen in Tables 4.1 and 4.2:

$$\text{Accuracy} = \pi \cdot \text{Recall} + (1 - \pi) \cdot \text{Specificity} \implies \tag{4.19}$$

$$\text{Specificity} = \frac{\text{Accuracy} - \pi \cdot \text{Recall}}{1 - \pi}. \tag{4.20}$$

It is worth noting that while ROC-AUC was employed as the optimisation target during hyperparameter tuning (see Section 4.5.2), it was not included among the primary test metrics. The rationale is that it is less intuitive compared to the above metrics, which makes it less interpretable to end users (clinicians and patients). It also averages over thresholds, which can obscure disparities and overstate performance in imbalanced datasets. In this respect, the threshold tuning was deliberately not applied in this study. Adjusting the classification threshold can optimise sensitivity or specificity depending on stakeholder priorities, but it complicates fairness comparisons, as group-specific thresholds can artificially inflate equity metrics while reducing transparency [102]. Instead, the standard threshold of 0.5 for probabilistic models was retained to ensure comparability across all experimental settings. This choice was consistent with the study's emphasis on fairness mechanisms at the representation and grouping level, rather than outcome post-processing.

4.6.2 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 each input feature's contribution to model predictions. Given an input x and model f (see Equation (2.1)), the SHAP value for feature i is formally defined as:

$$\phi_i = \sum_{S \subseteq F \setminus i} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup i}(x) - f_S(x)], \tag{4.21}$$

where F denotes the full set of features, and S denotes a subset not containing i. This formulation captures the marginal contribution of feature i, averaged across all possible coalitions, thereby ensuring the axiomatic properties of consistency, local accuracy, and fairness across correlated variables [74].

In the ungrouped configuration, SHAP values were computed directly for each original feature. 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 (4.11). These latent attributions were then decomposed to feature-level importances using the decoder weight matrix $W_{\text{dec}}^{(k)}$, yielding Equation (4.14), as described in Subsection 4.4.6. This proportional redistribution ensures that the contribution of each latent factor is faithfully allocated across its constituent features. Importantly, this procedure preserves the auditability of feature-level attributions even after dimensionality reduction, thereby addressing a critical limitation of traditional post-hoc explanations that often become uninterpretable once features are aggregated [1, 127].

The integration of SHAP into SHIELD thus has a dual role. On one hand, it provides faithful explanations of model predictions at both latent and feature levels, enabling clinicians to scrutinise individual decisions. On the other hand, 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 features 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.

4.6.3 Fairness Metrics

Ensuring algorithmic fairness in clinical machine learning goes beyond verifying that a model's predictions are accurate. It requires a principled understanding of how different sources of bias can arise and propagate through the modelling pipeline. Motivated by this, the fairness evaluation framework adopted multiple complementary perspectives that systematically uncovered both direct prediction disparities and deeper structural explanations of unfairness. This aligned with recent literature calling for fairness audits that addressed not only outcomes but also how models internally justified those outcomes [54, 60, 126, 38].

Group Fairness: Equal Opportunity and Equalised Odds

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 as formalised by Hardt *et al.* [54].

Equal Opportunity requires that true positive rates (TPRs) be equal across groups:

$$\Pr(\hat{Y} = 1 \mid Y = 1, A = 1) = \Pr(\hat{Y} = 1 \mid Y = 1, A = 0). \tag{4.22}$$

This ensured that individuals in all groups had 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 strengthens this by also requiring equal false positive rates (FPRs):

$$\Pr(\hat{Y} = 1 \mid Y = 0, A = 1) = \Pr(\hat{Y} = 1 \mid Y = 0, A = 0). \tag{4.23}$$

However, it was noted that Equalised Odds can 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 [11]. This design choice demonstrated a balance between fairness and respecting the clinical integrity of ground truth labels.

Statistical Normalisation: The N-Sigma Index

While group fairness metrics expose mean differences between groups, they do not account for uncertainty due to small sample sizes or high variance in subgroup distributions. As Chong *et al.* argue, fairness improvements that appear large in percentage terms can be statistically insignificant when sample sizes are small [38]. N-Sigma index was computed to safeguard against overinterpreting noisy fairness estimates:

$$N\text{-}\sigma = \frac{|\epsilon_1 - \epsilon_0|}{\sqrt{\frac{\sigma_1^2 + \sigma_0^2}{2}}}, \tag{4.24}$$

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 is an important step when working with health records, where minority group sizes could be limited in real-world hospital datasets.

Explanation-Level Bias: Bias Quadrant Analysis

Prediction parity alone did not guarantee that a model's internal reasoning is fair. As Jain *et al.* highlight, models that appeared fair at the prediction level could still produce biased explanations, which undermine trust in contexts where interpretability was critical, such as patient-specific risk scores or feature-driven diagnostic rules [60].

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

$$B_j^{\text{exp}} = \left| \mathbb{E}[\phi_j \mid A = 1] - \mathbb{E}[\phi_j \mid A = 0] \right|, \tag{4.25}$$

where ϕ_j denotes the Shapley value for feature j. 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 protected features 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 features. 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 feature 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.

This dual perspective clarified that proxy variables and latent representations can still amplify bias, even when fairness constraints are applied solely at the prediction level.

Integrated View: A Balanced Fairness Objective

No single fairness test was sufficient in isolation. Prediction-level parity alone could mask hidden explanation bias, explanation-level auditing alone may have ignored structural error. By combining group fairness (EO, EOdds), statistical normalisation (N-Sigma), local explanation bias, and structural error decomposition, the fairness framework aimed to expose different pathways through which unfairness could arise and persist.

The final composite fairness score was therefore expressed as

$$\text{Fairness Overview} = \gamma \cdot \text{Group Parity} + (1-\gamma) \cdot \text{Explanation Parity} \tag{4.26}$$

$$= \gamma \cdot \left(\frac{\text{Equal Opportunity} + N\text{-}\sigma}{2} \right) + (1-\gamma) \cdot (\text{Bias Quadrant}). \tag{4.27}$$

These components were balanced with tunable weights γ that reflected stakeholder priorities, whether equal opportunity was paramount, or explanation consistency was critical. This overview score ensured that SHIELD preserves fairness across various perspectives that complement one another.

Chapter 5

Results and Discussion

5.1 Data Preprocessing

The experimental adjustment in data preprocessing explained in Section 4.2 instigated the following results. Table 5.1 indicates how one-hot encoding led to an increase in feature dimensionality for datasets with multiple categorical attributes. For instance, the Obesity dataset expanded from 16 raw variables to 25 encoded features, while Diabetes retained 47 but with significantly more sparse encoding internally. This expansion was relevant because it affected both computational complexity and the grouping process, as groups were constructed in the encoded feature space.

Dataset	**Original Rows**	**Final Rows**	**Retention**	**Feature Count Change**
Breast Cancer	569	569	100.0%	$30 \rightarrow 30$
Heart Disease	303	303	100.0%	$13 \rightarrow 13$
Obesity	2,111	2,111	100.0%	$16 \rightarrow 25$
Diabetes	101,766	98,053	96.35%	$47 \rightarrow 47$

Table 5.1: Post-preprocessing dataset overview, including feature expansion due to one-hot encoding.

Table 5.2 reveals distinct class-balance characteristics across datasets after preprocessing. The Obesity dataset maintained a balanced distribution across seven BMI categories by design, whereas the remaining datasets exhibited notable class imbalance. Such imbalance is a well-documented phenomenon in healthcare data and often arises naturally because the general population is predominantly composed of individuals without the target condition. In real-world epidemiological contexts, disease prevalence is typically low, which leads to a healthy class dominating the dataset. However, this trend can reverse or become less pronounced in clinical or hospital-based datasets. This is because data collection in these settings is conditional on healthcare-seeking behaviour, meaning participants are more likely to present with symptoms that prompt a medical visit. Consequently, the observed distributions highlight the significance of contextualising imbalance in healthcare datasets, where they often exhibit a conditional sampling bias.

This bias underscores the importance of carefully interpreting class distribution in predictive modelling, as the level of imbalance is influenced not only by disease prevalence but also by the context and setting of data acquisition.

Dataset	Split	# Samples	Class Distribution (%)
Breast Cancer	Train	455	Benign: 62.6, Malignant: 37.4
	Test	114	Benign: 64.0, Malignant: 36.0
Heart Disease	Train	242	No Disease: 54.5, Disease: 45.5
	Test	61	No Disease: 55.7, Disease: 44.3
Obesity	Train	1,689	Classes balanced across 7 BMI categories
	Test	422	Same proportion as training
Diabetes	Train	77,700	Class 0: 11.3, Class 1: 35.3, Class 2: 53.4
	Test	20,353	Class 0: 11.2, Class 1: 35.5, Class 2: 53.3

Table 5.2: Train-test split and class distribution for all datasets after preprocessing.

Clearly, the Diabetes dataset posed the greatest challenge in terms of its size and missing data. While most features had negligible missingness, several key attributes exhibited extreme sparsity, making naive imputation infeasible and risking bias propagation or data leakage. Table 5.3 summarises these features, their missingness rates, and the corrective actions taken.

Feature	Not Missing	Missing	% Missing	Action Taken
`weight`	3,197	98,569	96.9%	Feature removed
`max_glu_serum`	5,346	96,420	94.7%	Imputed (XGBoost)
`A1Cresult`	17,018	84,748	83.3%	Imputed (XGBoost)
`medical_specialty`	51,817	49,949	49.1%	Imputed (XGBoost)
`payer_code`	61,510	40,256	39.5%	Imputed (XGBoost)
`race`	99,493	2,273	2.23%	Dropped rows with NA
`diag_3`	100,343	1,423	1.40%	Dropped rows with NA
`diag_2`	101,408	358	0.35%	Dropped rows with NA
`diag_1`	101,745	21	$< 0.1\%$	Dropped rows with NA

Table 5.3: Summary of features with severe missingness in Diabetes dataset and preprocessing decisions.

These decisions were guided by three considerations: (i) extremely sparse feature, namely

`weight`, was removed to prevent injecting noise through imputation; (ii) features with moderate to high missingness but with clinical relevance (e.g. `A1Cresult`, `payer_code`) were imputed using an XGBoost classifier to preserve an ability to predict without discarding large portions of the dataset; (iii) low-missing features (`diag_1-diag_3`) were handled by row-wise removal to maintain simplicity and avoid introducing imputation bias. Consequently, the retention ended up being incomplete, but still high (96.35%) despite discarding severely incomplete records.

5.2 Feature grouping

This section focuses on empirical results of the grouping step. First, the CMI-derived dissimilarity structure of the feature graph is characterised with a heatmap (see Figure 5.1). Second, the effect of varying the number of groups K is examined by plotting the two anticlustering criteria defined in Subsection 4.4.1, illustrated as Figure 5.2. Finally, the immediate implications of these patterns for the grouping step are summarised in Subsection 5.2.3, with discussion confined to how the induced partitions constrain correlated signals within the feature space.

5.2.1 CMI-based pairwise dissimilarity

Across datasets, the CMI-based pairwise dissimilarity matrices consistently exhibited clinically coherent structure. Figure 5.1 shows the normalised dissimilarity matrix for the Heart Disease dataset. Several patterns aligned with the domain knowledge. Pairs with lower dissimilarity score, such as `age-trestbps` (`trestbps` stands for resting blood pressure) and `age-chol` (`chol` stands for serum cholestoral) exhibited $d \approx 0.62$, which is clinically plausible given that age influences both blood pressure and lipid levels. Meanwhile, those with higher scores (weaker relation), such as `age-sex` with $d \approx 0.82$, reflect minimal shared information once conditioned on disease status. These relationships corroborate that CMI captured clinically relevant dependencies while revealing proxy risks, such as age being highly connected to multiple physiological measures.

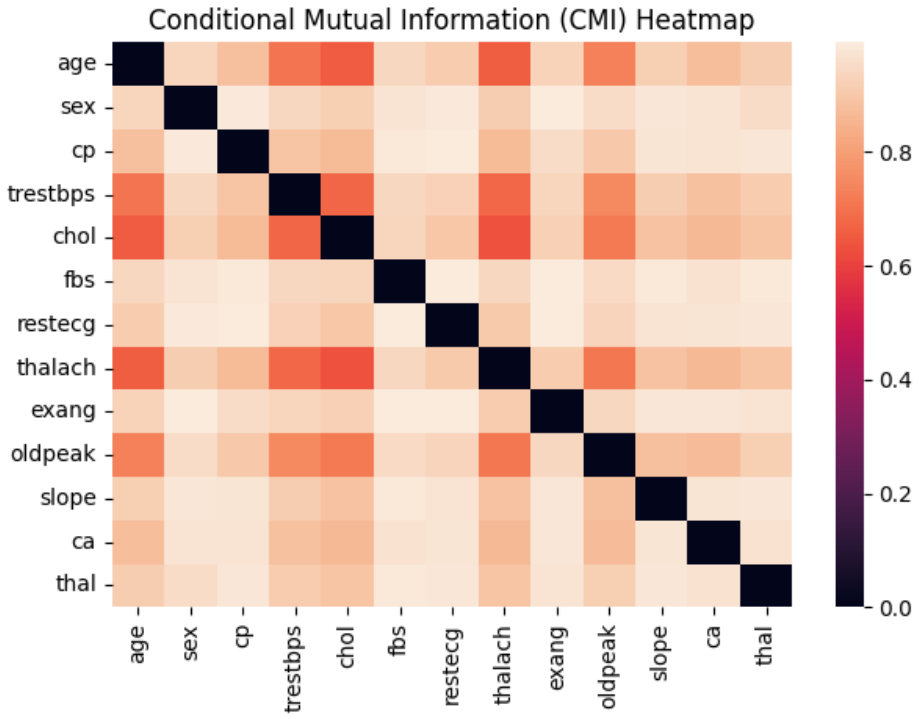
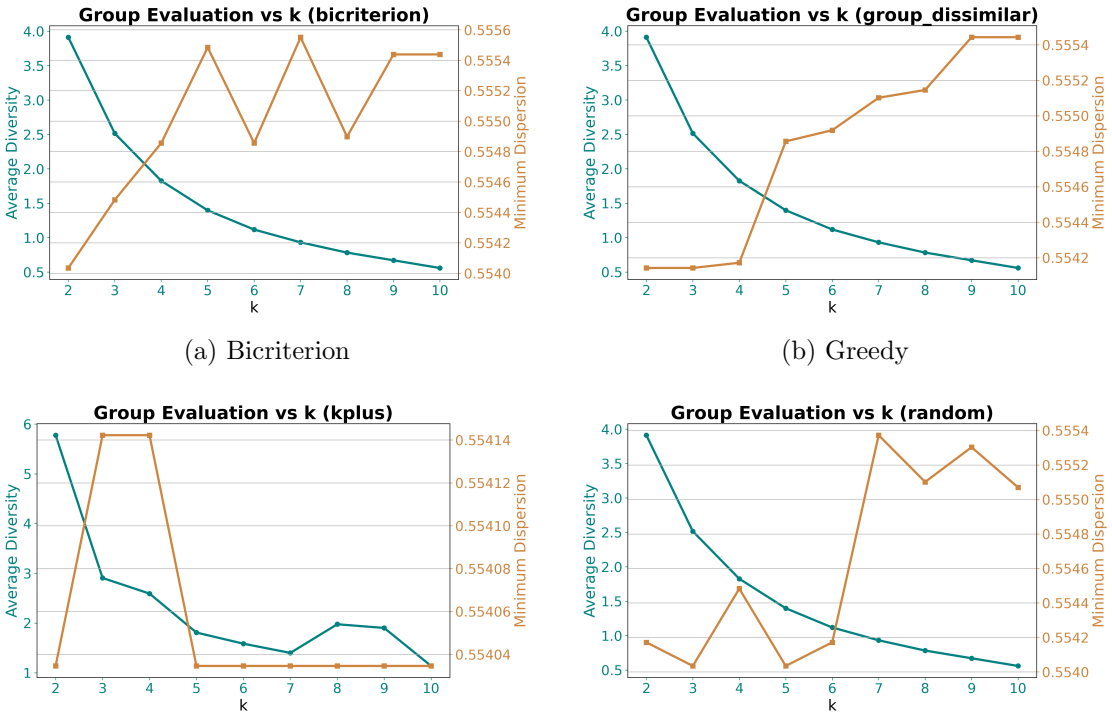


Figure 5.1: CMI-based dissimilarity matrix for Heart Disease dataset. Lower values indicate stronger dependency.

5.2.2 Trade-off between diversity and dispersion

The principal empirical finding was that increasing the number of groups K almost monotonically reduced diversity. Figure 5.2 illustrates the relationship between average diversity and minimum dispersion as K varies from 2 to 10. As can be seen, diversity, which measures the mean pairwise dissimilarity between groups, consistently declined as K increased because forming more groups reduced opportunities for inter-group separation. The rate of decline differed similarly across methods. K-plus maintained the highest diversity throughout, starting near 5.8 at $K = 2$ and remaining above 1.0 even at $K = 10$, whereas all the other methods started at around 3.9 and dropped below 0.6 by $K = 10$. This scale difference reflects K-plus' design objective of maximising global centroid separation without considering worst-case margins.



(a) Bicriterion

(b) Greedy

(c) K-plus anticlustering

(d) Random

Figure 5.2: Average diversity (left axis) and minimum dispersion (right axis) versus the number of groups k on the Breast Cancer CMI graph for four grouping strategies. Higher is better for both criteria. The plots are arranged side by side rather than stacked vertically to enable simultaneous comparison of all cases at once. The same rationale applies to Figures 5.3, 5.6, 5.7, 5.8, 5.9, 5.11, 5.12, 5.13, 5.14 and 5.15. Some plots are also enlarged to exceed the default margin for better readability.

Dispersion values, in contrast, were tightly clustered across methods and K values, remaining within the narrow band of approximately 0.554 to 0.556. For instance, Bicriterion peaked at 0.5556 around $K = 7$, and Greedy reached 0.5554 at $k = 9$, but these increments were negligible relative to the concurrent diversity loss. K-plus exhibited an almost flat dispersion curve near 0.5541 because its anti-clustering algorithm does not explicitly address minimum pairwise distances. Instead, it concentrated on maximising overall group spread. Consequently, improvements in dispersion at higher K offered minimal practical benefit and were not to drive the choice of K.

Sophisticated methods such as Bicriterion and K-plus achieved favourable trade-offs at smaller K values. At $K = 3 \sim 5$, for example, Bicriterion retained a diversity of approximately 2.1 while reaching dispersion close to its upper bound (0.5555). K-plus also remained exceptionally strong at low K, with diversity above 4.0 and dispersion

stable at 0.554. In contrast, Greedy and Random required $K \geq 8$ to approach their best dispersion, by which point diversity had diminished by over 85%. These trends confirmed that advanced strategies could cluster features into sufficiently dissimilar groups early, avoiding the need for excessive partitioning.

From a practical standpoint, the marginal gains in dispersion beyond $K = 5$ did not compensate for the steep diversity loss observed across all methods. This justified the choice of $K = 4$ as the default configuration throughout the experiments including other datasets, which exhibited the same trend, as it balanced strong diversity with near-maximal dispersion, supporting both effective separation and computational efficiency.

5.2.3 Implications of grouping

The results above corroborate that grouping features based on dissimilarity effectively reduces the influence of proxy variables. In healthcare datasets, some features, such as `age`, exhibit strong correlations with multiple clinical indicators like cholesterol and blood pressure. When these variables appear together in an ungrouped model, their combined effect can disproportionately shape predictions, creating hidden pathways for sensitive attributes to leak into the decision-making process. By forcing such features into separate groups, dissimilarity-driven grouping disrupts these correlations at the latent representation level. As a result, no single latent factor can fully reconstruct the protected information, mitigating risks of indirect discrimination. As already noted, the CMI heatmap in Figure 5.1 supports this rationale by showing that `age-trestbps` and `age-chol` pairs exhibited relatively low dissimilarity scores (approximately 0.62), while weakly related pairs like `sex-cp` approached 0.91. These observations confirm that grouping prioritises clinically meaningful independence.

Feature grouping promotes a more balanced usage of the available predictors, which directly supports equitable learning. Without grouping, models frequently over-rely on a few dominant variables, leaving many others underutilised or completely inactive, as evidenced by steep importance drop-offs in SHAP plots (see Figure 5.6). Grouped representations flatten this distribution by ensuring each latent factor combines signals from multiple, diverse features. This reduces the dominance of any single attribute, particularly those correlated with sensitive characteristics. This also distributes predictive responsibility more evenly across the feature set. The SHAP analysis confirms that grouped configurations reduce the prevalence of zero-attribution features and increase participation of mid-importance variables, aligning with fairness objectives by limiting systemic biases encoded in individual predictors.

Beyond fairness benefits, grouping introduces practical advantages for computational efficiency when operating on grouped representations instead of the full feature set. Reducing dimensionality from dozens of raw variables to a handful of groups simplifies the complexity of model fitting, especially for algorithms sensitive to feature dimensionality such as logistic regression, SVM, and MLP. Although grouping incurs additional cost

for encoding and decoding latent representations, this overhead is minimal compared to the gains during repeated training and hyperparameter optimisation. For example, compressing the Obesity dataset from 25 encoded features to $K = 4$ groups significantly reduced input dimensionality without incurring major accuracy penalties (see Section 5.4.1). This dimensionality reduction shortens training times, lowers memory usage, and decreases the risk of overfitting, making grouped pipelines more scalable for high-dimensional healthcare applications.

5.3 Tuned hyperparameters

As listed in Table 5.4, sophisticated partitions (Bicriterion, K-plus) produced hyperparameters that closely tracked the ungrouped optimum, whereas naive partitions (Random, Greedy) forced larger, compensatory deviations. The table summarises the tuned hyperparameters for all models on the Obesity dataset across grouping strategies. Similar patterns were observed in the other datasets, but Obesity was chosen to be a representative case for closer analysis.

A key theme is the contrast between grouped and ungrouped configurations. Sophisticated grouping strategies such as Bicriterion and K-plus often produced hyperparameter values close to those of the ungrouped baseline, while Random and Greedy diverged more strongly. This convergence suggests that principled grouping can preserve much of the inductive bias of the ungrouped feature space, whereas ad hoc partitions disrupt signal structure, forcing models to compensate with more extreme tuning.

For linear models, both Logistic Regression and SVM showed substantial variability in the regularisation parameter C. Values spanned two orders of magnitude, from as low as $C = 0.82$ (random) to the maximum allowed $C = 100$ (ungrouped and K-plus SVM). The fact that optimal C was often at the boundary of the search range highlights two important points: (i) the search space may not have fully captured the true optimum, and (ii) grouping could have fundamentally altered how much regularisation the model required. Random grouping, for instance, needed much stronger penalisation to prevent overfitting on disrupted feature signals, while Bicriterion and K-plus aligned more closely with the ungrouped settings, implying that their structured partitions better preserved useful information.

Neural models (MLP) were far less sensitive to grouping. The hidden layer size consistently tuned to the smallest option (50 units) across all groupings, and both α and the learning rate remained stable at their lower search bounds (0.0001 and 0.001). This consistency suggests that MLPs internally absorbed feature redundancy and correlation, making them less reliant on hyperparameter adjustments. This is corroborated in Figures 5.4 and 5.5, where the MLP model yields the most consistent performance across

different grouping methods. However, the repeated selection of boundary values indicates that the model's capacity could potentially be further optimised if larger hidden sizes or different learning rates were explored.

Tree-based models illustrate a different dynamic. Random Forest consistently favoured moderately large ensembles (178-248 trees) and shallow depths (8-15), with the `sqrt` setting for maximum features selected in every case. This stability across grouping strategies demonstrates the robustness of bagging and feature subsampling in counteracting grouping perturbations. K-plus and Bicriterion again picked the closer (higher) value to that of Ungrouped compared to other methods. Furthermore, XGBoost exhibited systematic shifts, where the number of estimators reached the maximum (500) and depths were also near the upper bound (14-15) in ungrouped, K-plus, and Bicriterion settings. This boundary-hitting behaviour shows that gradient-boosted trees increasingly demanded complexity when feature groups were introduced, particularly with structured partitions, whereas Random grouping converged on fewer estimators (278) and a higher learning rate ($\eta = 0.025$). Such divergence reveals a trade-off between shallow, high-rate learners that quickly adapt to noisy groupings and deeper, slower learners that exploit structured signals.

Hyperparameter	**Ungrouped**	**Random**	**Greedy**	**K-plus**	**Bicriterion**
LR C	2.98	15.1	10.7	8.38	9.90
SVM C	100	0.82	9.56	100	47.2
SVM γ	0.021	0.0012	0.327	0.028	0.020
MLP hidden layer size	50	50	50	50	50
MLP α	0.0001	0.0001	0.0001	0.0001	0.0001
MLP learning rate	0.001	0.001	0.001	0.001	0.001
RF n_estimators	215	182	178	248	198
RF max depth	10	9	8	15	9
RF max features	sqrt	sqrt	sqrt	sqrt	sqrt
XGB n_estimators	500	278	281	500	500
XGB max depth	15	8	9	15	14
XGB learning rate (η)	0.001	0.025	0.012	0.00103	0.00130
XGB subsample	0.88	0.99	0.99	0.60	0.60
XGB colsample by tree	0.88	0.99	0.99	0.60	0.60

Table 5.4: Tuned hyperparameters for all models on Obesity.

5.4 Train and test results

This section exemplifies that dissimilarity-based grouping exhibited a consistent trade-off. There was a modest reductions in predictive performance relative to ungrouped features with mean decreases of 3.43% in accuracy, 3.82% in recall, 6.41% in precision, and 5.16% in F1-score (see Figure 5.3 in Subsection 5.4.1). In return, SHAP analysis demonstrated the equitable distribution of feature contribution as well as more productive use of instances when grouped. Furthermore, grouping led to systematic overview fairness gains by 2.42% on average and remarkable 9.47% improvement on average distance from origin of the bias quadrant (see Figure 5.15 in Subsection 5.4.2). Benefits tended to be more visible in smaller datasets where individual instances exerted more influence, reflecting the significance of broader distribution of feature attributions. This section concludes by discussing the practical implications of the results for clinical deployment (see Subsection 5.4.3), such as improvement of sample efficiency and reduction of participant burden.

5.4.1 Performance metrics

Across all datasets and grouping methods, accuracy consistently scored the highest among the four metrics. This pattern is often expected in medical classification problems where class imbalance or conservative decision thresholds may inflate overall correctness without necessarily optimising sensitivity to the positive class [56]. While high accuracy suggests robust general classification performance, the relatively lower F1-scores and recalls in certain datasets (particularly Heart Disease and Diabetes) reveal that positive class detection may still be challenging. In the medical context, this trade-off is critical. A model that maintains high accuracy but underperforms in recall risks failing to identify patients with the condition, which can have serious clinical consequences.

Before examining the variability of predictive performance across grouping methods in Figure 5.3, it is worth highlighting some dataset-specific patterns that emerged from the aggregated results. The Breast Cancer dataset exhibited near-ceiling performance across all metrics and grouping methods, indicating that the classification problem was relatively well-posed and robust to feature grouping. The Obesity dataset similarly showed high accuracy and balanced F1-score, precision, and recall, despite having the most number of classes (7), suggesting stable model behaviour across groupings. In contrast, the Diabetes dataset demonstrated greater metric variability, particularly in recall, where some grouping strategies suffered from notable performance drops. The Heart Disease dataset also stood out as the most challenging, since not only were average scores lower across all metrics, but the gap between accuracy and recall was wider, signalling potential limitations in capturing true positive cases. Across all datasets, the choice of grouping method appeared to have a smaller impact than the intrinsic nature of the dataset, as metrics for each dataset remained relatively consistent regardless of grouping strategy.

Quantitatively, when comparing grouped methods against the ungrouped baseline, it was observed that accuracy, recall, F1-score, and precision decreased on average by 3.43%, 3.82%, 5.16%, and 6.41% respectively. This consistent downward shift indicates a modest but systematic performance trade-off associated with grouping. The largest observed performance gap was in the case of Obesity with the K-plus grouping method for recall, which showed a 20.14% decrease relative to the ungrouped configuration. These results suggest that while grouping may offer computational or equitability benefits, it tends to introduce small yet measurable reductions in predictive performance, with the magnitude of this reduction varying across datasets and metrics.

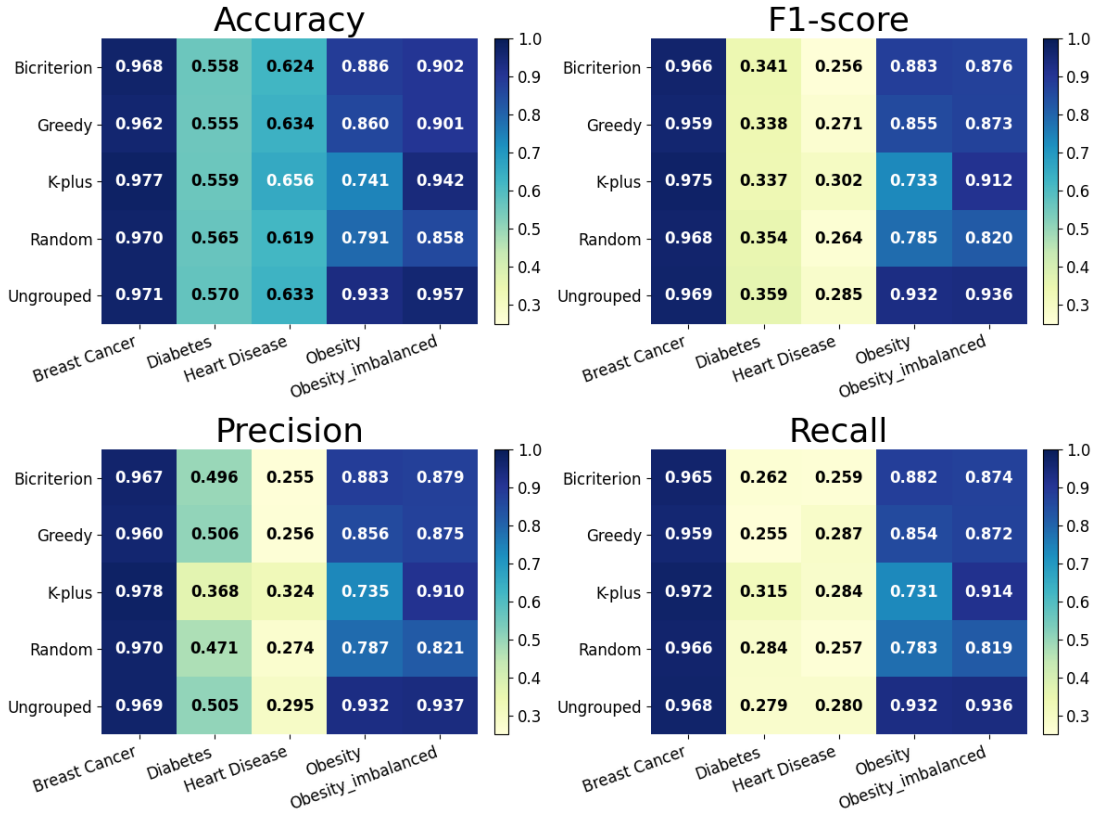


Figure 5.3: Mean accuracy, F1-score, precision, and recall of models across datasets and grouping methods.

While the overall trend indicates that grouped methods tended to underperform the ungrouped baseline, an important exception arose with the K-plus strategy in more challenging datasets. Notably, K-plus outperformed the ungrouped configuration across all four metrics for Heart Disease, with gains of up to 9.87% in precision and 5.89% in F1-score. This suggests that extreme (compared to bicriterion) feature grouping could enhance signal extraction and model discrimination in scenarios where feature redundancy or noise hampered learning. Interestingly, this advantage was most pronounced in

imbalanced datasets (all but Obeisty), which were generally harder due to skewed class distributions and the difficulty of detecting minority-class cases. In contrast, the Obesity dataset was relatively balanced across its seven classes and exhibited near-ceiling performance without grouping. Meanwhile, K-plus performed substantially worse, with drops of up to 21.57% in recall compared to the ungrouped case. This stark contrast highlights that the benefits of grouping were highly context-dependent. While they could have mitigated challenges in complex, imbalanced datasets by improving minority-class sensitivity, they may have disrupted well-established feature-class relationships in simpler, balanced problems. Consequently, grouping strategies like K-plus should be deployed selectively, informed by dataset balance and class separability.

Beyond the aggregated metric view, this experiment also conducted a more granular benchmark comparison for the Breast Cancer dataset. Figures 5.4 and 5.5 overlay the results of each grouping method on top of benchmark baseline distributions, which capture the minimum, mean, and maximum performance across multiple runs for each classifier as provided in UCI ML Repository [120]. This allows direct visual comparison of grouped approaches against the variability range of an established baseline.

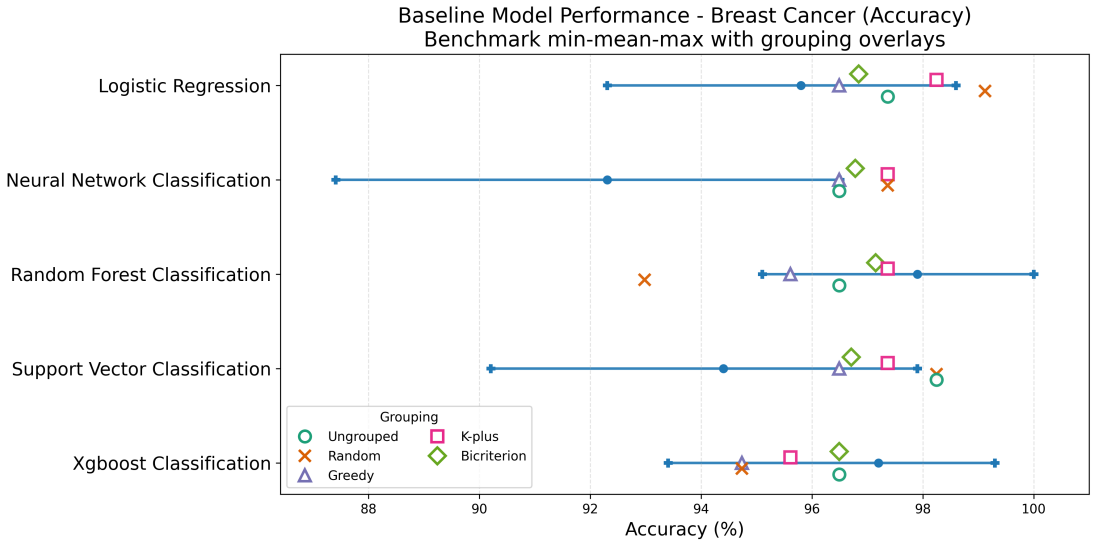


Figure 5.4: Accuracy results of grouping methods against Breast Cancer benchmark, represented by blue line and dot.

From the benchmark overlays in Figures 5.4 and 5.5, it was evident that K-plus and Bicriterion generally produced results closest to the ungrouped baseline across classifiers. Greedy consistently performed slightly worse than the ungrouped case, while Random exhibited the largest deviations, sometimes outperforming and sometimes underperforming relative to the baseline. For example, in accuracy, the mean absolute deviation from the ungrouped baseline across all classifiers was only 0.38% for K-plus and 0.41% for Bicriterion, compared to 1.13% for Greedy and 2.46% for Random. A similar pattern

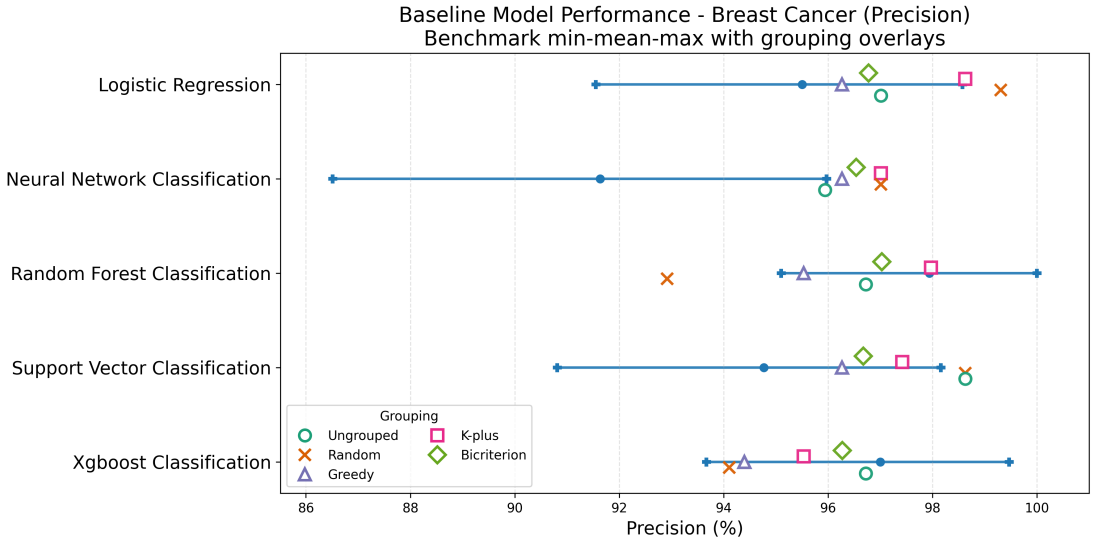


Figure 5.5: Precision results of grouping methods against Breast Cancer benchmark, represented by the blue line and dot.

was observed for precision, where K-plus and Bicriterion differed by 0.35% and 0.39% from the ungrouped mean respectively, while Greedy had a deviation of 1.02% and Random of 2.11%. Importantly, most of the grouping results lay within the benchmark variability range, indicating that performance differences were well within the expected stochastic variation. An exception was the MLP classifier, where several groupings, particularly K-plus, achieved results near the upper benchmark bound, suggesting particularly favourable interaction between grouping and neural architectures.

Random grouping underperformed the method averages of XGBoost and Random Forest by 1.4% and 3.3%, respectively. Considering their standard deviations (SDs) were only 0.9% and 1.5%, these gaps equated to 1.5 and 2.2 SDs, indicating statistically meaningful differences. This relatively weaker performance of Random grouping on tree-based models could be explained by the way these models exploited structured feature relationships. Tree-based methods depend heavily on early, high-gain splits formed by correlated or interacting features [52, 129]. Hence, random grouping may have disrupted these structures by separating mutually informative features or mixing them with low-importance variables in non-systematic or meaningless way, thereby reducing individual split gains. This could have led the model to select suboptimal splits in the upper tree levels, propagating noise into downstream nodes and lowering predictive performance. In contrast, models like MLPs or SVMs, which learn distributed decision boundaries more globally, were more resilient to the disruption of feature structure caused by random grouping.

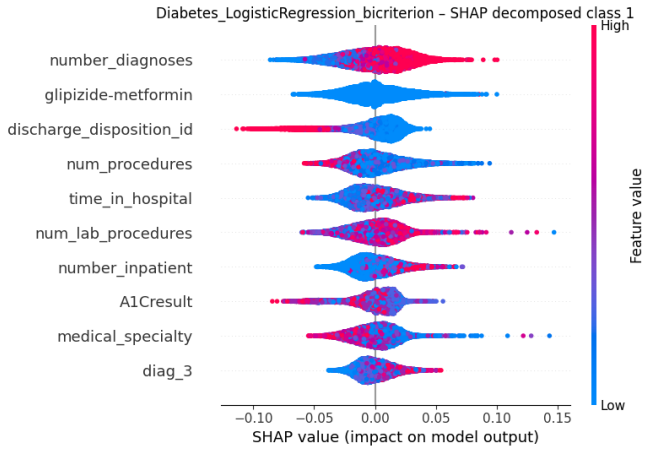
5.4.2 SHAP and fairness metrics

A central goal of this study was to assess how feature grouping affects a model's equitable learning in addition to predictive performance. On the surface level, SHAP values were used to quantify each feature's contribution to model predictions. Beyond this explainability, SHAP here served to evaluate whether grouping encourages the model to rely more evenly on the available features and instances, rather than concentrating decision-making power in a small dominant subset. This is particularly important in medical applications, where each feature often represents clinically relevant information and each patient instance is valuable in terms of availability and relevancy. With this in mind, the SHAP and fairness results of SHIELD are discussed in this subsection coherently as follows: (i) the overall impact of grouping compared to ungrouped cases, then (ii) comparison across grouping methods and ML models, followed by (iii) further examination of other insightful results such as instance-level explanations, SHAP range statistics and bias quadrant, and ending with (iv) more detailed fairness metrics comparison between grouping methods.

Grouping versus not grouping

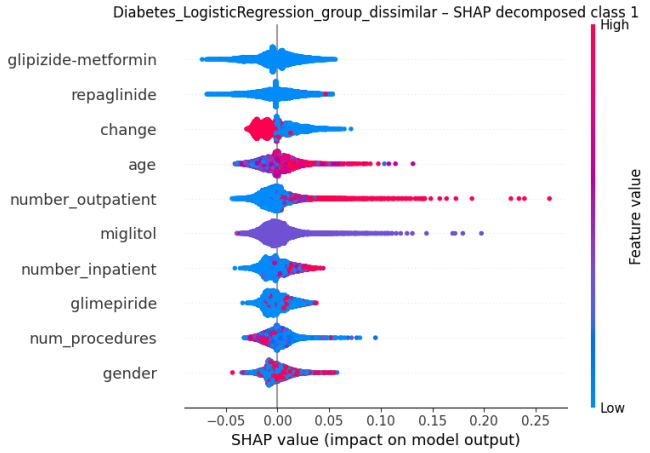
A consistent pattern across datasets was that grouped representations led to more equitable use of features and instances compared to the ungrouped baseline. As illustrated by SHAP plots including Figure 5.6, the ungrouped case was dominated by a small subset of features, producing steep drop-offs in importance and leaving many features with near-zero contribution. This concentration implies that large portions of the feature 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 features were assigned moderate levels of importance and fewer instances were associated with zero SHAP values. In practice, this means that grouped models make more use of the available data, essentially reducing 'waste'.

The more balanced reliance on features also translated into improvements in fairness metrics. For example, in Equal Opportunity, the ungrouped Obesity model was recorded with a disparity of 0.520, whereas grouping reduced this to 0.373 with K-plus and 0.232 with Bicriterion (see Figure 5.15). Similarly, in Equalised Odds, grouping lowered the disparity in Obesity from 0.520 to 0.373 and 0.214 with K-plus and Bicriterion, respectively. These reductions indicate that grouping prevented single attributes, particularly sensitive ones, from becoming disproportionately influential in determining positive outcomes. By mixing privileged and unprivileged samples within latent groups, membership of a protected attribute ceased to dictate outcomes in a deterministic manner.

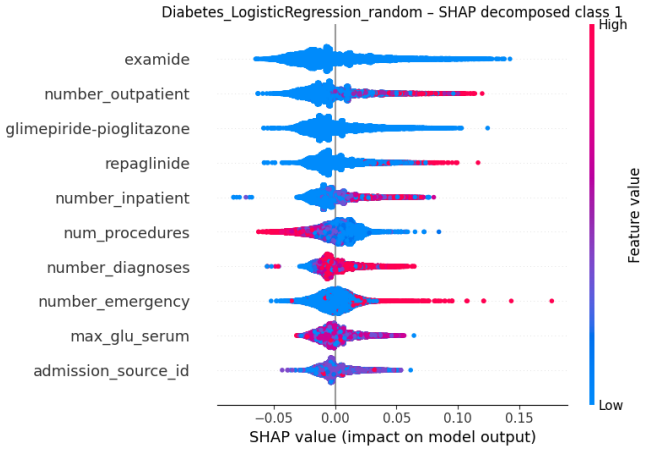


(a) Grouped via bicriterion method

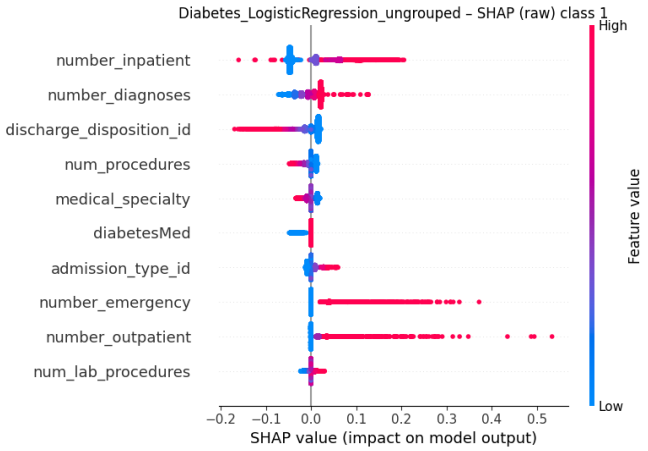
(b) Grouped via K-plus anticlustering



(c) Grouped via greedy method



(d) Grouped randomly



(e) Ungrouped

Figure 5.6: Comparison of SHAP plots for Diabetes classification by Logistic Regression between different grouping methods including ungrouped case.

The effect of grouping was especially prominent in smaller datasets where each instance carried greater weight. In Heart Disease (303 instances), grouping reduced Equal Opportunity disparity from 0.520 in the ungrouped case to 0.232 with Bicriterion. In Obesity (2111 instances), the improvement was smaller but still measurable, with Bicriterion at 0.091 compared to 0.030 for the ungrouped case. By contrast, in Diabetes (101,766 instances), grouping had little effect because the sheer volume of data already ensured more stable decision boundaries, so fairness disparities were low in both grouped and ungrouped cases. This confirms that the benefits of grouping for fairness were inversely proportional to dataset size, with the strongest gains observed when data was scarce and each observation was more influential. Similarly, its benefits diminished in very large datasets, where data abundance already regularised feature contributions.

It is also important to note the difference between the balanced and imbalanced Obesity datasets. In the synthetically imbalanced version (Obesity_imbalanced), created by oversampling minority classes with SMOTE, fairness disparities were amplified. For instance, Equal Opportunity worsened from 0.103 in balanced Obesity to 0.142 in Obesity_imbalanced under Random grouping, and Average Distance from Origin increased from 0.216 to 0.313 in the ungrouped case. Grouping, especially Bicriterion, partly mitigated this degradation, achieving 0.281 for Average Distance compared to 0.313 ungrouped. This suggests that while class imbalance inherently increased fairness risks, sophisticated grouping offered some resilience by enforcing more equitable contribution of features and instances.

Grouping method comparisons

While grouping generally improved fairness relative to the ungrouped baseline, the extent of improvement differed markedly between methods. Random grouping was the most variable: it achieved results comparable to more principled methods in some cases, but it also performed the worst in others. For example, in Heart Disease, Random achieved an Equal Opportunity disparity of 0.214, very close to Bicriterion at 0.232, and better than Greedy at 0.312. However, in the Obesity dataset, Random performed poorly with Equal Opportunity disparity of 0.142 compared to Bicriterion's 0.091 and K-plus' at 0.219. This sensitivity reflects the probabilistic nature of Random grouping: when the feature set was small (e.g. 13 features in Heart Disease), there was a non-trivial probability of forming useful combinations by chance, but as the feature space grew (47 features in Diabetes), the likelihood of randomly forming balanced groups diminished.

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 Diabetes, Greedy resulted an Equalised Odds disparity of 0.025 compared to 0.016 and 0.013 with Bicriterion and ungrouped, respectively. A similar trend was observed in Predictive Parity, where Greedy recorded 0.850 in Heart Disease, which was the worst among all methods and worse than the ungrouped baseline at 0.647. This indicates that the heuristic strategy of

locally maximising dissimilarity did not guarantee globally fairer or more balanced group structures.

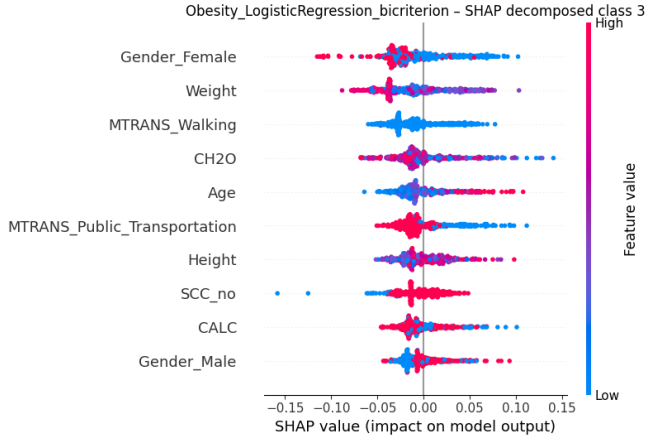
Bicriterion consistently produced the strongest and most stable fairness outcomes across datasets. By explicitly optimising both diversity (high average dissimilarity between groups) and dispersion (ensuring no group is too similar), it achieved lower disparities in almost all fairness metrics. For example, in Obesity, Equal Opportunity was reduced to 0.091 with Bicriterion compared to 0.219 with K-plus, 0.103 with Random, and 0.128 with Greedy. Across datasets, Bicriterion also produced the lowest Fairness Overview scores, such as 0.126 in Obesity compared to 0.192 with K-plus. These results confirm that an explicit optimisation of diversity and dispersion provided a systematic advantage over heuristic or random strategies.

K-plus occupied an interesting middle ground. It often performed comparably to Bicriterion in some metrics but diverged in others. For example, in the Average Distance from Origin, which measures bias magnitude in the two-dimensional fairness quadrant, K-plus achieved the best score of 0.104 in Diabetes compared to 0.136 ungrouped and 0.113 Bicriterion. However, it performed poorly in Predictive Parity, with 0.750 disparity in Heart Disease compared to 0.660 for Bicriterion and 0.647 for ungrouped. This suggests that while K-plus could have spread groups well in terms of geometric distance from fairness-neutral points, it did not guarantee balanced performance across multiple fairness criteria.

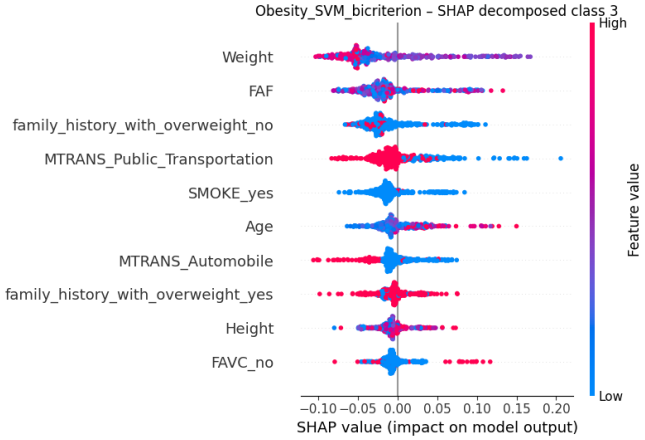
Taken together, these results indicate that Bicriterion was the most reliable method, delivering consistent fairness gains across datasets and metrics. Random could sometimes perform well but lacked robustness, while Greedy added little value relative to its computational cost, and K-plus showed promise in bias-magnitude reduction but struggled in consistency. This suggests that Bicriterion should be the preferred grouping strategy in fairness-critical applications, while K-plus may be valuable as a complementary method in contexts where geometric balance is prioritised.

Cross-model perspective

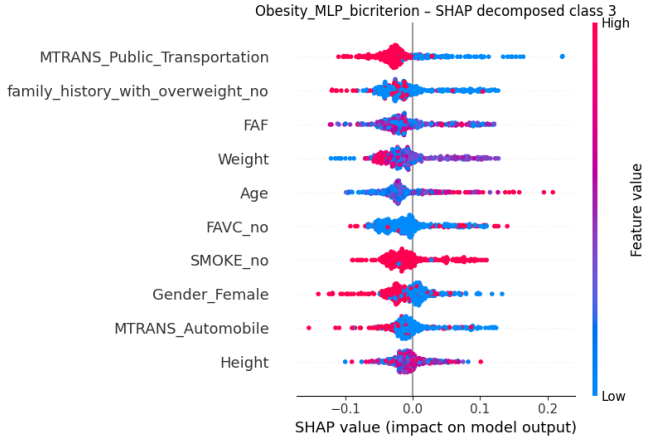
The model-wise SHAP summaries reinforce that the grouping effect was not model-specific. Under Bicriterion (Figure 5.7), linear and margin-based learners (Logistic Regression, SVM) showed visibly flatter importance spectra than their ungrouped counterparts (Figure 5.8), suggesting less reliance on a few dominant variables. It should also be noted that the most contributing features (`weight`, `height` and `age`) remained unchanged across different models for ungrouped cases, while they were not fixed for grouped cases. This implies that grouping adjusted the use of features and instances 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.



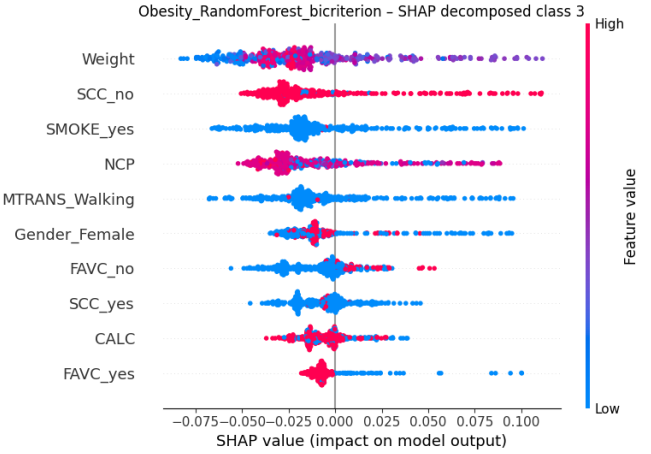
(a) Logistic regression



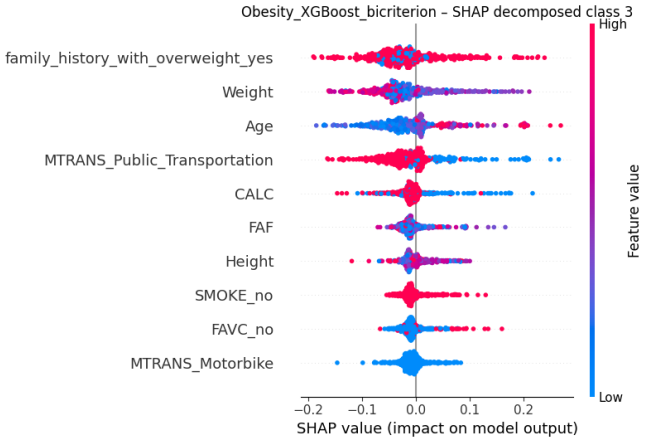
(b) SVM



(c) MLP



(d) Random Forest

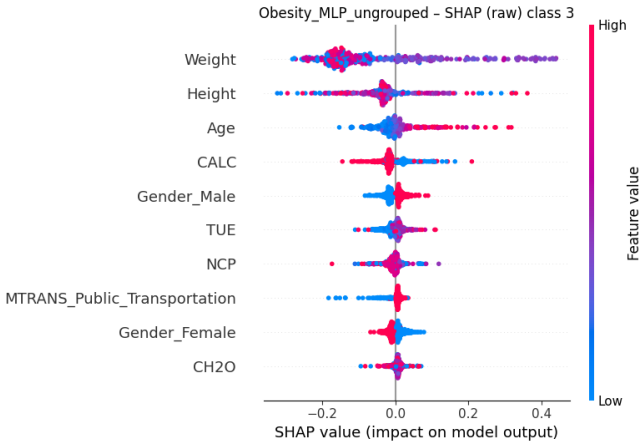


(e) XGBoost

Figure 5.7: Comparison of SHAP plots from Overweight I classification with bicriterion grouping across different models.

(a) Logistic regression

(b) SVM



(c) MLP

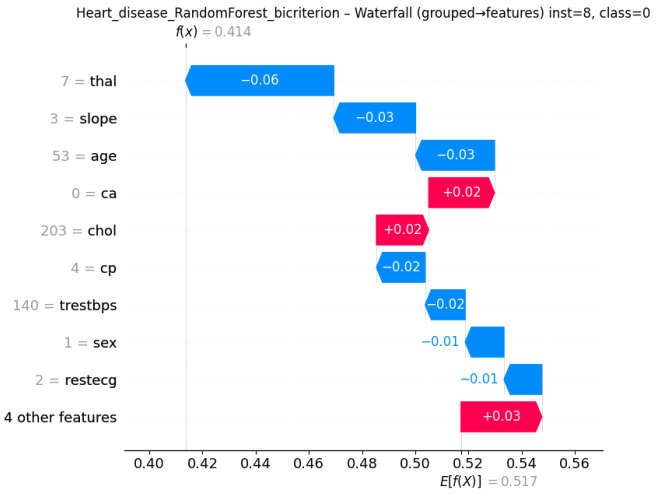
(d) Random Forest

(e) XGBoost

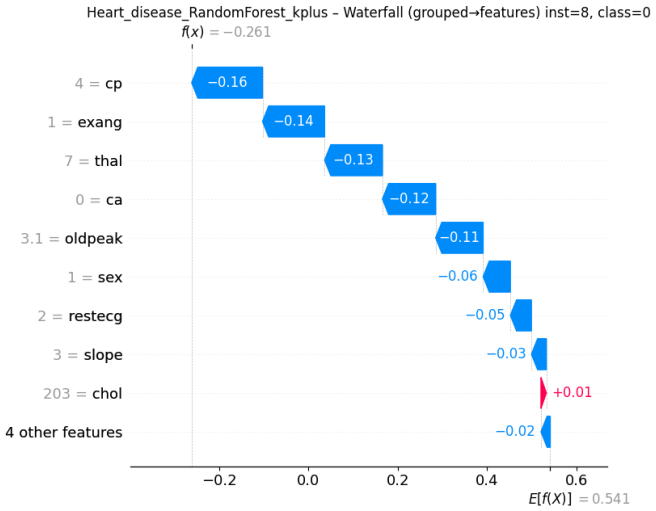
Figure 5.8: Comparison of SHAP plots from ungrouped Overweight I classification across different models.

Per-instance explanations

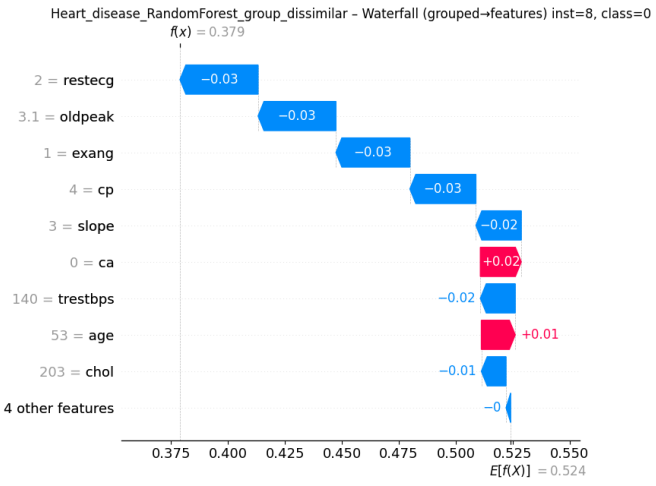
Complementing the global SHAP summaries, the waterfalls in Figure 5.9 provide an instance-level view of how grouping redistributed contributions. For the Heart disease task with Random Forest, grouped variants typically exhibited shorter extreme bars compared to the ungrouped baseline, indicating reduced dominance by a small set of correlated features. Bicriterion and K-plus, in particular, showed smoother step-down profiles from the base value to the final logit/probability, consistent with the use of more balanced features observed in the global SHAP plots. These per-instance patterns mirror the aggregated results in Figure 5.6, where grouping flattened the importance distribution, and they foreshadowed the reductions in group disparity visible in the fairness heatmap (Figure 5.15).



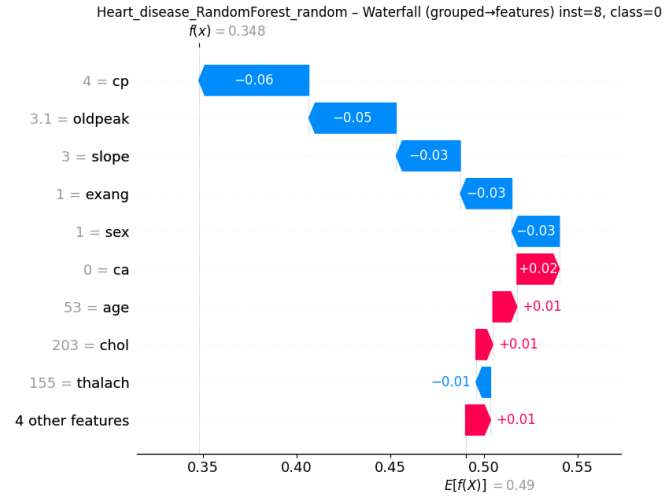
(a) Bicriterion method



(b) K-plus anticlustering

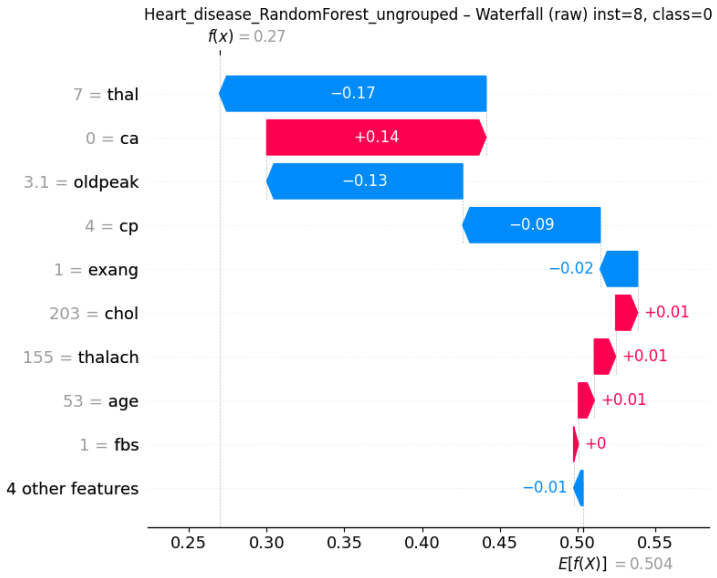


(c) Greedy method



(d) Random grouping

Figure 5.9: Comparison of SHAP plots for Heart disease classification by Random Forest between different grouping methods including ungrouped case (part 1).



(e) Ungrouped

Figure 5.9: Comparison of SHAP plots for Heart disease classification by Random Forest between different grouping methods including ungrouped case (continued).

Beyond the Heart Disease example, the Breast Cancer panels also make the cross-model effect of grouping clear. Under K-plus (Figure 5.11), all five learners produced waterfalls with shorter extremes and a smoother, stepped progression from the base value to the prediction, indicating that no single feature dominated the local decision. By contrast, the ungrouped counterparts (Figure 5.12) often contained one or two large jumps with almost no contribution from the non top-9 features, particularly for the tree ensembles, signalling heavier reliance on a small subset of raw variables. This clear trend across all models suggests the attenuation of extreme local attributions was induced by the grouped representation rather than any single model class.

This qualitative analysis is quantified and illustrated in Figure 5.10 by aggregating each instance's absolute SHAP magnitude and decomposing it into the sum over the top 9 features versus the remaining features. Grouping systematically shifted mass from the few largest contributors into the long tail. Continuing with the exemplar configuration in Figures 5.11 and 5.12, the "rest" portion always increased when grouped across all models. For tree-based models, it even escalated from 0.06 to 0.48 for Random Forest and from 0.08 to 0.57 for XGBoost. Indeed, the increased total absolute attribution was another consistent benefit of grouping. For example, Logistic Regression case also yielded the smallest increase of 10%, but such increase was even approximately double for SVM ($0.46 \rightarrow 0.88$) and MLP ($0.47 \rightarrow 0.96$).

Overall, Figure 5.10 supports the findings from Figures 5.11 and 5.12 by confirming that grouping did not merely rescale the same few features, but it also broadened participation

across many features at the instance level. First, the flatter global SHAP spectra seen earlier under grouping manifest locally with smoother paths with fewer extreme bars (compare Figures 5.11 and 5.12). Second, the larger "rest" component in Figure 5.10 suggests improved robustness to spurious correlations. When explanatory mass was more evenly distributed, perturbations to any single (possibly proxy) feature had less leverage on the prediction. This redistribution was consistent with the fairness improvements reported in the heatmap (see Figure 5.15), where methods that flattened local waterfalls and expanded the tail of contributing features also tended to reduce group disparity.

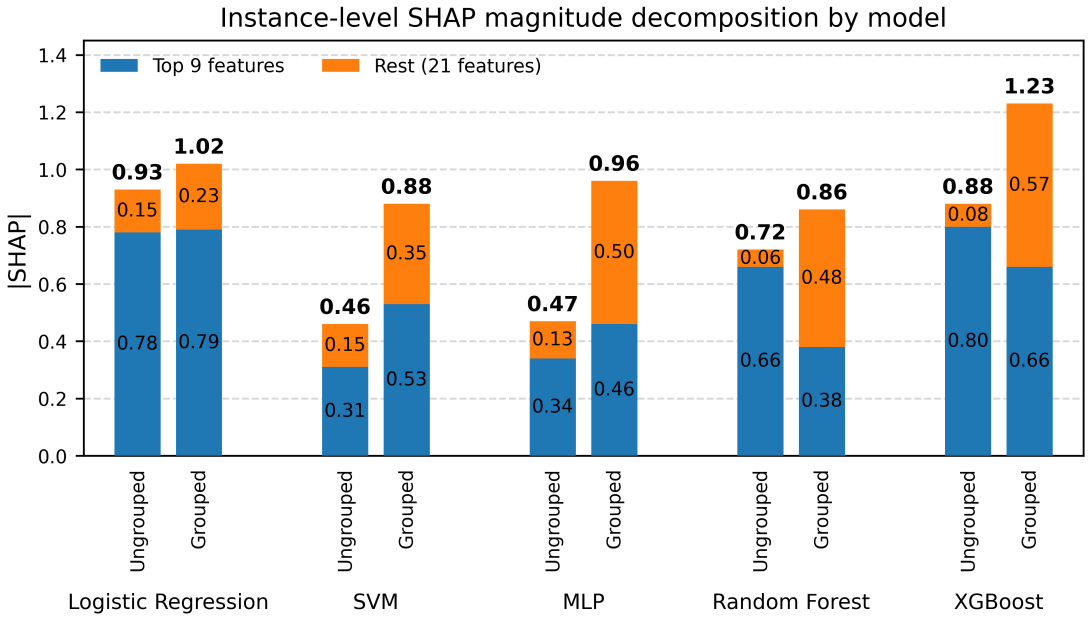
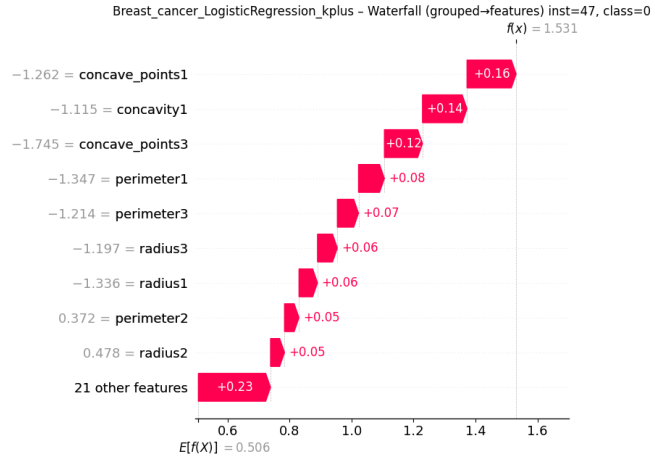
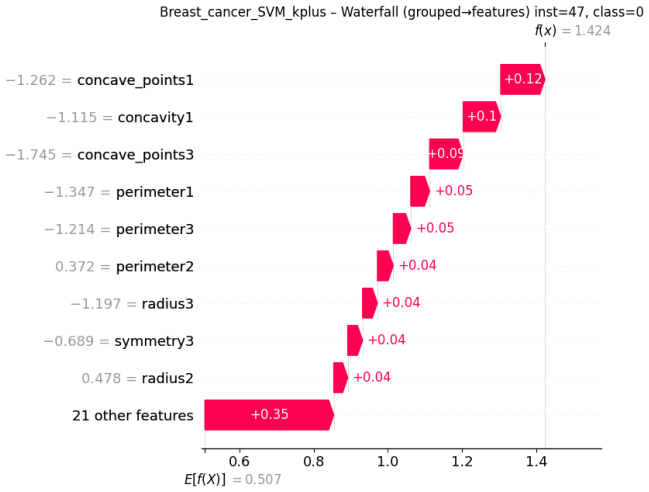


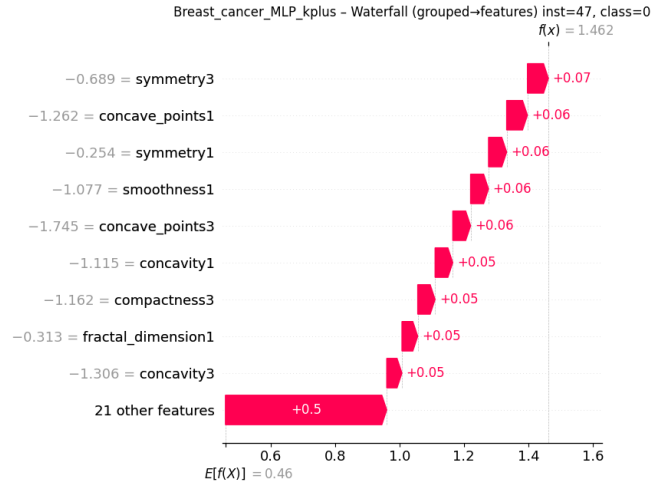
Figure 5.10: Stacked bar chart of the total absolute SHAP attribution to compare the configurations of Figures 5.11 and 5.12.



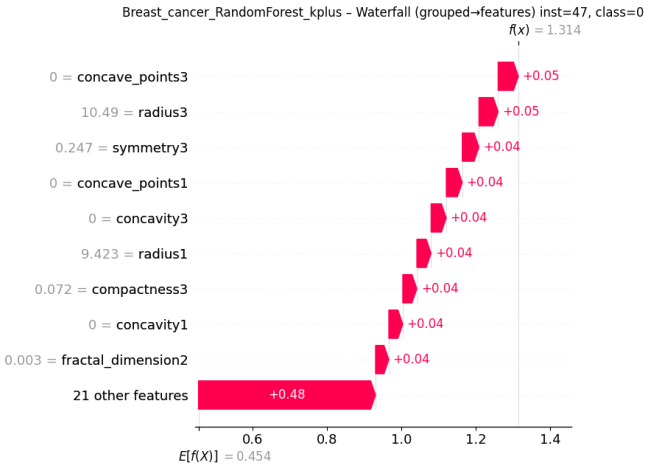
(a) Logistic regression



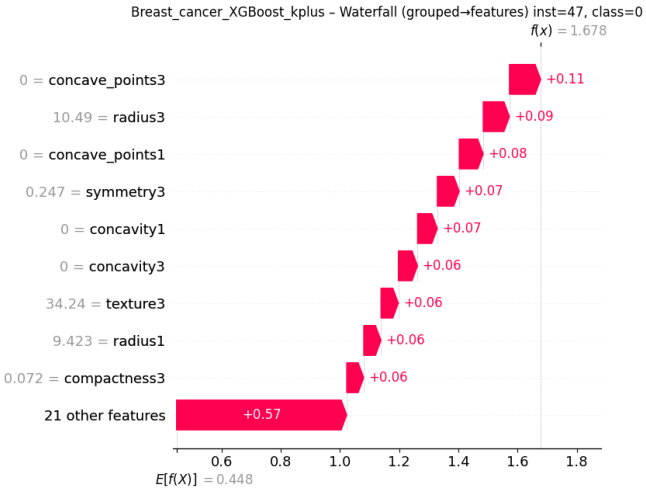
(b) SVM



(c) MLP

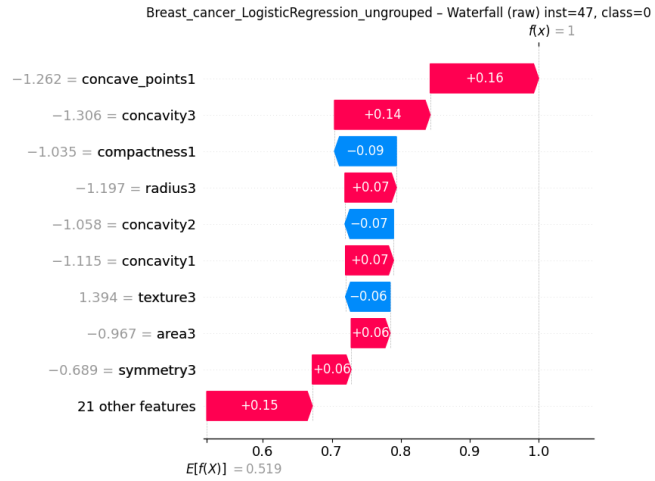


(d) Random Forest

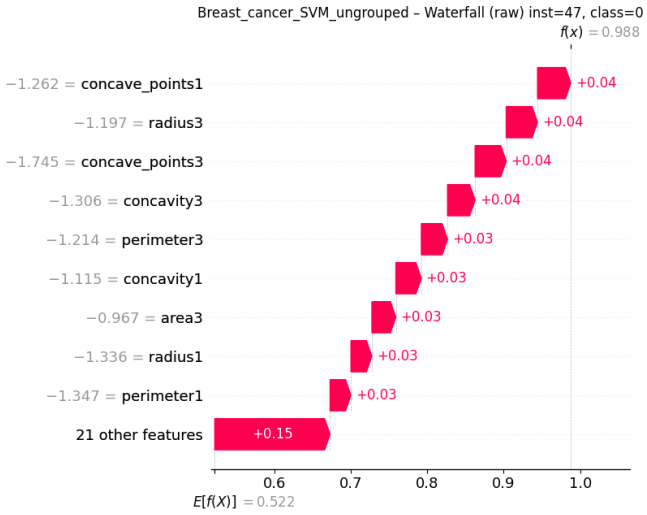


(e) XGBoost

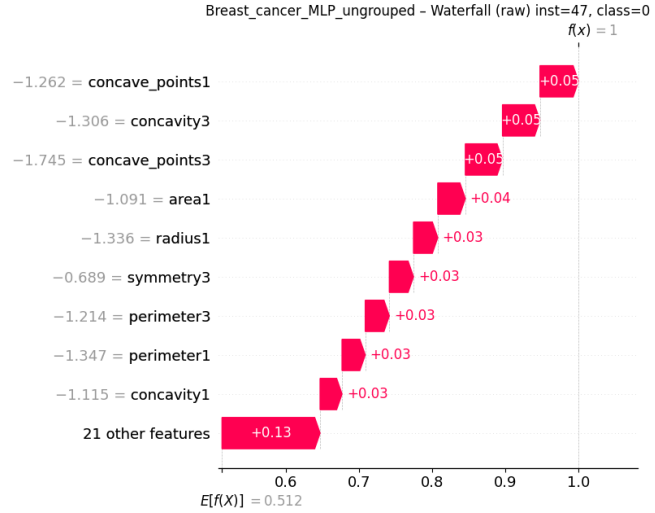
Figure 5.11: Comparison of SHAP waterfall plots from Breast cancer classification with K-plus grouping across different models.



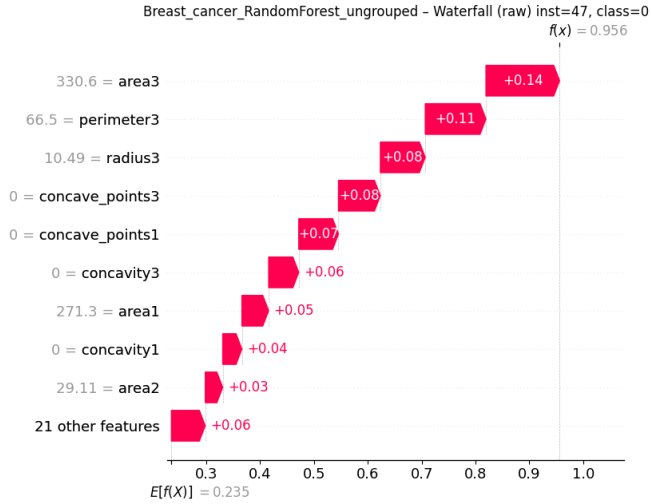
(a) Logistic regression



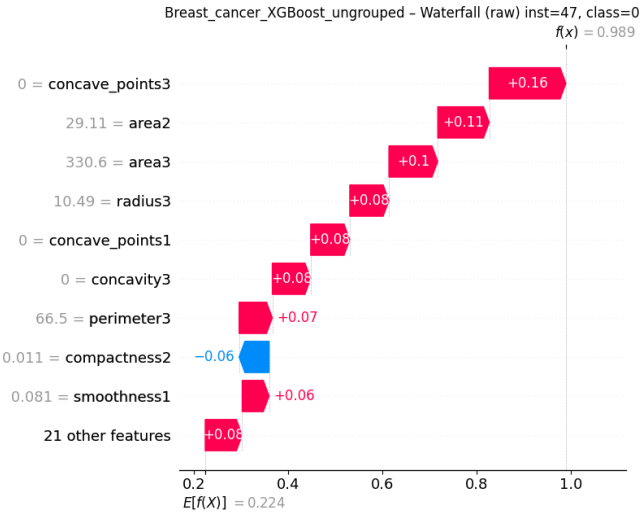
(b) SVM



(c) MLP



(d) Random Forest

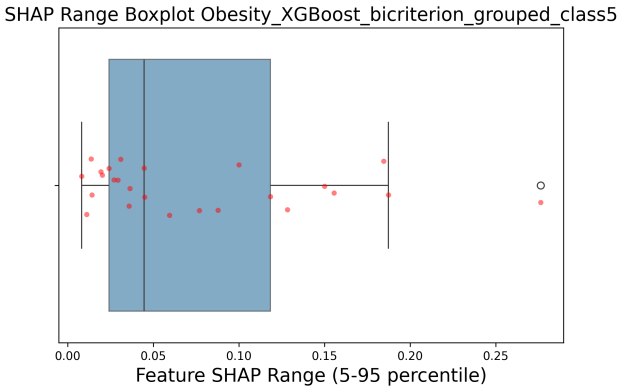


(e) XGBoost

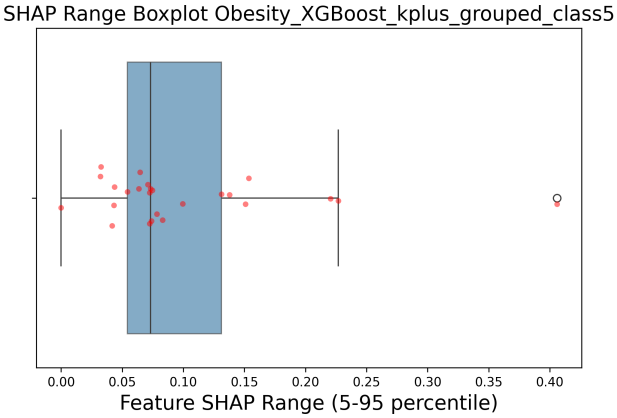
Figure 5.12: Comparison of SHAP plots from ungrouped Breast cancer classification across different models.

SHAP range statistics

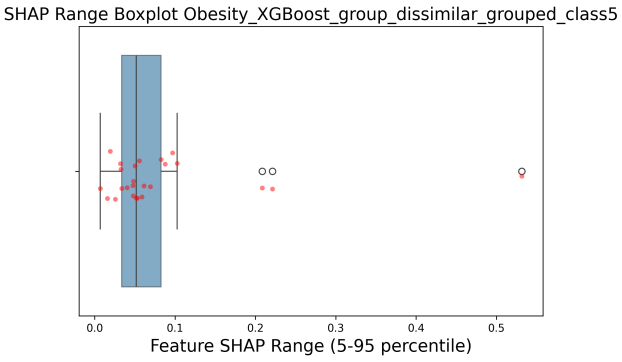
The spread of per-feature SHAP values was summarised by the range boxplots in Figure 5.13 (XGBoost on Obesity Type II) to offer a geometric complement to the SHAP summary and waterfall views. Grouped settings clearly increased the median range (about 0.05-0.07) relative to the ungrouped baseline (0.0), with Bicriterion showing the most stable spread overall. The interquartile range across different grouping methods remained similar, but greedy and K-plus yielded more extreme outliers, but still not as much as when ungrouped. This contraction aligned with the per-instance waterfalls (Figure 5.9), where the cumulative contribution path from the base value contained fewer extreme steps. Together with the global summaries (Figures 5.6, 5.7, and 5.8), these range statistics substantiate the claim that grouping dampened variance in local attributions while broadening participation across features.



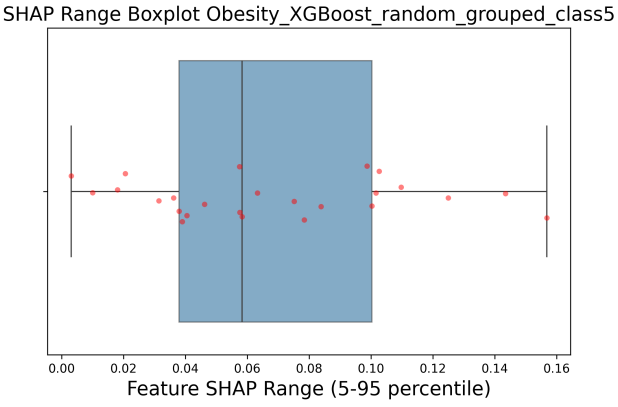
(a) Bicriterion method



(b) K-plus anticlustering

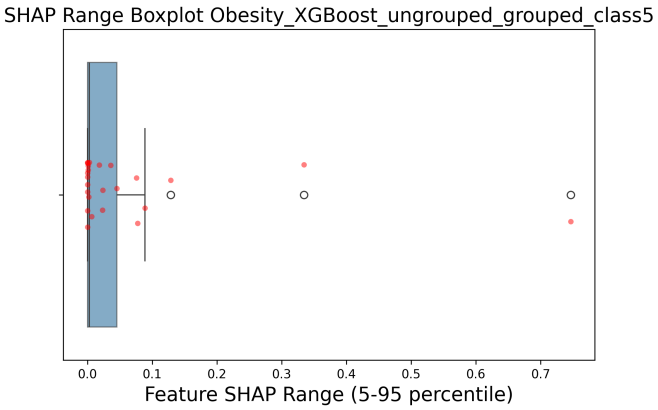


(c) Greedy method



(d) Random grouping

Figure 5.13: Box and whisker plots of the SHAP ranges for classifying Obesity Type II via XGBoost (part 1).



(e) Ungrouped

Figure 5.13: Box and whisker plots of the SHAP ranges for classifying Obesity Type II via XGBoost (continued).

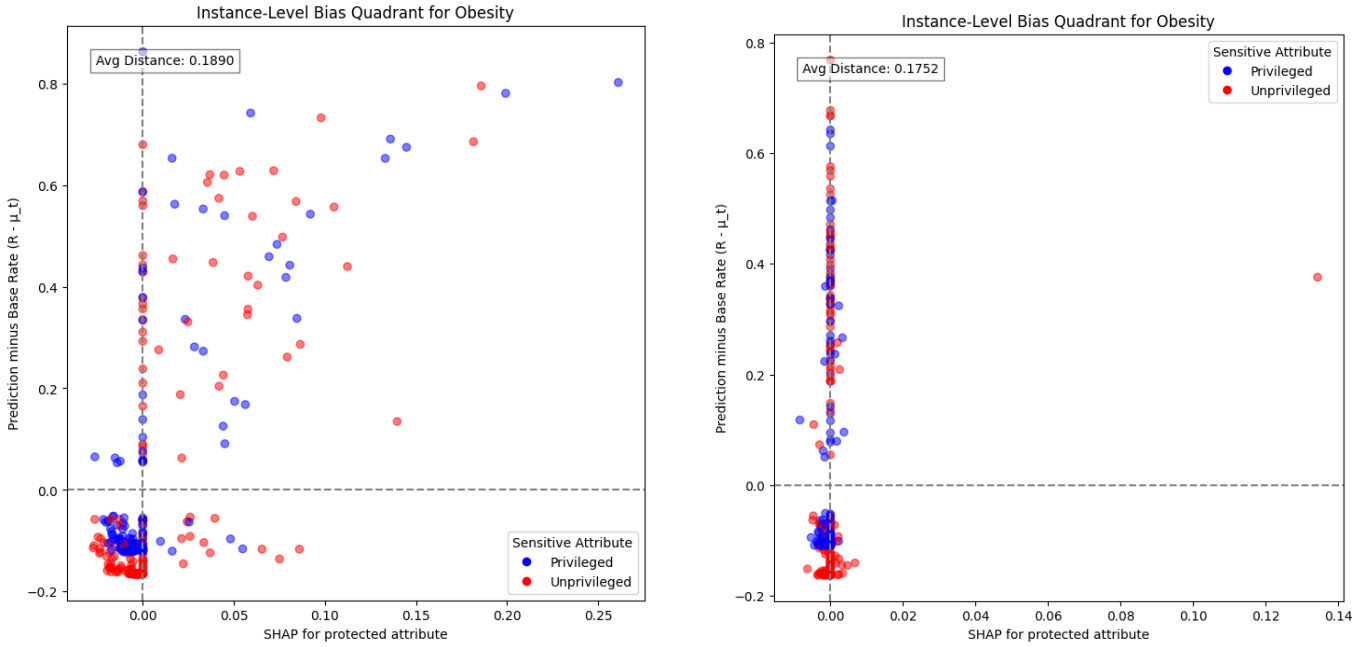
Bias quadrant analysis

The bias-quadrant plots juxtapose group-parity at the prediction level with explanation-parity from local SHAP attributions on protected attributes, so points near the origin indicate models that are fair both in outcomes and in how those outcomes are justified. Figure 5.14 illustrates how grouping consistently pulled configurations towards the origin along the y-axis compared with the ungrouped baseline, indicating reduced explanation disparity. This corroborates previous findings on how SHAP values were more evenly distributed for the grouped cases. Qualitatively, this movement corresponds to fewer instances where the protected attribute exerted asymmetric local influence (i.e. lower explanation bias). Meanwhile, the range and distribution of prediction bias was not so influenced by grouping, as they all deviated from -0.2 to 0.8. This suggests there still existed predictive disparity between different groups of sensitive attribute, but such difference in outcome was not necessarily due to its membership. Even when grouped, the unprivileged instances were still systemically a bit below the privileged instances in terms of y coordinates for quadrants III and IV. This is illustrated in Figure 5.14, where clustered red dots in bottom quadrants were consistently below the clustered blue dots by prediction base rate of approximately 0.1. This indicates the outcome of unprivileged instances were still somewhat suppressed or more encouraged to be negative despite the grouping technique mitigating the explanation bias.

Across grouping methods, K-plus led to the most stable contraction towards the origin. Its points were clustered in the low-bias region and avoided the quadrant I, where both prediction and explanation biases are high. However, it should be noted that this stability did not necessarily mirror its performance across other fairness metrics, since other metrics purely focused on the disparity of outcome. Bicriterion also moved instances

towards the origin and, in several tasks, attained the small geometric bias magnitude (Average Distance from Origin). One insightful observation was how it did better job in pulling the instances with negative SHAP values towards origin than that of positive values, as can be seen from Figure 5.14a with more clustering in quadrants II and III than quadrants I and IV. This distinguishable performance in clustering indicates Bicriterion was more effective in reducing the explanation disparity of the underprivileged instances with lower SHAP than regularising the privileged instances with higher SHAP. This is likely to be from the fact that Bicriterion accounts for maximising dispersion along with diveristy, which inherently sets more conservative lower threshold. Greedy and Random were more variable as their instances often landed in quadrant II (fair-looking predictions but residual explanation bias) or quadrant IV (prediction gaps without attributional evidence on the sensitive feature), patterns that are visible when explanation-level disparities do not track group-level ones.

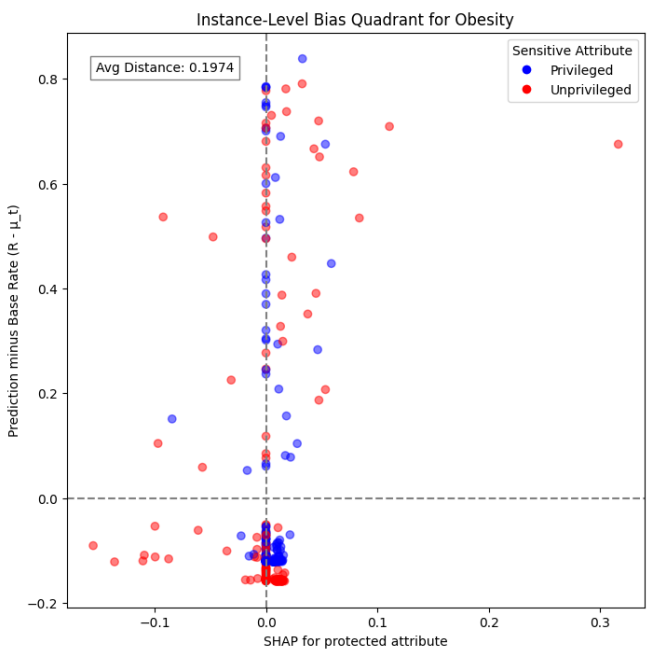
Overall, the quadrant view complemented the above SHAP analysis. Methods that flattened SHAP distributions and reduced reliance on a few dominant variables were the same ones that moved points inward, reducing both outcome disparity and explanation disparity on protected attributes. This integrated lens protects against 'apparent fairness' at only one level by requiring progress on both axes, as encoded in the final composite objective that emphasises explanation parity (7:3) to reflect the significance of explainability in clinical use.



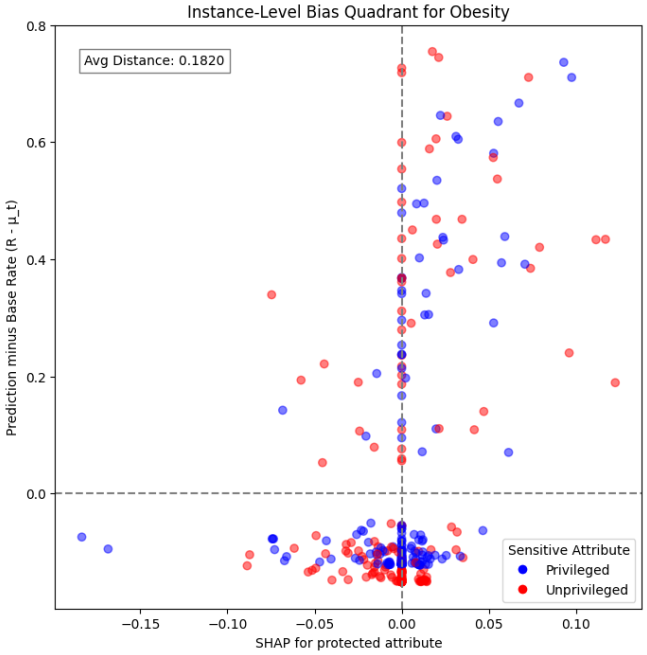
(a) Bicriterion method

(b) K-plus anticlustering

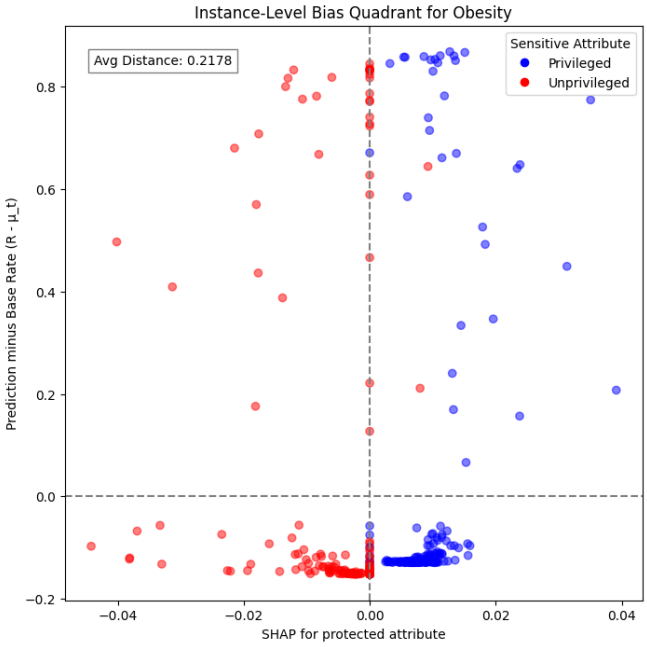
Figure 5.14: Comparison between bias quadrant plots of Obesity via SVM across different grouping methods (part 1).



(c) Greedy method



(d) Random grouping



(e) Ungrouped

Figure 5.14: Comparison between bias quadrant plots of Obesity via SVM across different grouping methods (continued).

Fairness metrics across grouping methods

Beyond explainability, this study systematically evaluated fairness outcomes across the five grouping strategies and the ungrouped baseline using six standard metrics: Equal Opportunity, Equalised Odds, Predictive Parity, N-Sigma (Error Rate), Average Distance from Origin, and the aggregated Fairness Overview score. The heatmaps in Figure 5.15 provide a comparative visualisation, with each cell representing the metric value for a given dataset and grouping method.

Ungrouped models exhibited substantial disparities for Equal Opportunity, particularly in the Obesity dataset, where the unfairness score reached 0.520, far higher than any grouped configuration. Grouping mitigated this imbalance, with the K-plus method achieving a markedly lower value of 0.373. Bicriterion and Random did even better with 0.232 and 0.214, respectively. These results highlight that grouping reduced the extent to which models unfairly privileged one class in true positive rates, a particularly important consideration in healthcare applications where equitable sensitivity is critical.

A similar pattern emerged for Equalised Odds, with the ungrouped Obesity model showing the highest disparity at 0.520. Grouping consistently reduced this value, with K-plus again achieving the highest among grouping methods at 0.373. Notably, the Greedy method recorded a moderately low disparity of 0.281, showing that while grouping in general improved fairness, the choice of strategy significantly influenced the degree of improvement.

Predictive Parity exhibited greater variance across datasets. For Heart Disease, Greedy yielded the highest unfairness at 0.850, followed by K-plus at 0.750. In contrast, ungrouped models performed relatively well in this case, with Heart Disease at 0.647. These results suggest that predictive parity may worsen under grouping for certain datasets, emphasising the trade-off between different fairness objectives.

The Obesity dataset showed the highest error rate disparity under Greedy (0.359), while ungrouped models performed better at 0.175. The more advanced approaches such as Bicriterion (0.233) and Random (0.315) fell in between. This indicates that grouping did not universally improve error rate fairness and may have exacerbated imbalances depending on the dataset and grouping method in some cases.

The average distance from origin of the bias quadrant captured overall deviation from ideal fairness across all protected attributes. Obesity and Obesity_imbalanced consistently recorded higher values, indicating these tasks were more fairness-challenging. For Obesity_imbalanced, ungrouped and K-plus both showed the highest unfairness at 0.313, whereas Bicriterion performed the best at 0.281. These results suggest that while grouping often redistributed feature contributions more equitably, it may have not always minimised systemic unfairness in model predictions.

Aggregating across all metrics, Greedy (0.363 for Heart Disease) and K-plus (0.346 for Heart Disease) showed strong but dataset-specific improvements, while Bicriterion

showed a consistent performance (0.271 for Heart Disease, 0.126 for Obesity). Ungrouped models demonstrated competitive results in some cases (e.g. Heart Disease 0.359), but consistently underperformed on Obesity, underscoring the importance of grouping for fairness in multi-class settings.

A crucial part of this study involved comparing the balanced Obesity dataset against the synthetically imbalanced Obesity_imbalanced variant created using SMOTE. The imbalanced version exhibited noticeably higher unfairness in metrics such as Equal Opportunity (e.g. Random achieving 0.142 in imbalanced, compared to 0.103 in balanced) and Average Distance from Origin (e.g. Ungrouped achieving 0.313 in imbalanced, compared to 0.216 in balanced). This confirms that increasing label skewness amplified disparities, with grouped methods partially but not fully mitigating the effect. In particular, K-plus and Bicriterion retained relatively robust performance under imbalance, suggesting that grouping could buffer against some of the fairness degradation caused by unequal class distributions.

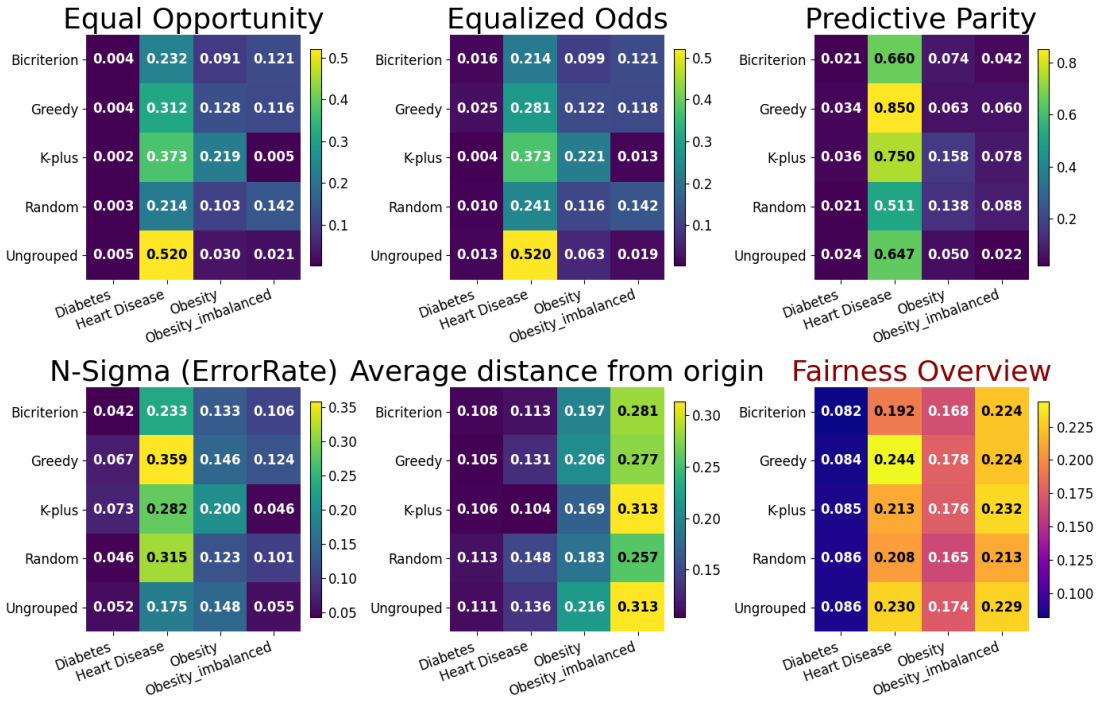


Figure 5.15: Equal opportunity, Equalised odds, Predictive parity, N-sigma and average distance from origin of bias quadrant of models across datasets and grouping methods. Fairness overview is calculated with more weight on explanation parity (7:3) compared to group parity (see Equation (4.27)). Note that lower is better for all fairness metrics above.

5.4.3 Practical implications for clinical deployment

The patterns observed, namely in Figures 5.6, 5.10 and 5.15, have operational value beyond metric improvements. By redistributing explanatory credit across a broader set of variables and instances, grouped models made more effective use of what had already been collected. In practice, this increases sample efficiency, since informative signal is drawn from a wider portion of the cohort rather than a narrow proxy-dominated subset. For prospective studies, this can ease recruitment pressure and reduce participant burden, since acceptable behaviour may be reachable without continually expanding the cohort solely to stabilise decision boundaries [94, 56]. Not only does SHIELD reduce the burden for clinicians with equitable distribution of feature importance, but it also makes the data collection more appealing for participants, since the framework assures the contribution of individual instances to the outcome, avoiding a waste of records that ungrouped cases have shown to exert.

A more even reliance on features also prevents from extravagant panel design. When dominance by a few correlated variables is reduced, the marginal value of adding further tests or questionnaires diminishes, which can improve data acquisition costs and streamline workflows. Decoder-mapped SHAP further helps clinical teams to verify that retained features remain clinically plausible, and to retire redundant or low-yield variables without losing traceability to the native feature space [3, 80].

These advantages do not license indiscriminate data reduction. Any operational savings must be conditioned on non-inferior predictive performance, stability under external validation, and domain review of feature rankings. In imbalanced or rare-event settings, the same caution applies to fairness: apparent gains from dispersion should be corroborated by parity on outcome metrics and by sensitivity analyses that test robustness to shifts in class prevalence [102, 94]. Considering the above aspects, SHIELD offers a practical route to improve equity and usability, while respecting clinical constraints on accuracy and accountability, which still has room for improvment as identified in Section 5.5 and Chapter 6.

5.5 Limitations

This study demonstrates that dissimilarity-based grouping with decoder mapping can flatten SHAP distributions and improve several fairness measures, but it also has important limitations. These limitations do not negate the value of the findings, but they delineate the conditions under which they hold and the directions where further theoretical and empirical work is needed. Hence, this section summarises these to guide interpretation and motivate future work.

The first limitation concerns the breadth of model selection and hyperparameter search. Although the number of groups K was initially planned to be searched alongside group-

ing weights and model parameters using five-fold cross-validation, the search budget was unfortunately limited [106, 17]. In particular, the range explored for K and the weighting between performance and fairness may not fully cover the space of Pareto-optimal trade-offs. The methodology in this thesis treats K as a tunable variable within Bayesian optimisation in principle, but in practice several configurations were selected by inspection and then held fixed to compromise with feasible runtime, which risks settling on locally adequate rather than globally optimal settings. The net effect is that some reported improvements could be sensitive to search scope and seed choice.

An alternative splitting strategy not explored in this thesis is adversarial data splitting [9], which intentionally constructs more challenging testing set to stress generalisation under distribution shift. Because SHIELD relied on conventional random stratified splits, it may understate model fragility when faced with edge cases. Future work should evaluate adversarial or hard-split schemes in clinical contexts to provide a more rigorous test of robustness.

The decoder-mapped explainability pipeline trades exactness for traceability. To project latent attributions back to original features, SHIELD uses normalised absolute decoder weights. This preserves a clear bridge from latent coordinates to raw variables, but it discards signs and compresses attribution through a linear weighting that is not itself a Shapley solution. In correlated or highly non-linear regimes this mapping can blur sparsity and directionality, so feature-level explanations should be read as principled approximations rather than literal decompositions of the latent-space SHAP values. While this choice enables end-to-end auditability in grouped spaces, it can understate dominant effects that are concentrated in a small number of latent coordinates or overstate diffuse ones when decoder columns are broad.

Fairness assessment is another area where design choices limit generality. The evaluation emphasised a holistic perspective, which consider group-based metrics such as Equal Opportunity and Equalised Odds together with an N-Sigma normalisation of error gaps, and an explanation-level audit via bias quadrants. These metrics capture salient aspects of parity but embody assumptions about acceptable trade-offs when base rates differ. As discussed in the methodology, Equalised Odds can conflict with clinical realities if groups genuinely have different prevalences, making Equal Opportunity a more defensible target in some settings. The bias-quadrant analyses further show that grouping reliably reduces explanation disparity along the attribution axis, yet prediction disparities can persist across the full range, indicating that more equitable rationales do not automatically guarantee equitable outcomes. These choices leave out individual-fairness notions and counterfactual or causal criteria. As such, the conclusions pertain to the specific set of statistical metrics adopted.

The study scope also constrains external validity. All experiments were conducted on structured tabular datasets with supervised classification endpoints. This thesis did not evaluate regression, time-to-event outcomes, or free-text modalities, and there were no external, prospective clinical validation. Performance targets and fairness constraints

were optimised under five-fold cross-validation rather than assessed with feedback from domain experts, which limits claims about deployment robustness. The domain knowledge was only used to cross-check the findings. Moreover, many visual analyses present single trained instances per configuration for clarity, so sensitivity to random seeds and reinitialisation is only partially explored, although the randomness still supports generalisability of the framework.

Finally, the findings are empirical. They align with conceptual expectations that grouping reduces over-reliance on a few correlated features and diffuses attributional bias, but formal guarantees are not provided in this thesis. In particular, the author has not proved conditions under which the risk of the grouped model is bounded relative to the ungrouped baseline, nor characterise when decoder-mapped attributions are provably faithful at the original feature level. The next Chapter 6 outlines how such guarantees might be approached, including bounds on risk inflation and stability of fairness metrics under grouping-induced perturbations. Until then, the results should be interpreted as credible patterns in the examined settings rather than universal laws.

Chapter 6

Future works

6.1 Theoretical validation of grouped representations

The empirical study suggests that grouping by conditional dissimilarity and decoding back to the original variables can flatten attribution spectra and often improve fairness with only modest performance cost. A natural next step is to formalise when these effects should be expected. This section outlines several concrete propositions and proof strategies that can be developed into rigorous results. Each target is stated with minimal assumptions, an illustrative example, and a sketch of the argument, so that they can serve as starting points for formal analysis.

The first target is to relate predictive risk before and after grouping. Let g denote the encoder from raw features to the grouped latent space and D be the linear decoder used to map latent attributions back to the original variables. For a learner h acting on the latent space, define the composed predictor $f(x) = h(g(x))$ and the reference predictor $\tilde{f}$ trained on the raw features. Under a Lipschitz loss ℓ, a bound of the form

$$\mathcal{R}(f) - \mathcal{R}(\tilde{f}) \;\le\; C_1 \,\mathrm{ReconErr}(D \circ g) \;+\; C_2 \,\Delta_{\mathrm{cap}}(h, \tilde{f}) \tag{6.1}$$

is expected, where $\mathcal{R}$ denotes expected risk, ReconErr measures encoder-decoder distortion, and Δ_{cap} captures the difference in effective capacity between h and $\tilde{f}$ for some constants C_1, C_2 [110]. The proof strategy follows standard stability or Rademacher-complexity arguments [13]. If the decoder is near-isometric on the data manifold, the latent hypothesis class is not more complex than the raw one, and reconstruction error is small, then excess risk is controlled. In the linear case with an orthonormal decoder and a convex loss, the risk can even be preserved exactly by projecting the raw optimum into the latent span, which explains why the empirical accuracy often remains stable when grouping is applied.

The second target is to establish stability of decoded SHAP attributions. Let $\phi_j^D(f, x)$ denote the decoded contribution assigned to original feature j at instance x. When two decoders D and D' are close in operator norm and the explainer is locally Lipschitz in the model parameters, one expects a perturbation bound

$$\left|\phi_j^D(f, x) - \phi_j^{D'}(f, x)\right| \;\le\; L_{\mathrm{expl}} \,\|D - D'\| \,\|z(x)\|, \tag{6.2}$$

where $z(x) = g(x)$ is the latent representation and L_{expl} summarises the explainer's sensitivity. For linear models, this reduces to a simple product of weight and decoder differences, and for tree ensembles, the path probabilities of TreeExplainer give an analogous control. This helps to justify that changes from the global SHAP profiles are driven by representation changes rather than instabilities of the decoding step itself [74, 60].

The third target connects grouping by conditional dissimilarity to fairness leakage. Recall that the grouping aims to reduce $I(X_j; A \mid Y)$, which is the information a feature X_j carries about a protected attribute A once the clinical state Y is known. Let $Z = g(X)$ be the grouped representation. If the grouping achieves $I(Z; A \mid Y)$ that is uniformly lower than $I(X; A \mid Y)$, then for any thresholded score with calibrated class-conditional distributions, Pinsker-type inequalities imply that disparities in true and false positive rates are bounded above by constants times $\sqrt{I(Z; A \mid Y)}$ plus a calibration term [35]. This would formalise the intuition that reducing conditional dependence between representation and protected attribute limits the room for equalised-odds violations, and explains the systematic improvement of the origin in the bias-quadrant plots when grouping is applied [34, 60].

The fourth and final target concerns identifiability of proxy variables after grouping. Suppose a small subset of original features retains large $I(X_j; A \mid Y)$ even after grouping and decoding, and these features repeatedly receive high decoded SHAP across instances in the unprivileged group. Under mild regularity, one can show that such features are flagged with high probability by a simple test that combines estimation of CMI with a dominance score based on the waterfall top-k mass. This provides a theoretically justified screening rule for proxy auditing that complements the visual diagnostics already used in the previous chapter.

Together, these targets outline a coherent goal. By bounding excess risk in terms of encoder-decoder distortion and capacity, proving stability of decoded attributions, and tying conditional information leakage to outcome disparities, SHIELD would rest on clear theoretical pillars. Each result is directly connected to a component already implemented in this thesis, so that formal proofs can build on the proposed pipeline rather than start from scratch. As these theorems are developed, they will also guide practical choices, for example how to better set K, how to constrain decoders for attribution stability, and when grouping is most likely to deliver fairness gains without unacceptable loss of accuracy.

6.2 Instance level extension

The interpretability of SHIELD could be extended to the instance level. While this study has shown how grouping features by conditional dissimilarity and mapping latent representations back to the original space enhances fairness and interpretability mainly at the feature level, the explanations still treat all training instances uniformly. A natural

next step is to quantify how individual training points themselves contribute to model behaviour. Early work in this direction includes Data Shapley [50], which evaluates the value of each training instance to overall model performance. Building on this line of work, more recent approaches such as residual Shapley decomposition (RSHAP) [72] provide scalable approximations by focusing on residual error contributions rather than retraining-based value estimation. In this sense, RSHAP can be viewed as an extension that retains the Shapley-theoretic grounding of Data Shapley while making it feasible for larger datasets and complex models.

Integrating these approaches with the proposed pipeline would allow a more comprehensive view of unfairness. For example, combining the decoder-weight mapping with residual instance attributions could uncover whether certain features' undue influence is driven by a handful of mislabelled or systematically biased training points, an insight that would not be visible from feature-level SHAP alone. This is particularly relevant in fairness auditing with bias quadrant plots, where instance-level scores could show whether unfair prediction disparities originate from a small subset of influential examples rather than the model structure or grouping method. Ultimately, attribution analyses at both feature-level and instance-level would support more robust, fair, and trustworthy pipelines that align with clinical demands for transparent and justifiable decision-making.

6.3 Application of the framework to regression problems

Another intriguing direction for future research is to test SHIELD on regression problems. This thesis conducted classification experiments, which grouped and ungrouped models often achieved similar predictive accuracy despite exhibiting notably different distributions of feature importance across grouping methods. This stability may be due to the fact that classification tasks rely on a discrete decision boundary. That is, even if feature attributions are redistributed, the most probable class remains unchanged as long as its margin is preserved. In regression, however, the target is continuous and prediction errors scale directly with contribution magnitudes. Hence, small differences in how grouped versus ungrouped features perform may aggregate to result in larger discrepancies in prediction accuracy.

Applying SHIELD to regression tasks would reveal whether the same benefits observed here, including flattened SHAP distributions and improved fairness, carry over when predictions are sensitive to fine-grained feature weightings. It would also open opportunities to evaluate fairness under continuous error metrics such as mean absolute error disparities or mean residual difference [132, 30], complementing the classification-focused measures used in this thesis. The author hypothesises that while grouping may still reduce over-reliance on dominant variables, its impact on accuracy-fairness trade-offs will be more pronounced in regression, particularly in cases with skewed feature distributions or outlier-heavy targets.

Chapter 7

Conclusion

This thesis investigated whether grouping features by conditional dissimilarity and mapping grouped representations back to the original space via a decoder could make machine learning models both more explainable and equitable to support clinical decision. The proposed pipeline integrated three core components: a dissimilarity-driven grouping stage, a decoder that localises latent effects to original variables, and an audit suite that couples attributional analyses with group-parity metrics. SHIELD was evaluated across various clinical tabular datasets and multiple classifiers, including linear, margin-based, neural and tree ensembles, to test for model-agnostic effects.

Empirically, the approach consistently reduced 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 features in the decision process. Per-instance waterfall plots further revealed fewer extreme steps from base value to prediction, aligning with narrower ranges in local attributions for grouped models. Taken together, these results corroborated the claim that grouping dampens variance in local attributions and encouraged more use of information of a given data.

Fairness analyses complemented the explainability findings. Using six standard metrics, including Equality of Opportunity, Equalised Odds, Predictive Parity, N-Sigma error, Average Distance from Origin in the bias quadrant, and a composite Fairness Overview, ungrouped cases often exhibited substantial disparities, which were improved to a different degree across grouping methods. This was especially pronounced when the dataset was scarce and imbalanced. This exemplified that grouping could lessen the degree to which models privilege one group, which is especially relevant in healthcare screening and triage.

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 delivered the most consistent overall improvements.

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.

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 regression, 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.

The thesis closed by outlining three promising avenues for consolidation and extension. First, theoretical foundations could establish conditions under which grouped representations preserve predictive risk and stabilise fairness metrics, yielding explicit bounds and diagnostics that connect back to practice. Second, extending the framework to instance-level analysis via residual Shapley decomposition would help to identify individual instances that propagate bias or instability into the latent space, enabling targeted data remediation. Third, applying the pipeline to regression would test whether the accuracy-fairness trade-offs observed here persist when errors vary continuously and can be evaluated with continuous fairness criteria.

In practical terms, the evidence suggested that dissimilarity-based grouping with decoder mapping could deliver benefits beyond metric gains. By drawing signal from a broader share of features 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 features also supports parsimonious testing designs and data collection protocols, saving acquisition and processing costs while keeping explanations traceable to native clinical features through decoder-mapped SHAP. These operational benefits should always be conditioned on non-inferior predictive performance, external validation, and clinical plausibility of the induced feature rankings, which SHIELD has empirically corroborated.

In conclusion, the results showed that grouping dissimilar features 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 contributions of this thesis provide a concrete step from concept to practice. SHIELD

is lightweight, model-agnostic, and compatible with existing pipelines. It also points to a responsible way forward in which accuracy, explainability, and fairness are pursued together rather than heavily compromising each other. In doing so, this work synergised the four research areas in Figure 1.1 and contributed to the virtuous circle outlined in Section 1.1: methodological advances build more trustworthy and equitable explanations, which foster confidence in clinical application, and this in turn motivates further methodological refinement.

Bibliography

[1] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximateions to Shapley values. *Artificial Intelligence*, 298, 2021. [Cited on pages 2 and 46.]

[2] Charu C. Aggarwal. *Outlier Analysis*. Springer, 2nd edition, 2017. [Cited on page 30.]

[3] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99:101805, 2023. [Cited on pages 3, 9, 10, and 78.]

[4] American Diabetes Association Professional Practice Committee. Standards of Care in Diabetes-2024. *Diabetes Care*, 47(Suppl 1), 2024. [Cited on page 19.]

[5] Valentin Amrhein, Sander Greenland, and Blake McShane. Scientists rise up against statistical significance. *Nature*, 567:305–307, 2019. [Cited on page 8.]

[6] Australian Institute of Health and Welfare. Heart, stroke and vascular disease: Australian facts - Coronary heart disease. https://www.aihw.gov.au/reports/heart-stroke-vascular-diseases/hsvd-facts/contents/all-heart-stroke-and-vascular-disease/coronary-heart-disease, 2024. [Cited on page 18.]

[7] Australian Institute of Health and Welfare. Cancer data in Australia. https://www.aihw.gov.au/reports/cancer/cancer-data-in-australia/contents/overview#breast, 2025. [Cited on page 18.]

[8] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49, 2011. [Cited on page 33.]

[9] Amanda S. Barnard. BenchMake: turn any scientific data set into a reproducible benchmark. *Machine Learning: Science and Technology*, 6(3), 2025. [Cited on page 79.]

[10] Solon Barocas and Andrew D. Selbst. Big Data's Disparate Impact. *California Law Review*, 104(3):671–732, 2016. [Cited on pages 1, 4, 8, and 35.]

[11] Caleb J. S. Barr, Olivia Erdelyi, Paul D. Docherty, and Randolph C. Grace. A Review of Fairness and A Practical Guide to Selecting Context-Appropriate Fairness Metrics in Machine Learning, 2025. [Cited on pages 9 and 47.]

[12] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, 58:82–115, 2020. [Cited on pages 10 and 11.]

[13] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher Complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005. [Cited on page 81.]

[14] Jenny M. Bauer and Martin Michalowski. Human-centered explainability evaluation in clinical decision-making: a critical review of the literature. *Journal of the American Medical Informatics Association*, 2025. [Cited on page 1.]

[15] Daniel Beechey, Thomas M. S. Smith, and Özgür Şimşek. Explaining reinforcement learning with shapley values. In *Proceedings of the 40th International Conference on Machine Learning*, pages 2003–2014, 2023. [Cited on page 21.]

[16] Asa Ben-Hur and Jason Weston. A User's Guide to Support Vector Machines. *Data Mining Techniques for the Life Sciences*, pages 223–239, 2009. [Cited on page 42.]

[17] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13:281–305, 2012. [Cited on pages 8, 43, and 79.]

[18] James Bergstra, Daniel Yamins, and David Cox. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 115–123, 2013. [Cited on page 43.]

[19] Gérard Biau and Erwan Scornet. A random forest guided tour. *TEST*, 25(2):197–227, 2016. [Cited on page 42.]

[20] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. [Cited on pages 3, 7, and 13.]

[21] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152. ACM, 1992. [Cited on page 42.]

[22] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: Cancer Journal for Clinicians*, 68(6):394–424, 2018. [Cited on page 17.]

[23] Leo Breiman. Random Forests. *Machine Learning*, 45:5–32, 2001. [Cited on pages 3 and 42.]

[24] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The Journal of Machine Learning Research*, 13(1):27–66, 2012. [Cited on pages 15, 16, and 35.]

[25] Michael J Brusco, Dennis J Cradit, and Douglas Steinley. Combining diversity and dispersion criteria for anticlustering: A bicriterion approach. *British Journal of Mathematical and Statistical Psychology*, 73(3):375–396, 2020. [Cited on page 37.]

[26] Julia Buss. Limitations of body mass index to assess body fat. *Workplace Health Saf.*, 62(6), 2014. [Cited on page 17.]

[27] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1721–1730, 2015. [Cited on page 8.]

[28] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, 8(1), 2018. [Cited on page 32.]

[29] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016. [Cited on pages 3, 13, and 42.]

[30] Jianfeng Chi, Yuan Tian, Geoffrey J. Gordon, and Han Zhao. Understanding and Mitigating Accuracy Disparity in Regression. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1866–1876, 2021. [Cited on page 83.]

[31] John Clore, Krzysztof Cios, Jon DeShazo, and Beata Strack. Diabetes 130-US Hospitals for Years 1999-2008. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5230J. [Cited on pages 19, 28, and 30.]

[32] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995. [Cited on pages 13 and 42.]

[33] Raphael Couronné, Philipp Probst, and Anne-Laure Boulesteix. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, 19, 2018. [Cited on page 13.]

[34] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2005. [Cited on pages 15, 35, 36, and 82.]

[35] Imre Csiszár and János Körne. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2015. [Cited on page 82.]

[36] Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. Proxy Non-Discrimination in Data-Driven Systems, 2017. [Cited on pages 2, 4, 16, and 35.]

[37] Daniel DeAlcala, Ignacio Serna, Aythami Morales, Julian Fierrez, and Javier Ortega-Garcia. Measuring Bias in AI Models: An Statistical Approach Introducing N-Sigma. *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1167–1172, 2023. [Cited on page 9.]

[38] Daniel DeAlcala, Ignacio Serna, Aythami Morales, Julian Fierrez, and Javier Ortega-Garcia. Measuring Bias in AI Models: An Statistical Approach Introducing N-Sigma. *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1167–1172, 2023. [Cited on pages 46 and 47.]

[39] Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H. Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 64(5):304–310, 1989. [Cited on pages 19 and 30.]

[40] Amit Dhurandhar, Karthikeyan Shanmugam, and Ronny Luss. Leveraging Simple Model Predictions for Enhancing its Performance, 2019. [Cited on page 28.]

[41] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012. [Cited on page 7.]

[42] Fabio Henrique Kiyoiti dos Santos Tanaka and Claus Aranha. Data Augmentation Using GANs, 2019. [Cited on pages 18 and 28.]

[43] Finale Doshi-Velez and Been Kim. Towards a Rigorous Science of Interpretable Machine Learning, 2017. [Cited on page 11.]

[44] Joshua Elliott, Barbara Bodinier, Matthew Whitaker, Rin Wada, Graham Cooke, Helen Ward, Ioanna Tzoulaki, Paul Elliott, and Marc Chadeau-Hyam. Sex inequalities in cardiovascular risk prediction. *Cardiovascular Research*, 120(11):1327–1335, 2024. [Cited on page 8.]

[45] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019. [Cited on pages 1 and 7.]

[46] Ludwig Fahrmeir, Thomas Kneib, and Susanne Konrath. Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing*, 20:203–219, 2009. [Cited on page 23.]

[47] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 329–338. ACM, 2019. [Cited on pages 2, 8, 9, and 41.]

[48] Matthew W. Gardner and Stephen R. Dorling. Artificial Neural Networks (the Multilayer Perceptron)—A Review of Applications in the Atmospheric Sciences. *Atmospheric Environment*, 32(14–15):2627–2636, 1998. [Cited on page 42.]

[49] Marc Ghanem, Abdul Karim Ghaith, Victor Gabriel El-Hajj, Archis Bhandarkar, Andrea de Giorgio, Adrian Elmi-Terander, and Mohamad Bydon. Limitations in Evaluating Machine Learning Models for Imbalanced Binary Outcome Classification in Spine Surgery: A Systematic Review. *Brain Sciences*, 13(12), 2023. [Cited on page 44.]

[50] Amirata Ghorbani and James Zou. Data Shapley: Equitable Valuation of Data for Machine Learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 2242–2251. PMLR, 2019. [Cited on page 83.]

[51] David C. Goff, Donald M. Lloyd-Jones, Glen Bennett, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk. *Circulation*, 129, 2013. [Cited on page 19.]

[52] Baptiste Gregorutti, Bertrand Michel, and Philippe Saint-Pierre. Correlation and variable importance in random forests. *Statistics and Computing*, 27(3):659–678, 2016. [Cited on pages 14 and 60.]

[53] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003. [Cited on page 7.]

[54] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc. [Cited on pages 9, 23, and 46.]

[55] Zengyou He and Weichuan Yu. Stable feature selection for biomarker discovery. *Computational Biology and Chemistry*, 34(4):215–225, 2010. [Cited on page 23.]

[56] Steven A. Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A. Riegler, Pål Halvorsen, and Sravanthi Parasa. On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, 12(1), 2022. [Cited on pages 1, 3, 4, 9, 10, 45, 57, and 78.]

[57] David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant. *Applied Logistic Regression*. Wiley, 3 edition, 2013. [Cited on pages 13 and 42.]

[58] Adela Hruby and Frank B Hu. The epidemiology of obesity: a big picture. *Pharmacoeconomics*, 33(7):673–689, 2015. [Cited on page 17.]

[59] John P. A. Ioannidis. The Importance of Predefined Rules and Prespecified Statistical Analyses: Do Not Abandon Significance. *JAMA*, 321(21):2067–2068, 2019. [Cited on page 8.]

[60] Aditya Jain, Manish Ravula, and Joydeep Ghosh. Biased models have biased explanations, 2020. [Cited on pages 2, 9, 14, 24, 46, 47, and 82.]

[61] Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. Heart Disease. UCI Machine Learning Repository, 1989. DOI: https://doi.org/10.24432/C52P4X. [Cited on pages 18, 19, and 29.]

[62] Julie Josse, Jacob M. Chen, Nicolas Prost, Gaël Varoquaux, and Erwan Scornet. On the consistency of supervised learning with missing values. *Statistical Papers*, 65:5447–5479, 2024. [Cited on page 32.]

[63] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012. [Cited on page 2.]

[64] William B. Kannel, Thomas R. Dawber, Abraham Kagan, Nicholas Revotskie, and Joseph Stokes. Factors of risk in the development of coronary heart disease: six year follow-up experience. the framingham study. *Annals of Internal Medicine*, 55:33–50, 1961. [Cited on page 19.]

[65] Michael Kearns and Aaron Roth. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, 2019. [Cited on page 35.]

[66] John D Kelleher, Brian Mac Namee, and Aoife D'Arcy. *Fundamentals of Machine Learning for Predictive Data Analytics*. MIT Press, 2 edition, 2020. [Cited on page 7.]

[67] Seo-Hee Kim, Sun Young Park, Hyungseok Seo, and Jiyoung Woo. Feature selection integrating Shapley values and mutual information in reinforcement learning: An application in the prediction of post-operative outcomes in patients with end-stage renal disease. *Computer Methods and Programs in Biomedicine*, 257:108416, 2024. [Cited on page 21.]

[68] Sotiris B Kotsiantis, Dimitris Kanellopoulos, and Panagiotis E Pintelas. Data Preprocessing for Supervised Learning. *International Journal of Computer Science*, 1(2):111–117, 2006. [Cited on pages 29 and 31.]

[69] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer, 2013. [Cited on pages 7, 29, 30, and 34.]

[70] Cihan Kuzudisli, Burcu Bakir-Gungor, Nurten Bulut, Bahjat Qaqish, and Malik Yousef. Review of feature selection approaches based on grouping of features. *PeerJ*, 2023. [Cited on page 22.]

[71] Roderick Little and Donald Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 3rd edition, 2019. [Cited on pages 32 and 33.]

[72] Tommy Liu and Amanda Barnard. Shapley based residual decomposition for instance analysis. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023. [Cited on page 83.]

[73] Scott Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2:56–67, 2020. [Cited on page 13.]

[74] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. [Cited on pages 2, 8, 12, 13, 42, 45, and 82.]

[75] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning Adversarially Fair and Transferable Representations. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3384–3393. PMLR, 2018. [Cited on page 35.]

[76] Kiran Maharana, Surajit Mondal, and Bhushankumar Nemade. A review: Data preprocessing and data augmentation techniques. *Global Transitions Proceedings*, 3(1):91–99, 2022. International Conference on Intelligent Engineering Approach(ICIEA-2022). [Cited on page 30.]

[77] Vasanti S Malik, Barry M Popkin, George A Bray, Jean-Pierre Després, and Frank B Hu. Sugar-sweetened beverages, obesity, type 2 diabetes and cardiovascular disease risk. *Circulation*, 121(11):1356–1364, 2010. [Cited on page 17.]

[78] G. Manikandan and S. Abirami. An efficient feature selection framework based on information theory for high dimensional data. *Applied Soft Computing*, 111:107729, 2021. [Cited on page 21.]

[79] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):1–35, 2021. [Cited on page 7.]

[80] Christoph Molnar. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/, 2025. Online book, version 2025-03-13. [Cited on pages 9, 42, and 78.]

[81] Marie Ng, Tom Fleming, Margaret Robinson, et al. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*, 384(9945):766–781, 2014. [Cited on page 17.]

[82] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. [Cited on pages 2 and 9.]

[83] Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *The Lancet*, 358(9291):1389–1399, 2001. [Cited on page 18.]

[84] Luca Oneto, Nicolò Navarin, Alessandro Sperduti, and Davide Anguita. *Fairness in Machine Learning*, pages 155–196. Springer, Cham, 2020. [Cited on page 8.]

[85] Jaganathan Palanichamy and Kuppuchamy Ramasamy. An improved feature selection algorithm with conditional mutual information for classification problems. In *2013 International Conference on Human Computer Interactions (ICHCI)*, pages 1–5, 2013. [Cited on pages 16 and 21.]

[86] Fabio Mendoza Palechor and Alexis de la Hoz Manotas. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data in Brief*, 25:104344, 2019. [Cited on page 30.]

[87] Fabio Mendoza Palechor and Alexis de la Hoz Manotas. Estimation of Obesity Levels Based On Eating Habits and Physical Condition . UCI Machine Learning Repository, 2019. DOI: https://doi.org/10.24432/C5H31Z. [Cited on pages 17 and 28.]

[88] Martin Papenberg. K-Plus anticlustering: An improved k-means criterion for maximizing between-group similarity. *British Journal of Mathematical and Statistical Psychology*, 77(1):80–102, 2024. [Cited on pages 35 and 38.]

[89] S. Gopal Krishna Patro and Kishore Kumar Sahu. Normalization: A Preprocessing Stage, 2015. [Cited on page 32.]

[90] Hanchuan Peng, Fuhui Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005. [Cited on page 15.]

[91] Charles M Perou, Therese Sørlie, Michael B Eisen, et al. Molecular portraits of human breast tumours. *Nature*, 406:747–752, 2000. [Cited on page 17.]

[92] Kedar Potdar, Taher Pardawala, and Chinmay Pai. A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. *International Journal of Computer Applications*, 175(4):7–9, 2017. [Cited on page 31.]

[93] David M. W. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, 2020. [Cited on pages 4, 8, and 45.]

[94] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine Learning in Medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019. [Cited on pages 1, 2, 3, 4, 8, 10, 45, and 78.]

[95] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*, pages 1135–1144, 2016. [Cited on pages 2, 8, and 12.]

[96] Juan C Rojas, John Fahrenbach, Sonya Makhni, Scott C Cook, James S Williams, Craig A Umscheid, and Marshall H Chin. Framework for Integrating Equity Into Machine Learning Models: A Case Study. *Chest*, 161(6):1621–1627, 2022. [Cited on page 8.]

[97] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. [Cited on page 39.]

[98] Sayanti Roy, Emily Kieson, Charles Abramson, and Christopher Crick. Mutual Reinforcement Learning, 2019. [Cited on page 21.]

[99] Daniel J. Rubin. Hospital readmission of patients with diabetes. *Current Diabetes Reports*, 15(4), 2015. [Cited on pages 19 and 20.]

[100] David E. Rumelhart and James L. McClelland. Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, volume 1: Foundations, pages 318–362. MIT Press, 1987. [Cited on page 42.]

[101] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007. [Cited on page 22.]

[102] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), 2015. [Cited on pages 1, 3, 4, 45, and 78.]

[103] Wenqian Shang, Houkuan Huang, Haibin Zhu, Yongmin Lin, Youli Qu, and Zhihai Wang. A novel feature selection algorithm for text categorization. *Expert Systems with Applications*, 33(1):1–5, 2007. [Cited on page 22.]

[104] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. [Cited on pages 4 and 14.]

[105] Jenni A. M. Sidey-Gibbons and Chris J. Sidey-Gibbons. Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, 19, 2019. [Cited on pages 18 and 28.]

[106] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical Bayesian Optimization of Machine Learning Algorithms. In *NIPS'12: Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 2, pages 2951–2959, 2012. [Cited on pages 26, 43, and 79.]

[107] Daniel Stamate, Wajdi Alghamdi, Daniel Stahl, Doina Logofatu, and Alexander Zamyatin. PIDT: A Novel Decision Tree Algorithm Based on Parameterised Impurities and Statistical Pruning Approaches. In *Artificial Intelligence Applications and Innovations*, pages 273–284, 2018. [Cited on page 18.]

[108] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. *BioMed Research International*, 2014. [Cited on page 20.]

[109] Hong Sun, Pouya Saeedi, Suvi Karuranga, et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Research and Clinical Practice*, 183, 2022. [Cited on page 19.]

[110] Ambuj Tewari and Sougata Chaudhuri. On Lipschitz Continuity and Smoothness of Loss Functions in Learning to Rank, 2016. [Cited on page 81.]

[111] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001. [Cited on page 33.]

[112] Australian National University. Human research ethics committees. [Cited on page 28.]

[113] Irvine University of California. Uc irvine machine learning repository. https://archive.ics.uci.edu/, 2025. [Cited on page 27.]

[114] Jorge R. Vergara and Pablo A. Estévez. A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24:175–186, 2014. [Cited on pages 15, 16, 35, and 36.]

[115] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2):841–887, 2018. [Cited on page 12.]

[116] Zachary J Ward, Sara N Bleich, Angie L Cradock, et al. Projected U.S. State-Level Prevalence of Adult Obesity and Severe Obesity. *New England Journal of Medicine*, 381(25):2440–2450, 2019. [Cited on page 17.]

[117] Ethan H. Weissler, Tristan Naumann, Tommy Andersson, et al. The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*, 22, 2021. [Cited on page 7.]

[118] Jenna Wiens, Suchi Saria, Mark Sendak, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9):1337–1340, 2019. [Cited on page 1.]

[119] Peter W. F. Wilson, Ralph B. D'Agostino, Daniel Levy, Albert M. Belanger, Halit Silbershatz, and William B. Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998. [Cited on page 19.]

[120] William Wolberg, Olvi Mangasarian, Nick Street, and W. Street. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5DW2B. [Cited on pages 18, 29, and 59.]

[121] World Health Organization. Obesity and overweight, 2025. [Cited on page 17.]

[122] Jaehong Yu, Hua Zhong, and Seoung Bum Kim. An Ensemble Feature Ranking Algorithm for Clustering Analysis. *Journal of Classification*, 37(2):462–489, 2020. [Cited on page 22.]

[123] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 325–333. PMLR, 2013. [Cited on pages 16 and 35.]

[124] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018. [Cited on page 2.]

[125] Yahong Zhang, Yujian Li, Ting Zhang, Pius Kwao Gadosey, and Zhaoying Liu. Feature clustering dimensionality reduction based on affinity propagation. *Intelligent Data Analysis*, 22(2):309–323, 2018. [Cited on page 22.]

[126] Ningsheng Zhao, Jia Yuan Yu, Krzysztof Dzieciolowski, and Trang Bui. Error Analysis of Shapley Value-Based Model Explanations: An Informative Perspective. In *AI Verification: First International Symposium, SAIV 2024, Montreal, QC, Canada, July 22–23, 2024, Proceedings*, page 29–48, Berlin, Heidelberg, 2024. Springer-Verlag. [Cited on pages 2, 14, and 46.]

[127] Ningsheng Zhao, Jia Yuan Yu, Krzysztof Dzieciolowski, and Trang Bui. Error Analysis of Shapley Value-Based Model Explanations: An Informative Perspective. In *AI Verification: First International Symposium, SAIV 2024, Montreal, QC, Canada, July 22–23, 2024, Proceedings*, page 29–48, Berlin, Heidelberg, 2024. Springer-Verlag. [Cited on pages 14, 24, and 46.]

[128] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'03, page 321–328, Cambridge, MA, USA, 2003. MIT Press. [Cited on page 32.]

[129] Zhengze Zhou and Giles Hooker. Unbiased measurement of feature importance in tree-based methods. *ACM Transactions on Knowledge Discovery from Data*, 15(2):1–21, 2021. [Cited on pages 14 and 60.]

[130] Zhiguo Zhou, Zhi-Jie Zhou, Hongxia Hao, Shulong Li, Xi Chen, You Zhang, Michael Folkert, and Jing Wang. Constructing multi-modality and multi-classifier radiomics predictive models through reliable classifier fusion, 2017. [Cited on page 28.]

[131] Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, 5(5):363–387, 2012. [Cited on page 30.]

[132] Anna Zink and Sherri Rose. Fair regression for health care spending. *Biometrics*, 76(3):973–982, 2020. [Cited on page 83.]