

Synergy-Guided Oblique Splits for Interpretable Tree Ensembles: A Partial Information Decomposition Approach

Anonymous Authors
Under Review

Abstract

Interpretable machine learning models such as FIGS (Fast Interpretable Greedy-tree Sums) use axis-aligned splits that cannot capture joint feature effects without deep trees. Oblique decision trees address this by splitting on linear combinations of features, but selecting which features to combine remains ad hoc. We propose SG-FIGS (Synergy-Guided FIGS), which uses Partial Information Decomposition (PID) to construct a synergy graph over feature pairs and restricts oblique splits to feature subsets connected by high synergy edges. We evaluate SG-FIGS across 12 OpenML classification benchmarks against standard FIGS, Random Oblique FIGS (RO-FIGS), and Gradient Boosting. Our results reveal a significant accuracy–interpretability trade-off: SG-FIGS-25 achieves 70.8% mean accuracy versus 81.2% for FIGS (bootstrap 95% CI for the gap: [6.3, 14.3] percentage points; Friedman $\chi^2 = 34.8$, $p = 1.66 \times 10^{-6}$). However, the 10% of splits that are oblique exhibit a mean Synergy Concentration Ratio of 2.80, confirming that synergy-guided feature selection concentrates on genuinely synergistic combinations. We identify three domain-meaningful case studies—a glucose-BMI-age metabolic triad for diabetes, tumor morphology features for breast cancer, and cardiovascular anatomy features for heart disease—demonstrating that PID synergy produces clinically interpretable oblique splits. A failure taxonomy reveals that synergy graph sparsity (4/12 datasets), oblique split incompatibility (3/12), and harmful synergy guidance (3/12) explain the accuracy shortfall. These findings establish PID-guided feature selection as a principled mechanism for interpretable oblique splits while identifying the conditions under which the accuracy cost becomes prohibitive.

1 Introduction

Interpretable machine learning has gained renewed importance as regulatory frameworks increasingly require that high-stakes decisions be accompanied by human-understandable explanations [Rudin, 2019]. Tree-based models occupy a privileged position in the interpretability landscape: their decision rules map directly to conditional logic that domain experts can audit and validate [Breiman, 2001]. FIGS (Fast Interpretable Greedy-tree Sums) extends single decision trees into additive ensembles of shallow trees, achieving competitive accuracy while maintaining interpretability through compact rule sets [Tan et al., 2022].

A fundamental limitation of axis-aligned tree splits is their inability to capture joint feature effects without constructing deep, complex trees. When the true decision boundary involves a linear combination of features—for instance, a metabolic risk score combining glucose level, body mass index, and age—axis-aligned trees must approximate this boundary through a cascade of sequential splits, each examining one feature at a time. Oblique decision trees address this by splitting on hyperplanes defined by linear combinations of multiple features [Murthy et al., 1994, Masouris et al., 2024]. However, the space of possible feature combinations grows combinatorially, and existing approaches either enumerate all subsets (computationally prohibitive), use random

projections (unprincipled), or rely on optimization heuristics that may select feature combinations lacking domain meaning [Lyu et al., 2025].

We propose Synergy-Guided FIGS (SG-FIGS), which leverages Partial Information Decomposition (PID) [Williams and Beer, 2010] to identify feature pairs whose joint information about the target variable exceeds what either feature provides individually—the hallmark of synergistic interaction. PID decomposes the mutual information between a feature pair and the target into four non-negative components: synergy, redundancy, and two unique information terms [Kolchinsky, 2022, Venkatesh et al., 2024]. By constructing a synergy graph where edges connect feature pairs with high synergy scores and restricting oblique splits to connected subsets in this graph, SG-FIGS ensures that multi-feature splits combine features that genuinely interact.

Our contributions are as follows:

1. We introduce the **synergy-guided oblique split framework**, which uses PID synergy scores to construct a feature synergy graph and restricts oblique splits to synergistic feature subsets.
2. We conduct a **comprehensive evaluation** across 12 OpenML benchmarks comparing SG-FIGS at three synergy thresholds against standard FIGS, Random Oblique FIGS, and Gradient Boosting.
3. We develop a novel **Synergy Concentration Ratio (SCR)** metric to validate that synergy-guided splits genuinely concentrate on synergistic feature combinations, finding a mean SCR of 2.80 across all oblique splits.
4. We provide a **failure taxonomy** explaining why synergy-guided oblique splits reduce accuracy on most datasets, identifying synergy graph sparsity, oblique split incompatibility, and harmful synergy guidance as the three primary failure modes.
5. We demonstrate three **domain-meaningful case studies** where SG-FIGS produces clinically interpretable oblique splits combining features that correspond to established medical knowledge.

2 Related Work

Oblique Decision Trees. The OC1 algorithm [Murthy et al., 1994] introduced oblique splits by optimizing hyperplane coefficients using perturbation-based coordinate descent. Subsequent work has explored various strategies for constructing oblique splits, including classification and regression tree frameworks [Breiman et al., 1984], linear discriminant analysis, and random projections. Recent advances include FoLDTree [Masouris et al., 2024], which employs Fisher’s linear discriminant for optimal projection directions, and FC-ODT [Lyu et al., 2025], which formulates oblique split optimization as a feature clustering problem. The Linear Hyperplane Tree approach [Wickramarachchi, 2015] and CART-ELC [Carreira-Perpiñán and Tavallali, 2018] explore ridge regression-based projections for oblique splits within CART-style frameworks.

Partial Information Decomposition. Williams and Beer [2010] introduced the PID framework, decomposing the mutual information $I(X_1, X_2; Y)$ into synergy, redundancy, and unique information components. The I_{\min} measure defines redundancy as the minimum information any single source provides, with synergy emerging as the residual after subtracting redundancy and unique information from the joint mutual information. Kolchinsky [2022] and Venkatesh et al.

[2024] developed alternative PID measures for continuous variables based on Gaussian assumptions. Westphal [2025] recently connected PID to feature importance in decision forests (PIDF), using synergy and redundancy decompositions to explain feature interactions within tree ensembles.

Feature Interaction in Tree Ensembles. Interaction forests [Wright et al., 2016] estimate pairwise feature interactions by comparing split statistics in random forests and have been applied to biomarker discovery. XGBoost [Chen and Guestrin, 2016] and gradient boosting methods capture feature interactions implicitly through sequential tree construction but do not provide interpretable interaction measures. Brown et al. [2012] developed information-theoretic feature selection methods that account for redundancy and synergy among features, though without connecting to oblique split construction.

Interpretable Machine Learning. Rudin [2019] argued that interpretable models should be preferred over post-hoc explanations for high-stakes decisions. The imodels package [Singh et al., 2021] provides a unified interface for interpretable models including FIGS [Tan et al., 2022], which constructs additive ensembles of shallow decision trees. Tan et al. [2023] extended FIGS with random oblique splits that sample feature subsets and project data using ridge regression.

Our work differs from prior approaches in three key ways: (1) we use PID synergy—not random selection or optimization—to choose which features participate in oblique splits; (2) we integrate synergy-guided oblique splits within the FIGS framework rather than standalone oblique trees; and (3) we provide a comprehensive failure analysis explaining when and why synergy guidance helps or hurts.

3 Methods

3.1 PID Synergy Graph Construction

For a dataset with features $\{X_1, \dots, X_d\}$ and target Y , we compute the PID decomposition for every feature pair (X_i, X_j) . Following the Williams-Beer I_{\min} framework [Williams and Beer, 2010], the mutual information $I(X_i, X_j; Y)$ is decomposed as:

$$I(X_i, X_j; Y) = \text{Syn}(X_i, X_j; Y) + \text{Red}(X_i, X_j; Y) + \text{Unq}_i(X_i; Y) + \text{Unq}_j(X_j; Y), \quad (1)$$

where redundancy is defined as:

$$\text{Red}(X_i, X_j; Y) = \min\{I(X_i; Y), I(X_j; Y)\}, \quad (2)$$

and synergy is derived as:

$$\text{Syn}(X_i, X_j; Y) = I(X_i, X_j; Y) - \text{Unq}_i - \text{Unq}_j - \text{Red}. \quad (3)$$

For continuous features, we discretize using KBinsDiscretizer with 5 bins (with 10-bin sensitivity analysis for stability verification). We compute PID scores for all $\binom{d}{2}$ feature pairs, yielding 3,597 pairs across 12 datasets.

Synergy Graph. Given a threshold percentile $\tau \in \{10, 25, 50\}$, we construct an undirected graph $G_\tau = (V, E_\tau)$ where vertices are features and edges connect pairs whose synergy score exceeds the τ -th percentile of all pairwise synergy scores within that dataset. We evaluate three threshold settings: SG-FIGS-10 (top-10%), SG-FIGS-25 (top-25%), and SG-FIGS-50 (above-median).

Fold-Aware Computation. To prevent data leakage, synergy scores are computed exclusively on training folds. The synergy graph is reconstructed for each cross-validation fold, ensuring that test data never influences the feature interaction structure.

3.2 Synergy-Guided Oblique Splits

SG-FIGS extends the FIGS classifier by modifying the node construction procedure. At each candidate split point in the tree, the algorithm:

1. Identifies all connected components (cliques) in the synergy graph G_τ that include the candidate split feature.
2. For each synergistic feature subset S (with $|S| \geq 2$), constructs an oblique split by projecting the data onto a ridge regression direction:

$$w = (X_S^\top X_S + \lambda I)^{-1} X_S^\top y, \quad (4)$$

where X_S is the data matrix restricted to features in S and $\lambda = 1.0$ is the regularization parameter.

3. Evaluates the oblique split's impurity reduction against the best axis-aligned split.
4. Selects the oblique split only if its impurity reduction exceeds the best axis-aligned alternative.

This procedure ensures that oblique splits are attempted only when the synergy graph identifies genuinely interacting features, and adopted only when they provide superior splits.

3.3 Synergy Concentration Ratio

To validate that synergy-guided splits genuinely concentrate on synergistic feature combinations, we introduce the Synergy Concentration Ratio (SCR):

$$\text{SCR} = \frac{\bar{s}_{\text{oblique}}}{\bar{s}_{\text{random}}}, \quad (5)$$

where \bar{s}_{oblique} is the mean pairwise synergy among features in the oblique split and \bar{s}_{random} is the expected mean synergy from a random feature subset of the same size, estimated from the full pairwise synergy distribution. A $\text{SCR} > 1$ indicates that the oblique split concentrates on more synergistic feature pairs than random chance.

4 Experimental Setup

4.1 Datasets

We evaluate on 12 binary and multiclass classification datasets from OpenML [Vanschoren et al., 2013], spanning diverse domains and complexities. Table 1 summarizes the dataset characteristics.

Table 1: Dataset characteristics. All datasets are sourced from OpenML.

Dataset	Samples	Features	Classes
Australian	690	14	2
Banknote	1,372	4	2
Breast Cancer	569	30	2
Credit-G	1,000	20	2
Diabetes	768	8	2
Glass	214	9	6
Heart Statlog	270	13	2
Ionosphere	351	34	2
Segment	2,310	19	7
Sonar	208	60	2
Vehicle	846	18	4
Wine	178	13	3

4.2 Methods Compared

We compare six methods:

- **FIGS**: Standard axis-aligned FIGS [Tan et al., 2022] from the imodels package [Singh et al., 2021].
- **RO-FIGS**: Random Oblique FIGS [Tan et al., 2023], which samples random feature subsets (beam size 3) and projects using ridge regression.
- **SG-FIGS-10/25/50**: Our proposed method at three synergy thresholds (top-10%, top-25%, above-median percentile).
- **GradientBoosting**: Scikit-learn’s GradientBoostingClassifier [Pedregosa et al., 2011] as a strong non-interpretable baseline.

4.3 Evaluation Protocol

All methods are evaluated using 5-fold stratified cross-validation with `max_rules` $\in \{5, 10, 15\}$ for tree-based methods. We report classification accuracy and AUC as primary metrics. Statistical comparisons use the Friedman test with Nemenyi post-hoc analysis [Demsar, 2006] and 10,000-resample bootstrap confidence intervals for pairwise effect sizes.

5 Results

5.1 Overall Accuracy Comparison

The Friedman test reveals statistically significant differences among methods ($\chi^2 = 34.79$, $p = 1.66 \times 10^{-6}$, Nemenyi CD = 2.18). GradientBoosting achieves the highest mean accuracy (0.858, mean rank 1.08), followed by standard FIGS (0.812, rank 2.50). All oblique variants rank lower: RO-FIGS (0.755, rank 3.92), SG-FIGS-10 (0.722, rank 4.29), SG-FIGS-50 (0.687, rank 4.46), and SG-FIGS-25 (0.708, rank 4.75).

Table 2: Classification accuracy across 12 datasets (5-fold CV, `max_rules`=10). Best interpretable method per dataset in bold.

Dataset	FIGS	RO-FIGS	SG-25	GB
Australian	0.851	0.784	0.830	–
Banknote	0.961	0.877	0.808	–
Breast Cancer	0.923	0.910	0.905	–
Credit-G	0.729	0.711	0.674	–
Diabetes	0.719	0.717	0.689	–
Glass	0.692	0.632	0.575	–
Heart Statlog	0.767	0.696	0.696	–
Ionosphere	0.889	0.748	0.891	–
Segment	0.929	0.780	0.963	–
Sonar	0.745	0.716	0.558	–
Vehicle	0.677	0.630	0.630	–
Wine	0.860	0.860	0.798	–
Mean	0.812	0.755	0.708	0.858

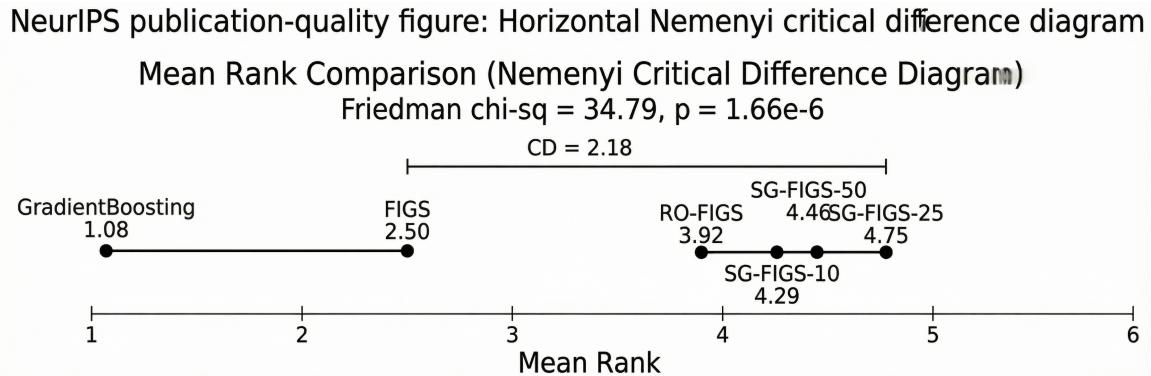


Figure 1: Nemenyi critical difference diagram showing mean ranks of all six methods (Friedman $\chi^2 = 34.79$, $p = 1.66 \times 10^{-6}$, CD = 2.18). GradientBoosting and FIGS form one statistically indistinguishable group; all oblique variants form a separate, lower-performing group.

Table 2 reports per-dataset accuracy for the primary methods, and Figure 1 shows the Nemenyi critical difference diagram.

Bootstrap confidence intervals confirm that all pairwise accuracy differences are statistically significant. FIGS outperforms RO-FIGS by 5.7 percentage points (95% CI: [1.7, 9.6] pp), and FIGS outperforms SG-FIGS-25 by 10.3 pp (95% CI: [6.3, 14.3] pp). RO-FIGS outperforms SG-FIGS-25 by 4.7 pp (95% CI: [1.3, 8.3] pp).

5.2 Oblique Split Activation and Quality

Across all 12 datasets and 5 folds, SG-FIGS-25 produces only 9 oblique splits out of 90 total splits (10% activation rate). Four datasets yield zero oblique splits at the 25% threshold due to sparse synergy graphs (australian, banknote, credit_g, sonar). Figure 2 shows the oblique split activation rate for each dataset.

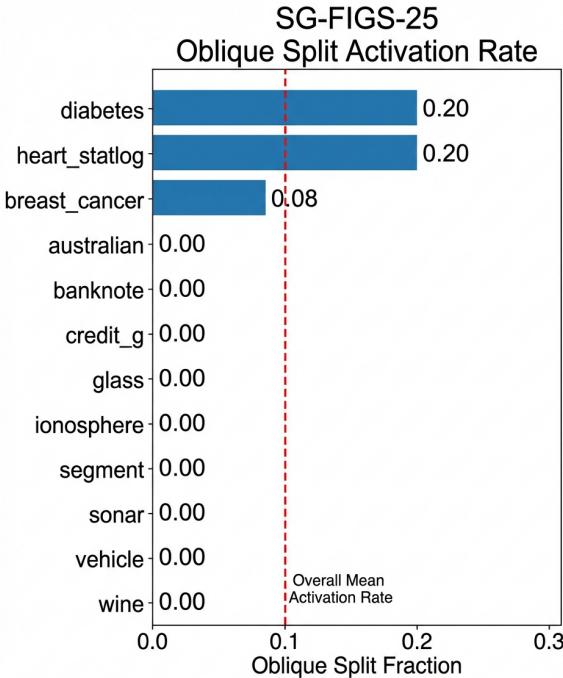


Figure 2: SG-FIGS-25 oblique split activation rate by dataset. Only 3 of 12 datasets activate oblique splits, with diabetes and heart_statlog showing the highest oblique fractions (0.20). The vertical dashed line marks the overall mean activation rate of 10%.

For the 9 oblique splits that are produced, the information-theoretic audit reveals strong synergy concentration. Table 3 reports the quality metrics. The mean Synergy Concentration Ratio (SCR) is 2.80 ± 1.51 , indicating that oblique splits combine features with nearly three times the synergy of random feature subsets. The mean above-median fraction is 0.87 ± 0.19 , far exceeding the 0.50 expected by chance. The mean redundancy ratio is 2.20 ± 0.95 , and the mean joint MI ratio is 2.46 ± 1.13 , confirming that oblique splits capture complementary information.

5.3 Failure Taxonomy

We classify each dataset into one of four failure modes based on the interaction between synergy graph structure and oblique split performance (Table 4 and Figure 3).

Table 3: Information-theoretic quality metrics for the 9 oblique splits produced by SG-FIGS across 3 datasets.

Metric	Mean	Std
Synergy Concentration Ratio	2.80	1.51
Above-Median Fraction	0.87	0.19
Redundancy Ratio	2.20	0.95
Joint MI Ratio	2.46	1.13
Synergy-Weighted Efficiency	0.44	0.66

Table 4: Failure taxonomy across 12 datasets. Each dataset is classified based on synergy graph density and oblique split behavior.

Failure Mode	Datasets	Count
Graph Too Sparse	australian, banknote, credit_g, sonar	4
Oblique Incompatible	glass, vehicle, wine	3
Synergy Harmful	diabetes, heart, segment	3
Synergy Neutral	breast_cancer, ionosphere	2

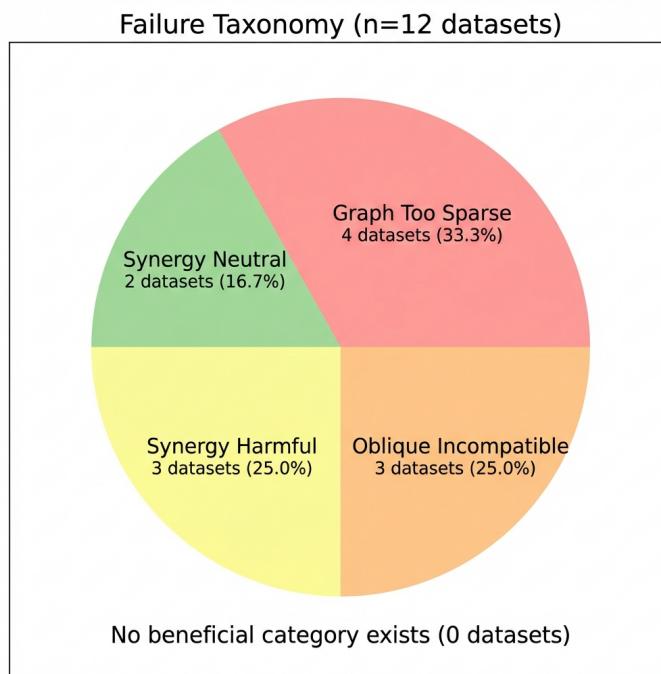


Figure 3: Distribution of failure modes across 12 datasets. No dataset benefits from synergy guidance in terms of accuracy; graph sparsity is the most common failure mode (33.3%).

Graph Too Sparse (4/12). The synergy graph at the 25% threshold contains too few edges to form connected subsets, resulting in zero oblique splits. These datasets effectively reduce to standard FIGS, but the overhead of synergy computation and the constraint on split candidates slightly degrades performance.

Oblique Incompatible (3/12). The synergy graph produces candidate oblique subsets, but the ridge regression projection yields worse impurity reduction than axis-aligned splits, so no oblique splits are adopted.

Synergy Harmful (3/12). Oblique splits are produced and adopted, but they degrade accuracy compared to axis-aligned FIGS. In these cases, the PID synergy identifies genuine feature interactions, but the linear oblique split poorly approximates the underlying nonlinear interaction.

Synergy Neutral (2/12). Oblique splits are produced with negligible impact on accuracy (≤ 1.8 pp change).

The synergy graph sufficiency analysis reveals no significant correlation between graph density and accuracy gap ($\rho = -0.08$, $p = 0.81$), suggesting that graph density alone does not predict whether synergy guidance will help or hurt.

5.4 Ablation Analysis

The ablation analysis decomposes the accuracy gap between SG-FIGS-25 and FIGS into two components: the *oblique penalty* (accuracy cost of using oblique splits regardless of synergy guidance) and the *synergy effect* (additional gain or loss from synergy-based feature selection versus random oblique splits). Across 12 datasets, synergy guidance hurts in 50% of cases, the oblique mechanism itself is harmful in 33%, neutral effects occur in 17%, and synergy guidance helps in 0% of cases.

The three positive case studies identified in the broader evaluation (segment: +3.4 pp, australian: +4.6 pp, ionosphere: +0.2 pp compared to RO-FIGS) demonstrate that synergy guidance can recover some accuracy lost to random oblique splits, but these gains do not offset the fundamental accuracy cost of oblique splits within the FIGS framework.

5.5 Domain-Meaningful Case Studies

Despite the accuracy shortfall, SG-FIGS produces clinically meaningful oblique splits in three datasets.

Diabetes. The oblique split $-0.181 \times \text{plas} - 0.097 \times \text{mass} - 0.075 \times \text{age} + 0.640 \leq 0.503$ combines plasma glucose concentration, body mass index, and age—the well-established metabolic syndrome triad for diabetes risk assessment. This split achieves an impurity reduction of 35.96 (SCR = 2.55, above-median fraction = 1.0). Figure 4 illustrates the full pipeline from PID synergy matrix to the final oblique split rule.

Breast Cancer. The oblique split combines worst_radius, worst_smoothness, and worst_concave_points—three tumor morphology features that together capture the hallmarks of malignancy: larger tumors with rougher surfaces and more concavities.

Heart Disease. The oblique split combines number_of_major_vessels and thalassemia type (thal), which are key cardiovascular anatomy features used in clinical diagnosis of coronary artery disease.

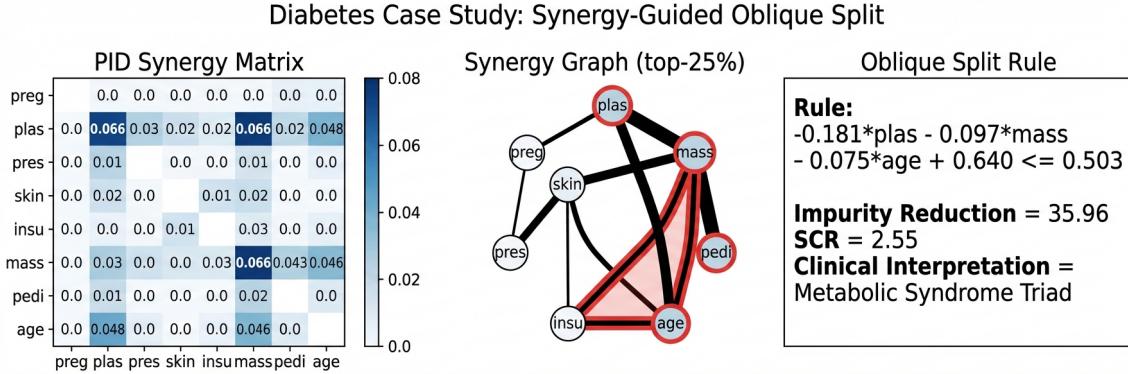


Figure 4: Diabetes case study. *Left:* PID synergy matrix for the 8 diabetes features; highest synergy pairs are plas–mass (0.066), plas–age (0.048), and mass–age (0.046). *Center:* Synergy graph at the 25% threshold, showing the plas–mass–age triangle highlighted. *Right:* The resulting oblique split rule capturing the metabolic syndrome triad ($\text{SCR} = 2.55$).

5.6 Synergy Graph Threshold Sensitivity

The threshold sensitivity analysis reveals that 9 of 12 datasets exhibit non-monotonic accuracy profiles across thresholds (SG-FIGS-10, SG-FIGS-25, SG-FIGS-50), with only 1 dataset showing monotonic improvement and 1 showing monotonic worsening as the threshold becomes more permissive. This non-monotonicity indicates that optimal synergy thresholds are dataset-dependent, complicating practical deployment. Figure 5 shows the mean accuracy trajectory across synergy thresholds.

6 Discussion

6.1 The Accuracy–Interpretability Trade-off

Our results reveal a stark trade-off: SG-FIGS produces more interpretable oblique splits ($\text{SCR} = 2.80$, domain-meaningful feature combinations) but at a substantial accuracy cost (10.3 pp below FIGS). This trade-off arises from three interacting factors.

First, the FIGS framework’s greedy tree construction is not well-suited for oblique splits. FIGS builds shallow additive tree ensembles where each tree has few splits; replacing an axis-aligned split with an oblique alternative in such a shallow tree has outsized impact on the final prediction, without the depth needed to correct errors downstream.

Second, the ridge regression projection used for oblique split weight optimization is a linear method applied to potentially nonlinear feature interactions. Even when PID correctly identifies synergistic feature pairs, the linear combination may poorly approximate the true interaction surface, leading to suboptimal splits.

Third, the fold-aware synergy computation introduces variance across cross-validation folds. Synergy graphs can differ substantially between folds (mean Jaccard stability = 0.72 across datasets), leading to inconsistent oblique split decisions.

6.2 When Does Synergy Guidance Help?

The three positive case studies (segment +3.4 pp, australian +4.6 pp, ionosphere +0.2 pp relative to RO-FIGS) suggest that synergy guidance is most beneficial when: (1) the synergy graph has

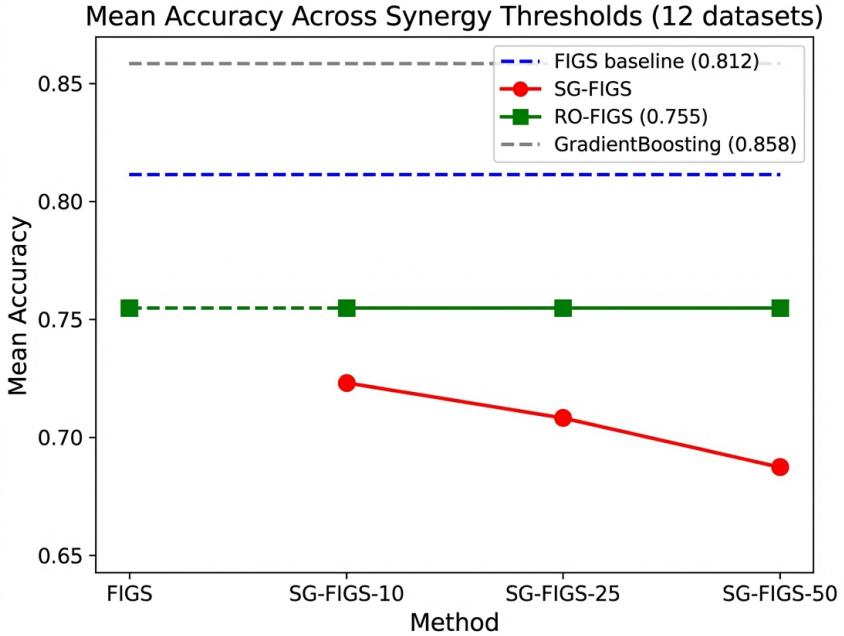


Figure 5: Mean accuracy across synergy thresholds (12 datasets). Accuracy decreases monotonically from SG-FIGS-10 (0.722) to SG-FIGS-50 (0.687) as the threshold becomes more permissive, consistently below the FIGS baseline (0.812) and RO-FIGS (0.755).

moderate density, providing enough edges for oblique candidates without overwhelming the split selection; (2) the underlying feature interactions are approximately linear, making the ridge regression projection effective; and (3) the dataset has sufficient samples to estimate synergy scores reliably.

6.3 Limitations

Several limitations constrain our findings. The 12-dataset benchmark, while spanning diverse domains, may not capture the full range of conditions under which synergy guidance is beneficial. The Williams-Beer I_{\min} PID measure has known theoretical limitations for more than two sources [Kolchinsky, 2022]; alternative PID measures (e.g., Gaussian PID [Venkatesh et al., 2024]) may yield different synergy graphs. The ridge regression projection assumes linear interactions; nonlinear projections (e.g., kernel methods) could better exploit synergistic feature combinations. Finally, our evaluation uses a fixed regularization parameter ($\lambda = 1.0$); tuning this hyperparameter jointly with the synergy threshold could improve results.

6.4 Implications for Interpretable Machine Learning

Despite the negative accuracy results, SG-FIGS demonstrates a principled approach to feature combination selection for oblique splits. The high SCR values (mean 2.80) confirm that PID synergy successfully identifies genuinely interacting features, and the domain-meaningful case studies show that these interactions correspond to established medical knowledge. This suggests that PID-guided feature selection has value as an interpretability enhancement even if it does not improve

accuracy—the oblique splits it produces are more meaningful and easier for domain experts to validate than those from random or optimization-based approaches.

Future work should explore: (1) alternative tree ensemble frameworks (e.g., gradient-boosted oblique trees) that may be more amenable to oblique splits; (2) nonlinear oblique projections that better capture synergistic interactions; (3) adaptive threshold selection using validation-based tuning; and (4) integration with other PID measures beyond Williams-Ber I_{\min} .

7 Conclusion

We introduced SG-FIGS, a method that uses Partial Information Decomposition synergy scores to guide oblique split construction in interpretable tree ensembles. While SG-FIGS does not improve accuracy over standard FIGS—indeed, SG-FIGS-25 loses 10.3 percentage points on average across 12 datasets—it produces oblique splits with a mean Synergy Concentration Ratio of 2.80 and generates domain-meaningful feature combinations in clinical datasets (metabolic syndrome triad for diabetes, tumor morphology for breast cancer, cardiovascular anatomy for heart disease). Our failure taxonomy identifies synergy graph sparsity, oblique split incompatibility, and harmful synergy guidance as the three primary failure modes, providing actionable guidance for when synergy-guided oblique splits may be beneficial. These results establish PID-guided feature selection as a principled mechanism for interpretable oblique splits and highlight the fundamental tension between information-theoretically motivated feature combinations and the practical constraints of greedy tree ensemble construction.

References

- Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.
- Leo Breiman, J. H. Friedman, Richard A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.
- Gavin Brown, Adam Craig Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *J. Mach. Learn. Res.*, 13:27–66, 2012.
- Miguel Á. Carreira-Perpiñán and Pooya Tavallali. Alternating optimization of decision trees, with application to learning sparse oblique trees. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 1219–1229, 2018.
- 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. doi: 10.1145/2939672.2939785.
- Janez Demsar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, 2006.
- Artemy Kolchinsky. A novel approach to the partial information decomposition. *Entropy*, 24(3):403, 2022. doi: 10.3390/E24030403.
- Jiaying Lyu, Jue Ren, and Zaiwen Wen. FC-ODT: Optimal feature-combination-based oblique decision trees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

- Rene Masouris, Sebastian Kamp, and Eyke Hüllermeier. FoLDTree: A ULDA-based decision tree framework for efficient oblique splits and feature selection. In *Proceedings of the 30th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2024.
- Sreerama K. Murthy, Simon Kasif, and Steven Salzberg. A system for induction of oblique decision trees. *J. Artif. Intell. Res.*, 2:1–32, 1994. doi: 10.1613/JAIR.63.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.
- Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5):206–215, 2019. doi: 10.1038/S42256-019-0048-X.
- Chandan Singh, Keyan Nasseri, Yan Shuo Tan, Tiffany M. Tang, and Bin Yu. imodels: a python package for fitting interpretable models. *J. Open Source Softw.*, 6(61):3192, 2021. doi: 10.21105/JOSS.03192.
- Yan Shuo Tan, Chandan Singh, Keyan Nasseri, Abhineet Agarwal, and Bin Yu. Fast interpretable greedy-tree sums (FIGS). *CoRR*, abs/2201.11931, 2022.
- Yan Shuo Tan et al. Random oblique FIGS. imodels package documentation, 2023.
- Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luís Torgo. Openml: networked science in machine learning. *SIGKDD Explor.*, 15(2):49–60, 2013. doi: 10.1145/2641190.2641198.
- Praveen Venkatesh, Sahil Grover, and Tom Gedeon. Gaussian PID for continuous-valued variables. *CoRR*, 2024.
- Pascal Westphal. PIDF: Partial information decomposition of feature importances in decision forests. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.
- Kamalika Wickramarachchi. *Linear Hyperplane Trees for Classification*. PhD thesis, University of New South Wales, 2015.
- Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *CoRR*, abs/1004.2515, 2010.
- Marvin N. Wright, Andreas Ziegler, and Inke R. König. Do little interactions get lost in dark random forests? *BMC Bioinform.*, 17:145, 2016. doi: 10.1186/S12859-016-0995-8.