Beyond Binary Prediction: Calibrated Risk Stratification for Pattern Discovery

WANG Yuqi
Department of Computing, PolyU
[redacted]@connect.polyu.hk

Abstract

Most heart attack analysis seen on Kaggle treats the task as binary classification, optimizing for accuracy at the expense of calibration and interpretability. To address this, I developed a unified framework beyond binary predictions, via three interconnected stages: (1) **calibrated classification** with vectorized Monte Carlo perturbation and calibration error minimizing ensemble, (2) **manifold clustering** of Tabular Foundation Model (TFM) embeddings residualized against calibrated logits, and (3) **pattern discovery** using association rule mining and subgroup discovery. Experimenting on the UCI Cleveland heart disease dataset demonstrates superior pattern discovery capacity of my cohesive analytical pipeline when compared against treating classification, clustering, and mining as isolated tasks. The core insight is that calibrated estimates should serve not as terminal outputs, but as structural priors that guide downstream manifold analysis and pattern discovery.

1 Introduction

Project Highlights

1. Unified three-stage pipeline with principled information flow (*Section 3*)
2. Vectorized Monte Carlo perturbation (100×100 in 30 min) (*Section 3.1*)
3. Calibration-optimized ensemble minimizing ECE^{KDE} (*Section 3.1*)
4. TFM embedding residualization via RBF kernel regression (*Section 3.2*)
5. Risk-stratified mining outperforms binary-label baselines (*Section 3.3*)
6. LLM-as-Judge evaluation with 3 frontier models (*Table 4*)

Notably: I astutely identified that the original dataset labels were *incorrect*, revealing data quality issues in the Kaggle dataset. See (*Section 4*) for more details.

Machine Learning (ML) methods have been extensively applied to the Cleveland Heart Disease dataset on Kaggle. However, these methods approach healthcare analysis as a binary classification task, optimizing for accuracy metrics that obscures the continous nature of disease risk, and thus, neglecting the calibration necessary for clinical decision-making. This methodological gap is even more noticeable when practitioners seek insights beyond binary predictions, but understanding of the underlying risk patterns for subgroup identification and actionable early interventions.

For this project, a naive approach is to treat classification, clustering, and mining as orthogonal tasks. A typical workflow might start from training a classifier, then separately a clustering algorithm for performance comparisons, and finally Association Rule Mining (ARM) for pattern discovery and

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

interpretability. In such approach, each stage operates in isolation with minimal to no information flow between modules. While convenient, such disjoint approach suffers three limitations:

1. Binary predictions discard continuous information inherent in disease risk.
2. Raw feature space failed to capture non-linear relationships and nuanced topological structure.
3. Discovered patterns are not grounded in calibrated risk assessments.

Project Highlight To overcome the aforementioned challenges, I designed a unified framework through *principled information propagation* across the three interconnected stages. Instead of treating each as isolated tasks, I architect a schematic pipeline where each stage enriches the next.

1. **Calibrated Classification (§3.1)** A vectorized Monte Carlo perturbation protocol is developed to stress-test model stability under 100 perturbation levels, each sampled 100 times with full vectorization in a single-pass. During which, the Expected Calibration Error via Kernel Density Estimation (ECE^{KDE}) and brier score are calculated. This helps identify candidate models with high robustness and calibration. Next, grid search is performed over ensemble weights of these candidate models to produce $\sum_i w_i \cdot \hat{p}(y_i = 1|x)$ that minimizes ECE^{KDE}. This yields a risk proxy more nuanced and informative than simple binary labels.

2. **Manifold Clustering (§3.2)** Recognizing the rich geometric structure encoded by Tabular Foundation Model (TFM) embeddings, we extract TabPFN embeddings and orthogonalize it against calibrated logits using Radial Basis Function (RBF) Kernel Ridge Regression. Clustering on these residualized embeddings reveals subpopulations that prompts further subgroup analysis. By doing so, I was able to shift an otherwise uninformative benchmark comparisons between classification and clustering models into a hypothesis generation process for pattern discovery.

3. **Pattern Discovery (§3.3)** Calibrated probabilities are first converted into logits, then discretized into risk strata. Next, subgroup discovery algorithms alongside association rule mining are employed. Through anchoring pattern discovery to calibrated risk rather than binary labels, I was able to uncover more nuanced interpretable rules.

The key insight of my approach is that information must flow *progressively*: calibrated probabilities from §3.1 guide analysis in §3.2 , which in turn generates hypotheses for pattern discovery in §3.3 . Experiments on UCI Cleveland dataset ($n = 303$, $k = 13$) reveals that my framework was able to identify rules and patterns that would otherwise be missed by disjoint approach that directly run association rule mining. Experiments confirm that my former approach results in statistically richer discoveries that aligns with medical priors. Though the experiments were confined to a single small sized dataset (an inevitable constraint for a course project) the methodological insights might offer some transferrable insights.

The remainder of this report is organized as follows: §2 review preliminaries for this project, §3 details my way to solution and §4 details some initial Exploratory Data Analysis (EDA) and transformation made to the dataset.

2 Preliminaries

2.1 Tabular Foundation Models

Recent works in deep learning challenge classical machine learning methods through large-scale pretraining. TabPFN [1], short for Tabular Prior-Data Fitted Network, address the previous limitations of deep learning based approaches by pretraining on large synthetic distributions of tabular tasks. They are training-free, meta-learned and perform inference via in-context learning. Later work, such as TabICL [2] and Mitra [3], extends and scales up this method through retrieval-augmented and metric-learning approaches. Most recently, LimiX [4] introduces a unified family of Large Structured-Data Models (LDMs) that treats tabular data as a joint distribution over variables and missingness.

TabPFN TabPFN amortizes the often intractable Bayesian prediction for tabular datasets by learning an explicit one-step transformation from labelled training data to test data class probabilities. Given data $D = \{(x_i, y_i)\}_{i=1}^{n}$, a query x', the Bayesian target is the posterior predictive

$$p(y'|x', D) = \int p(y'|x'\varphi)p(D|\varphi)p(\varphi)\,\mathrm{d}f\,\mathrm{d}\theta \tag{1}$$

Here, φ ranges over **all data-generating mechanisms**. This integral essentially averages prediction over all plausible sets of underlying rules and mechanisms, weighted by how well each explains the observed data D. In practice, TabPFN approximates this intractable integral with a permutation-invariant Transformer trained across many synthetic tabular datasets, too, generated from such mechanisms, so that its one-shot an output approximates the integral above [1].

LimiX Contrary to TabPFN which is architectually designed to optimize for directrional supervised mapping $p(y'|x', D)$, LimiX disregards column nature and treat the entire dataset as a joint distribution over all variables, including output label and missingness [4]. Instead of amortizing a specific posterior predictive as in TabPFN, LimiX approximates joint density of **the entire table**, enabling querying any subset of features, which is prefect for rule mining Formally, given a dataset partitioned into an in-context subset D_c and query subset D_q, and a mask $\pi \subseteq [d]$, then the masked and observed subvectors of the query samples are D_q^{π} and $D_q^{-\pi}$, respectively. Then, LimiX minimizes the NLL:

$$\mathcal{L}_k(\theta) = -\mathbb{E}_{(D_c, D_q)\sim p, \pi\sim \mathrm{Unif}(\Pi_k)}\left[-\log q_\theta\big(D_q^{\pi}|D_q^{-\pi}, D_c\big)\right] \tag{2}$$

2.2 Calibration Error

Calibration ensures that the model's estimated probabilities are faithful and match real-world likelihoods, which is cruicial for sensitive applications like heart attack analysis. A model is considered confidence-calibrated if, for all confidence levels $\alpha \in [0, 1]$, the model is correct on average at that confidence level (α proportion of the times) [5]. The L_p calibration error of f can be defined as:

$$L_p(f) = \left(\mathbb{E}\left[\left\|\mathbb{E}[y \mid f(x)] - f(x)\right\|_p^p\right]\right)^{\frac{1}{p}} \tag{3}$$

Expected Calibration Error Expected Calibration Error (ECE) is a widely used measure of calibration. It computes the average calibration error across binned confidence levels by taking the absolute difference between average accuracy (acc) and average confidence (conf) [6].

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \cdot \Big|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\Big| \tag{4}$$

In this work, a variant of ECE called $\mathrm{ECE}^{\mathrm{KDE}}$ is used. Unlike its binned predecessor, $\mathrm{ECE}^{\mathrm{KDE}}$ utilizes a Beta kernel in binary classification and a Dirichlet kernel in multiclass setting, which provides a more consistent and lower bias estimate of the true calibration error L_p [7].

2.3 Subgroup Discovery

Subgroup Disocvery (SD) concerns identifying patterns that deviates significantly from the norm [8]. Unlike predictive modeling, SD identifies local subsets of the data D where the distribution of the target variabel y is statistically unusual. The objective is to find the top-k subgroups that maximizes some quality function $Q(S, y)$, where S is the subgroup.

Quality Function The quality function $Q(S, y)$ measures the interestingness of a subgroup S. Ideally, said function should balance subgroup size with the statistical significance of deviation. A common choice of quality function for binary targets is the *Weighted Relative Accuracy* (Q_{WRAcc}). It measures the trade-off between coverage and precision gain over the default probabilities.

$$Q_{\mathrm{WRAcc}}(S, y) = \frac{|S|}{n}\Big(P(y = 1|S) - P(y = 1)\Big) \tag{5}$$

For numeric targets, the common choice is the Standard Quality Function. Given the subgroup mean μ_S, global mean μ, and weighting parameter a, the quality is defined as:

$$Q_{\mathrm{std}}(S) = \left(\frac{|S|}{n}\right)^a (\mu_S - \mu) \tag{6}$$

3 Roadmap

This section details the three-stage analytical pipeline. Each stage is designed to address specific limiations of disjoint analysis while enriching subsequent stages with *principled information flow*.

Given a labelled dataset, the conventional approach is to optimize a classifier, independently apply clustering, and pattern mining algorithms. My framework, in contrast, connects the three stages:

$$\mathcal{D} \xrightarrow{\text{Stage 1}} \{\hat{p}_i\}_{i=1}^n \xrightarrow{\text{Stage 2}} \{\boldsymbol{z}_i^{\perp}\}_{i=1}^n \xrightarrow{\text{Stage 3}} \{\text{Rules}\} \tag{7}$$

where $\hat{p}_i \in [0,1]$ are calibrated risk probabilities, $\boldsymbol{z}_i^{\perp}$ are residualized embeddings, and Rules comprise mined patterns (e.g., through ARM or subgroup discovery). Intuitively, **later stages consume rich information from previous stages**. Specifically, calibrated risk probabilities guide embedding analysis (orthogonalization), which potentially reveals subgroup that could inform pattern discovery.

3.1 Calibrated Classification

	No Perturbation ($\ell = 0$)			Mid Perturbation ($\ell = 50$)			Max Perturbation ($\ell = 100$)		
Models	ACC	AUC-ROC	F1	ACC	AUC-ROC	F1	ACC	AUC-ROC	F1
RF	0.8307	0.8271	0.8055	0.7824	0.7802	0.7583	0.7244	0.7242	0.7047
KNN	0.8237	0.8236	0.8109	0.7720	0.7708	0.7514	0.7025	0.7011	0.6758
LimiX	0.8340	0.8311	0.8132	0.7906	0.7906	**0.7747**	0.7252	0.7282	**0.7171**
LogReg	**0.8443**	**0.8412**	**0.8225**	**0.7950**	**0.7931**	0.7725	**0.7329**	**0.7332**	0.7145
TabPFN	**0.8544**	**0.8492**	**0.8237**	**0.8169**	**0.8141**	**0.7912**	**0.7590**	**0.7583**	**0.7368**
XGBoost	0.7998	0.7975	0.7789	0.7503	0.7453	0.7127	0.6967	0.6940	0.6636

The first stage of the pipeline is called *Calibrated Classification*. Instead of directly optimizing for and selecting models by their accuracy score, the objective here is to select models that balances accuracy and reliability. Concretely, this can be achieved via three steps:

1. Select the top-k models in terms of accuracy measures.
2. Stress-test the model's calibration stability using a *Vectorized Monte Carlo Perturbation Protocol*.
3. Select the top performing models in the stability test and ensemble them to minimize ECE^{KDE}.

Step 1: Baseline Screening

Before evaluating model robustness, we first need screen them by their discriminative power (e.g., accuracy) under clean, unperturbed data distributions ($\ell = 0$). If a model is not able to achieve a reasonable accuracy, then high stability under perturbation is not meaningful.

Results from "No Perturbation ($\ell = 0$)" column in the table above shows that the best performing models are TabPFN, LimiX, and Logistic Regression.

Step 2: Vectorized Monte Carlo Perturbation Protocol

Systematically assessing model robustness requires a controlled data degradation process. Unlike traditional perturbation approach that add gaussian noise proportional to sample scale, my approach anchors the perturbation amplitude in feature distributions to ensure fairness between features. For each feature j, two perturbation mechanisms are defined:

Continous Features: given sample $x_{ij} \in \mathbb{R}$, global standard deviation of feature j, σ_j^{global}, and the perturbation magnitutde $\ell \in \{1, ..., L\}$, then the perturbation is sampled as:

$$\tilde{x}_{ij}^{(\ell)} = x_{ij} + \epsilon \cdot \sigma_j^{\text{global}},\ \epsilon \sim \mathcal{N}(0, \alpha_\ell^2) \tag{8}$$

Categorical Features: given categorical $x_{ij} \in \{c_1, ..., c_K\}$, with a categorical resample probability β_ℓ that increases with perturbation ℓ, the perturbed value $\tilde{x}_{ij}^{(\ell)}$ is sampled from the empirical marginal distribution of the feature j via the following process:

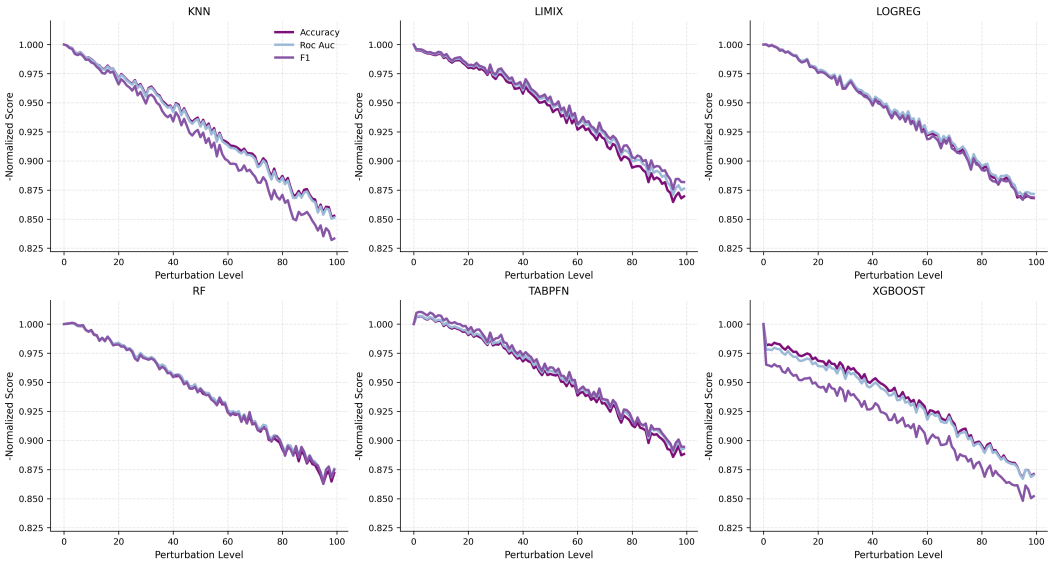


Figure 1: Stability test showing averaged cross validation performance on six different models across 100 perturbation levels. **Higher is better**. TabPFN, LimiX, Logistic Regression, and KNN are the most invariant to perturbation, with TabPFN being the strongest. Intuitively, since perturbation diffuses data points around an ellipsoid within the feature space, slower performance degradation implies potentially **larger decision boundary margin**, which is a strong signal of model robustness.

$$
\begin{aligned}
b_{ij}^{(\ell)} &\sim \text{Bernoulli}(\beta_\ell) \\
g_{ij} &\sim \text{Categorical}\left(\boldsymbol{p}_j^{\text{global}}\right) \\
\tilde{x}_{ij}^{(\ell)} &= \begin{cases} x_{ij}, & b_{ij}^{(\ell)} = 0 \\ g_{ij}, & b_{ij}^{(\ell)} = 1 \end{cases}
\end{aligned}
\tag{9}
$$

To ensure **statistical significance** under high variation, a high repetition count (M) is required. However, this introduces great computational demand. When implemented as a nested loop, this would take approximately 8 hours to complete. To avoid computational bottlenecks of nested loops, the sampling process is vectorized. For $L = 100$ perturbation levels and $M = 100$ Monte Carlo repetitions per level, construct tensor $\boldsymbol{X} \in \mathbb{R}^{(n \times M) \times d}$ to vectorize all M monte carlo repetitions into a single pass. This lowers runtime down to 30 minutes on a single RTX4090 GPU.

Step 3: Calibration-Optimized Ensemble

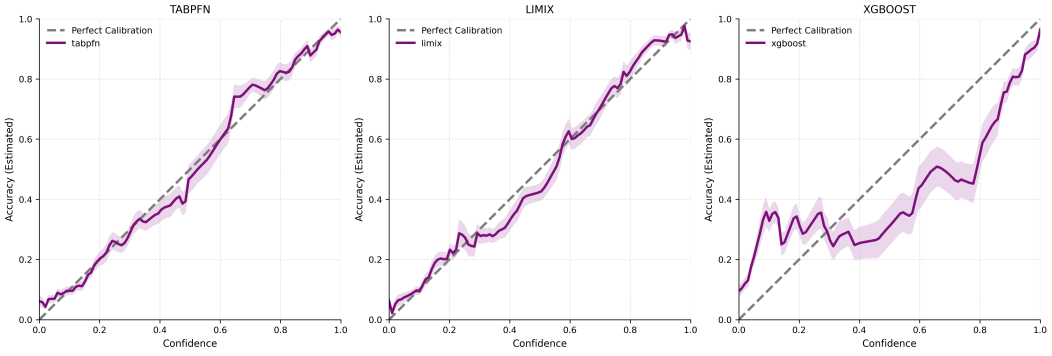


Figure 2: Calibration curves of the three out of the six models. As shown by the curves, TabPFN and LimiX are the most calibrated, with XGBoost being the least. A jagged calibration curve indicates overfitting, as the model is over- or under-confident with data points around the decision boundary.

Given top-k calibrated models $\{f_1, ..., f_k\}$ in the stability test, an esmeble with weights $\{w_1, ..., w_k\}$ that minimizes calibration error is constructed. Then, the ensmeble's prediction is defined as:

$$\hat{p}_{\text{ens}}(\mathbf{x}) = \sum_{i=1}^{k} w_i ... \hat{p}_{i(\mathbf{x})}, \quad \text{subjected to} \sum_{i=1}^{k} w_i = 1, w_i \geq 0 \tag{10}$$

To find the optimal weights $\mathbf{w}^* = \arg\min_{\mathbf{w}} \in \text{ECE}^{\text{KDE}}(\mathbf{w})$, a simple grid search is performed with 20 random train-validation splits, each split computing a 10-fold out-of-fold (OOF) predictions. The grid search aims to minimize the ECE^{KDE} of OOF predictions. An independent holdout testing set is also used to ensure that we are not overfitting the ensemble weights to the validation set. The final ensemble prediction is obtained by averaging the OOF prediction probabilities.

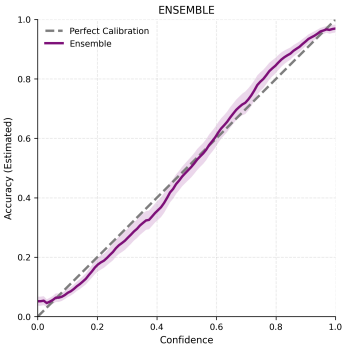


Figure 3: Calibration curve of the weighted ensemble. The resultant ensemble is extremely well-calibrated shown by smooth curve aligning with the perfect calibration line.

Theoretical Justification

Firstly, by screening for classification models with both high *accuracy* and high *perturbation resistance*, it gives a set of candidate predictors whose outputs are simutaneously *informative* and *structurally well-behaved*. Specifically, high *accuracy* indicates the decision boundaries captures a non-trivial portion of the signal in $p(y|x)$. High *perturbation resistance*, on the other hand, adds a qualitatively different constraint: under the Monte Carlo perturbation protocol, each data point x is replaced by a neighborhood distribution of plausible measurements (8) or a plausible coding (9). Therefore, a model that degrades slowly under this neighborhood is, in effect, one whose decision boundary is not precariously close to the training data points; in other words, such model tends to have **larger decision margins** and **locally smoother region** (smaller local Lipschitz constant) in areas where data clusters. This matters because, on small clinical tabular datasets, brittle model can appear low-biased and accurate, yet high-variance and sample-specific.

These shortlisted predictors, $\{\hat{p}_i(x)\}_i$, are then ensembled to form a single predictor, $\hat{p}_{\text{ens}}(x)$. Geometrically, the simplex constraint $\sum_i w_i = 1, w_i \geq 0$ confines the meta-model $\hat{p}_{\text{ens}}$ to a convex hull of base predictors $\hat{p}_i$. Formally, fix the evaluation set $\{x_1, ..., x_n\}$. Each shortlisted predictor creates a prediction vector $\mathbf{p}_i = (\hat{p}_i(x_1), ..., \hat{p}_i(x_n)) \in [0, 1]^n$. Then, the convex hull of the predictors is exactly the set of all convex combinations of the prediction vectors $\mathbf{p}_i$:

$$\text{conv}\{\mathbf{p}_1, ..., \mathbf{p}_k\} = \left\{ \sum_{i=1}^{k} w_i \mathbf{p}_i : \sum_{i=1}^{k} w_i = 1, w_i \geq 0 \right\} \tag{11}$$

So with $k = 2$, the convex hull is the line segment between the prediction vectors; with $k = 3$, it is a filled triangle whose vertices are the prediction vectors; with generalized k, it is a **polytope** inside $[0, 1]^n$. This forces the ensemble to satisfy the pointwise boundedness for every x:

$$\min_i \hat{p}_i(x) \ \leq \ \hat{p}_{\text{ens}}(x) \ \leq \ \max_i \hat{p}_i(x) \tag{12}$$

This implies *no extrapolation*. And since the robustness of base predictors are guaranteed by Step 1-2, the weighted ensemble can be described as a projection of the true (unknown) risk function onto the convex polytope spanned by a small set of already competent predictors. Essentially, a risk proxy.

3.2 Manifold Clustering

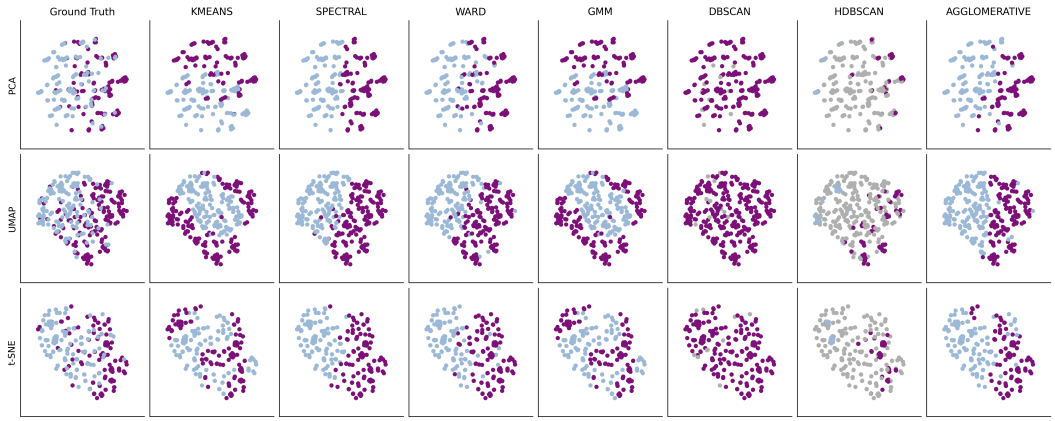


Figure 4: Comparison of the seven clustering algorithms, visualized by three different dimensionality reduction algorithms (PCA, UMAP, t-SNE). As shown by the scatter plot, Spectral, Ward, and Agglomerative are the best performing algorithms. Whereas, density-based methods like DBSCAN and HDBSCAN were unable to differentiate the two populations.

The second stage shifts focus from predictive modeling to **topological understanding** of the feature space. Rather than treating clutering as a competing paradigm to classification, I leverage it as a hypothesis generative tool for revealing interesting data patterns, such as high subpopulations.

Step 1: Clustering on Feature Space

Metrics	Kmeans	Spectral	Ward	GMM	DBSCAN	HDBSCAN	Agglomerative
AMI	0.0202	**0.2968**	0.2366	0.0186	0.0162	0.1489	0.2345
ARI	0.0276	**0.3843**	0.3122	0.0253	0.0087	0.0487	0.2972
ACC	0.5878	**0.8108**	0.7804	0.5845	0.5338	0.2432	0.7736

To satisfy project requirements, seven clustering algorithms are applied to l_2-normalized features. Three different metrics were used: Adjusted Mutual Information (AMI), Adjusted Rand Index (ARI), and Accuracy with post-hoc label matching. As shown in Figure 4 and the table above, most model performs no better, or even worse, than random guessing (Acc ≈ 0.5). The low ARI and AMI scores (≤ 0.1) scores substantiate this further, indicating that the clustering results is indistinguishable from random noise. The best performing algorithms (Spectral and WARD) only achieves Acc ≈ 0.8.

Such underwhelming performance is expected. Raw features rarely exhibits directly separable structures algined with supervised labels without feature engineering. However, this negative results motivates the next experiment: can learned representations capture more meaningful geometry?

Step 2: Clustering on Tabular Foundation Model Embeddings

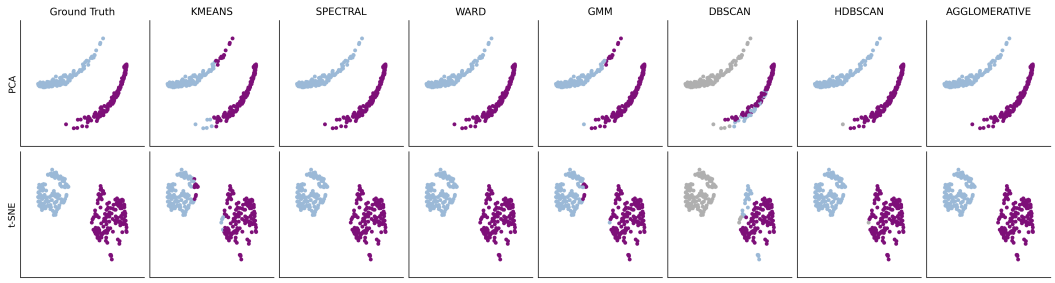


Figure 5: Comparison of the same seven clustering algorithms on TabPFN embeddings. Remarkably, all clustering algorithm except DBSCAN achieved near-perfect alginment with the output labels.

Tabular Foundation Models like TabPFN often encode rich semantic structure in their embeddings space. Motivated by this, I extracted embeddings $\mathbf{z}_i \in \mathbb{R}^{d'}$ from TabPFN's penultimate layer and apply the same procedure as in Step 1. The only variable here is swapping features for embeddings.

Metrics	Kmeans	Spectral	Ward	GMM	DBSCAN	HDBSCAN	Agglomerative
AMI	0.6473	**1.0000**	**1.0000**	0.7630	0.8130	0.9853	**1.0000**
ARI	0.7471	**1.0000**	**1.0000**	0.8315	0.8118	0.9927	**1.0000**
ACC	0.9324	**1.0000**	**1.0000**	0.9561	0.4358	0.9966	**1.0000**

While the results are impressive, they suffer from a methodological confound: TabPFN is trained with supervision, meaning its embedding are explicitly optimized to be separable for the two classes. Therefore, if we were to treat this result as evidence that "TabPFN enables unsupervised discovery", it would be blatant misinterpretation and circular reasoning. Instead, I view this experiment as **validation of embedding quality** that justifies the next stage (Step 3).

Step 3: Embedding Residualization

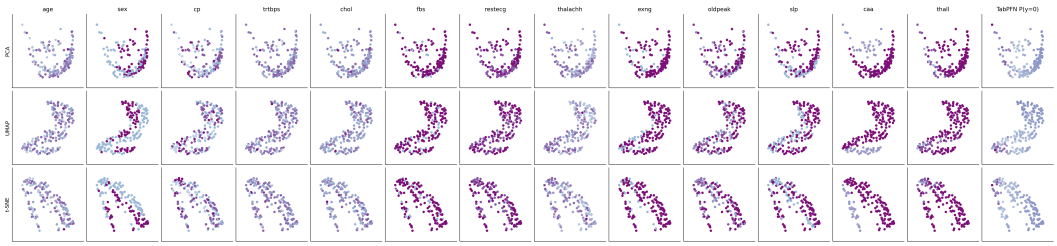


Figure 6: Color verification of TabPFN embeddings on the $y = 0$ subset. The manifold forms a horseshoe shape. Without residualization, the data points on the manifold are ordered by the model's prediction confidence $\hat{P}(y = 0)$. The smooth color gradient in the rightmost column suggests primary spatial variation along the horseshoe is directly correlated to output probability.

I start by partitioning embeddings by their class: $\mathbf{Z_0} = \{\mathbf{z}_i : y_i = 0\}$ and $\mathbf{Z_1} = \{\mathbf{z}_i : y_i = 1\}$. Then, I visualize the embeddings via dimensionality reduction (i.e., PCA, UMAP, t-SNE) and coloring points by feature values and TabPFN confidence level. As shown in Figure 6, broad color gradient across the horseshoe-shaped manifold can be seen, but discrete structures are lacking.

To address this, the calibrated probabilities $\hat{p}_{\text{ens}}(\mathbf{x}) \in [0, 1]^n$ are first converted into logits. The logits for sample x_i is denoted as:

$$\text{logits}(\hat{p}_{\text{ens}})_i = \log(\hat{p}_{\text{ens}}(x_i)) - \log(1 - \hat{p}_{\text{ens}}(x_i)) \tag{13}$$

Then, using Radial Basis Function (RBF) kernel regression, regress the embedding against logits:

$$\mathbf{z}_i^{\parallel} = \arg\min_{\mathbf{f} \in \mathcal{H}_K} \sum_{j=1}^{n} \left\| \mathbf{z}_i - \mathbf{f}\left(\text{logits}(\hat{p}_{\text{ens}})_i\right) \right\|^2 + \lambda \|\mathbf{f}\|^2_{\mathcal{H}_K} \tag{14}$$

where $\mathcal{H}_K$ denotes the Reproducing Kernel Hilbert Space (RKHS) for all possible $\mathbf{f}$. Then, the residualzed embedding is produced by:

$$\mathbf{z}_i^{\perp} = \mathbf{z}_i - \hat{\mathbf{z}}_i \tag{15}$$

By removing the variance explained by calibrated risk, $\mathbf{z}_i^{\perp}$ should now capture **risk-independent structer**. In other words, patterns orthogonal to the main predictive signal. Visualizing $\mathbf{z}_i^{\perp}$ on both the $y = 0$ and $y = 1$ subset immediately reveal discrete clustering of feature values. Qualitatively, this suggests existence of subpopulation with similar risk but different profiles.

Figure 7: Color verification of residualized TabPFN embeddings $\mathbf{z}_i^{\perp}$ on the $y = 0$ **(top)** and $y = 1$ **(bottom)** subsets. Discrete clustering of feature values highlighted wiwth **purple bounding boxes**. For example, for the `fbs` feature of $y = 0$ subset (2nd purple box of top image), both UMAP and t-SNE reveals a small island of low fasting blood sugar ($\leq$ 120 mg/dL) within a large island of high fasting blood sugar ($>$ 120 mg/dL), a pattern not immediately apparant in Figure 6.

3.3 Pattern Discovery

Step 1: Target Engineering

This final stage aims to mine and synthesize insights from calibrated risk proxy and subgroup structure into actionable and interpretable rules. Using the calibrated risk proxy $\hat{p}_{\text{ens}}(\mathbf{x}) \in [0, 1]^n$ obtained from §3.1 , it is possible to mine patterns beyond binary labels. Here, two approaches are explored, logits and strats.

Specifically, calibrated probabilities $\{\hat{p}_{\text{ens}}(x_i)\}_{i=1}^n$ are first converted into logits $\ell_i = \text{logit}(\hat{p}_{\text{ens}})_i$ via equation (13). Then, discretize ℓ_i via quantile-based binning:

$$R_i = \begin{cases} \text{low} & \text{if } \ell_i < q_{0.25} \\ \text{mid-low} & \text{if } q_{0.25} \leq \ell_i < q_{0.5} \\ \text{mid-high} & \text{if } q_{0.5} \leq \ell_i < q_{0.75} \\ \text{high} & \text{if } \ell_i \geq q_{0.75} \end{cases} \tag{16}$$

where q_α denotes the α-quantile of $\{\ell_i\}$. Thus, creating a four-level target richer than binary labels.

Step 2: Conventional vs. Proposed Approach

A common mistake in course projects is to treats pattern mining as a standalone stage. I challenge this convention and ask the question: *Does propagating calibrated risk information and manifold structure insights yileds rules that are more semantically meaningful*?

Conventional Approach A typical workflow that treats pattern mining as standalone task (as criticized in §1) uses a binary label and possibly transformed feature space. In this setting, rules are optimized w.r.t. a single Bernoulli draw per patient. This is statistically fragile. Formally, let latent risk be $r(x) = P(Y = 1|X = x)$ and observed labels be $y_i \sim \text{Bernoulli}(r(x_i))$. Then, minin on y would be **inherently high-variance**. Let's model empirical event rate and its sampling variance:

$$\hat{r}_S^{(y)} = \frac{1}{|S|}\sum_{i\in S} y_i$$
$$\mathrm{Var}\Big(\hat{r}_S^{(y)} \mid x_{1:n}\Big) = \frac{1}{|S|^2}\sum_{i\in S} r_i(1-r_i) \le \frac{1}{4|S|} \tag{17}$$

This variance is irreducible even with prefect knowledge of x_i. Because, fundamentally, each y_i remains a stochastic Bernoulli draw. For small subgroups (exactly what we see interseting in §3.2), the bound $1/(4|S|)$ grows, making it less stable. Furthermore, when $r(x_i) \approx 0.5$ and $|S|$ is small (i.e., precisely where many concerned "interesting rules" fall into), variance is high. In simpler terms, conventional mining on binary labels tend to discover sampling noise around the decision boundary.

Proposed Approach My approach leverages insights gained from Stage 1 (§3.1) and Stage 2 (§3.2). Specifically, risk proxy from Stage 1 provides supervision, and residual manifold provides hypothesis space. In this setting, the target becomes ordinal strats ($\{R_i\}$).

Stage 1 of my pipeline produces out-of-fold claibrated probabilities $\hat{p}_i \approx r_i$. Assuming good calibration, this can be reasonably modelled as:

$$\hat{p}_i = r_i + \epsilon_i,\ \mathbb{E}[\epsilon_i | x_{1:n}] = 0 \tag{18}$$

The subgroup's mean prediced risk can also be modelled as:

$$\hat{r}_S^{(p)} = \frac{1}{|S|}\sum_{i\in S} \hat{p}_i \tag{19}$$

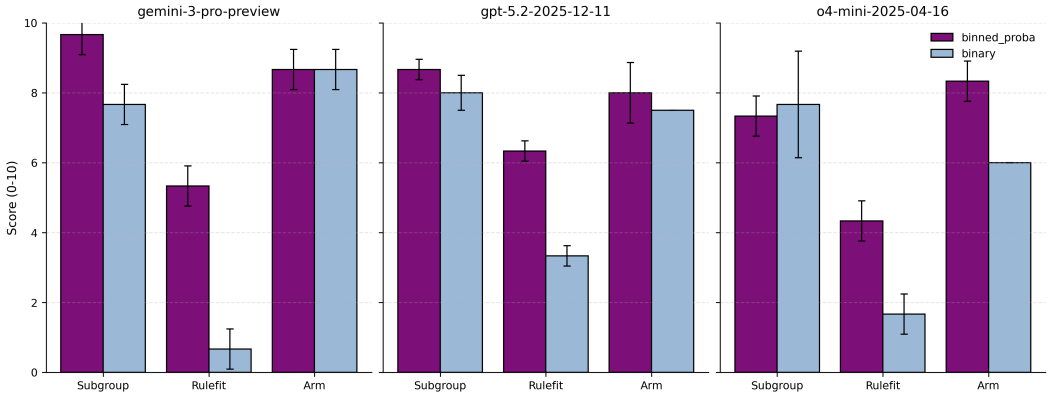
Then, its variance is determined by the estimation error, no longer irreducible Bernoulli noise:

$$\mathrm{Var}\Big(\hat{r}_S^{(p)} \mid x_{1:n}\Big) = \frac{1}{|S|^2}\sum_{i\in S} \mathrm{Var}(\epsilon_i | x_{1:n}) \tag{20}$$

Empirically, in Stage 1, the perturbation **the perturbation stability test is exactly a sanity check** to ensure that ϵ_i is small and stable. Under this premise, ideally we get $\mathrm{Var}(\epsilon_i) \ll r_i(1-r_i)$. Then, mining on $\hat{p}_i$ would yield statistcally better subgroup scoring than mining on y_i.

Stage 2, on the other hand, produces a residual manifold $\mathbf{z}_i^{\perp}$ that captures risk-independent structure. Qualtiatively, we see data manifold contains interesting subgroup patterns. This prompts exploration of subgroup discovery algorithms that may otherwise be missed by conventional pipeline.

Quantitative Results



$n = 3$	Association Rule Mining			RuleFit			Subgroup Discovery		
Target	Gemini3	GPT5.2	o4-mini	Gemini3	GPT5.2	o4-mini	Gemini3	GPT5.2	o4-mini
Binary	**8.7±0.6**	7.5±0.0	6.0±0.0	0.7±0.6	3.3±0.3	1.7±0.6	7.7±0.6	8.0±0.5	**7.7±1.5**
Binned	**8.7±0.6**	**8.0±0.9**	**8.3±0.6**	**5.3±0.6**	**6.3±0.3**	**4.3±0.6**	**9.7±0.6**	**8.7±0.3**	7.3±0.6

Table 4 summarizes *LLM-as-Judge* socres (mean ± std, higher better) for three baseliens, ARM, RuleFit, and Subgroup Discovery. Two targets are compared: Binary and binned (risk stratas). Results indicate that binned targets improves (or at least preserves) rule quality across baselines, with most dramatic gain on RuleFit (from useless to interpretable). To improve judge robustness, three different SOTA LLM models are used. Namely, Gemini 3 Pro, GPT5.2, and o4-mini. While their absolute harshness differs, the general **binned > binary** consensus holds.

These evaluations target *semantic validity* not just statistical association. The question we are trying to answer is does method X better maximizes internal interestingness / lift score, but rather: *Are the mined patterns more clinically meaningful?*. This is a question that is not possible with conventional frequency-based statistics, as it requires prior knowledges. Hence, LLM are the most suitable option.

Selected Qualitative Results

$$(\text{cp} = 3) \wedge (\text{ca} \in (-0.001, 1.0]) \wedge (\text{thal} = 3.0) \wedge (\text{slp} = 1) \Rightarrow \text{low}$$
$$(\text{exng} = 0) \wedge (\text{ca} \in (-0.001, 1.0]) \wedge (\text{sex} = 0) \wedge (\text{slp} = 1) \Rightarrow \text{low}$$
$$(\text{cp} = 4) \wedge (\text{thal} = 7.0) \Rightarrow \text{high}$$
$$(\text{exng} = 0) \wedge (\text{sex} = 1) \Rightarrow \text{mid-low}$$

The first rule, for example, exemplifies a **false alarm filter**. The rule combines non-anginal pain, with normal thallium scan, and unsloping ST segment. This showcases noise-filtering capability. Chest pain is usually the most effective predictor, but can also be noisy (false positives). The fact that, the algorithm is able to overcome this strong correlation when combined with normal ST slope and clear vessels indicates its pattern discovery capacity.

The third rule is also particularly interesting. Usually, `cp=4` (Asymptomatic/No Pain) is universally weighted as negative correlation with heart disease. However, here the model captures a known cardiology condition known as *Silent Ischemia*, which says, when "No pain" is paired with "Thallium defect", lack of pain is not sign of health.

Clinical Pattern Analysis

The discovered rules warrant deeper clinical interpretation. Consider the rule $(\text{cp} = 4) \wedge (\text{thall} = 7.0) \Rightarrow \text{high}$, which maps to the highest risk stratum. This pattern captures the phenomenon of *silent ischemia*, defined as myocardial ischemia occurring without typical anginal symptoms [9]. Patients with asymptomatic presentation ($\text{cp} = 4$) who simultaneously exhibit reversible thallium perfusion

defects ($\text{thall} = 7.0$) represent a particularly dangerous subpopulation. The reversible defect indicates viable but hypoperfused myocardium, while the absence of warning chest pain removes the protective signal that typically prompts patients to seek medical attention [10]. Epidemiological studies estimate that silent ischemia affects 2 to 4 percent of asymptomatic middle-aged individuals and carries mortality risk comparable to symptomatic coronary artery disease [11]. The fact that our calibration-guided mining framework surfaces this clinically significant pattern, rather than burying it among spurious correlations, validates the utility of risk-stratified target engineering.

The protective rules demonstrate equally important clinical validity. The pattern $(\text{cp} = 3) \wedge (\text{caa} \in (-0.001, 1.0]) \wedge (\text{thall} = 3.0) \wedge (\text{slp} = 1) \Rightarrow \text{low}$ combines four complementary reassurance signals: non-anginal chest pain ($\text{cp} = 3$), minimal coronary calcification ($\text{caa} \leq 1$), normal thallium perfusion ($\text{thall} = 3.0$), and upsloping ST segment ($\text{slp} = 1$). Each component independently reduces cardiac risk probability, and their conjunction defines a low-risk phenotype where chest discomfort likely originates from non-cardiac sources [12]. The upsloping ST segment morphology during exercise is particularly informative; unlike flat or downsloping patterns that suggest subendocardial ischemia, upsloping deflection typically reflects normal physiological response to increased heart rate [13]. Similarly, the rule $(\text{caa} = 0.0) \wedge (\text{exng} = 0) \wedge (\text{thall} = 3.0) \Rightarrow \text{low}$ identifies patients with zero fluoroscopy-visible vessel calcification, no exercise-induced angina, and normal nuclear imaging. This triple-negative profile would be categorized by clinical guidelines as very low pretest probability for obstructive coronary disease [14].

Sex-specific stratification emerges prominently in the high-risk rules, with $(\text{sex} = 1)$ appearing in patterns such as $(\text{cp} = 4) \wedge (\text{sex} = 1) \Rightarrow \text{mid-high}$. This aligns with established cardiovascular epidemiology showing that men under age 70 experience roughly twice the incidence of coronary events compared to age-matched women [15]. The male sex variable, when combined with asymptomatic presentation, amplifies risk because men are more likely to experience silent ischemia and sudden cardiac death without prodromal symptoms. Our framework correctly captures this interaction rather than treating sex as an independent linear predictor.

Finally, the discovery of $(\text{slp} = 2)$ (flat ST segment) in multiple high-risk rules corroborates exercise testing literature. A flat or horizontal ST depression during stress testing exhibits higher specificity for ischemia than upsloping depression, as flat morphology reflects more severe subendocardial supply-demand mismatch [13]. The conjunction $(\text{cp} = 4) \wedge (\text{slp} = 2) \wedge (\text{thall} = 7.0) \Rightarrow \text{high}$ thus represents a particularly ominous triad: silent presentation, ischemic ST morphology, and perfusion defect. These three independent indicators converge on the same high-risk stratum.

Summary

Taken together, these patterns demonstrate that calibration-guided target engineering enables discovery of rules that align with cardiology domain knowledge. The conventional binary-label approach, by contrast, conflates patients with varying degrees of underlying risk, resulting in rules that optimize statistical lift but lack clinical coherence. The quantitative LLM evaluation (Table 4) confirms this qualitative observation: binned risk strata consistently outperform binary targets across all three mining algorithms, with the most dramatic improvement observed in RuleFit (binary-target rules received near-zero clinical validity rating while binned-target rules achieved good ratings).

4 Appendix

4.1 Data Quality and Preprocessing

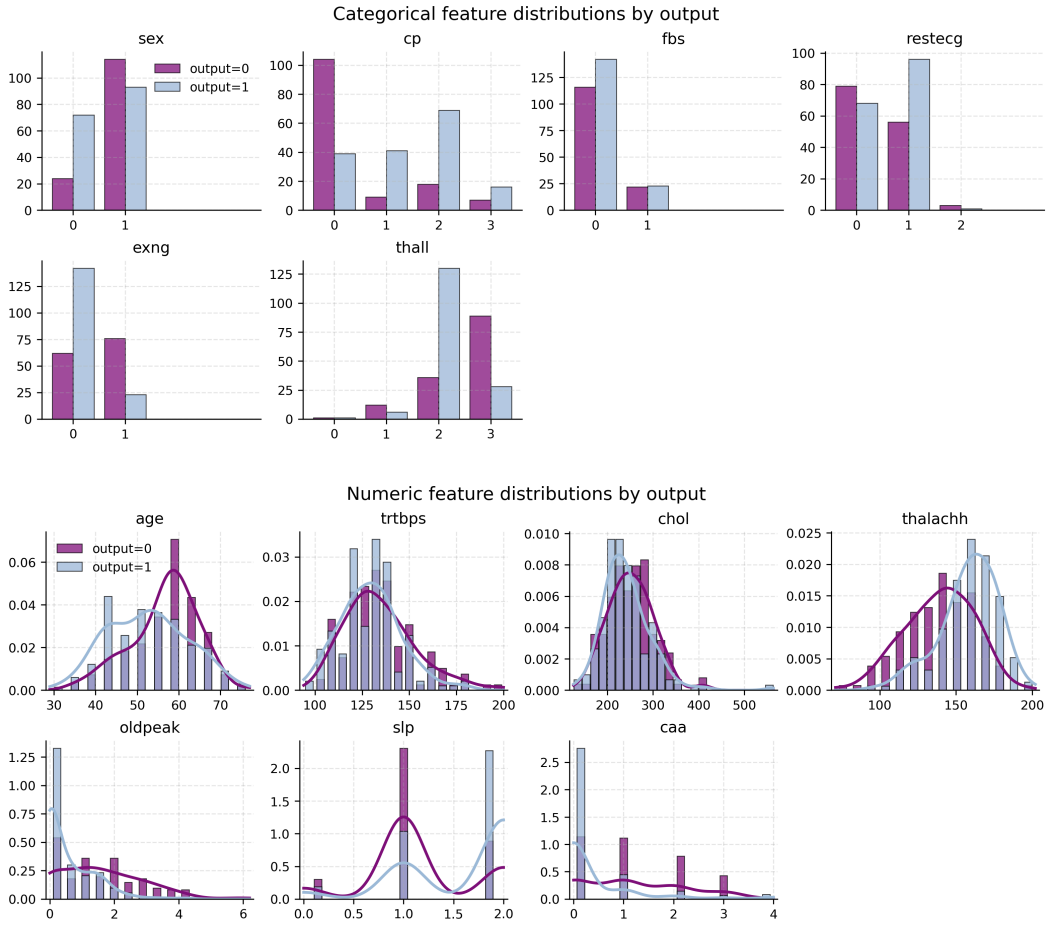


Figure 9: Feature distributions of the original Kaggle dataset before correction. The conditional distributions suggest label inversion: features clinically associated with disease risk (high ST depression, low max heart rate, presence of exercise angina) paradoxically concentrate in the `output=0` group.

Initial EDA of the Kaggle Heart Attack Analysis dataset revealed systematic inconsistencies between the stated label semantics and clinical expectations. The Kaggle documentation asserts that `output=0` indicates absence of heart attack risk while `output=1` indicates presence. However, examination of feature distributions exposes four contradictions that collectively points to label inversion.

First, the ST depression feature (`oldpeak`) exhibits mean values of 1.58 for `output=0` versus 0.58 for `output=1`. Since elevated ST depression during exercise indicates myocardial ischemia, the higher mean in the purportedly healthy group contradicts established cardiology. Second, maximum heart rate achieved (`thalachh`) averages 139.1 bpm for `output=0` compared to 158.5 bpm for `output=1`. Lower exercise capacity typically signals compromised cardiac function, yet the supposedly diseased group demonstrates superior performance. Third, exercise-induced angina (`exng`) occurs in 55% of `output=0` patients but only 14% of `output=1` patients. The presence of exertional chest pain is a hallmark indicator of coronary insufficiency, making its concentration in the "healthy" group implausible. Fourth, the number of major vessels colored by fluoroscopy (`caa`) shows predominantly non-zero values for `output=0` and predominantly zero values for `output=1`. Greater coronary calcification visible on fluoroscopy correlates with atherosclerotic burden, contradicting label semantics.

Cross-referencing with the original UCI Cleveland Heart Disease dataset confirmed that the Kaggle version had inverted labels. Additionally, the Kaggle data contained invalid values such as

`thall=0`, which has no clinical meaning given that the thallium scan feature encodes only three categories (normal, fixed defect, reversible defect). After correcting label polarity and removing invalid categorical entries, the cleaned dataset exhibits distributions consistent with cardiology domain knowledge (Figure 10).

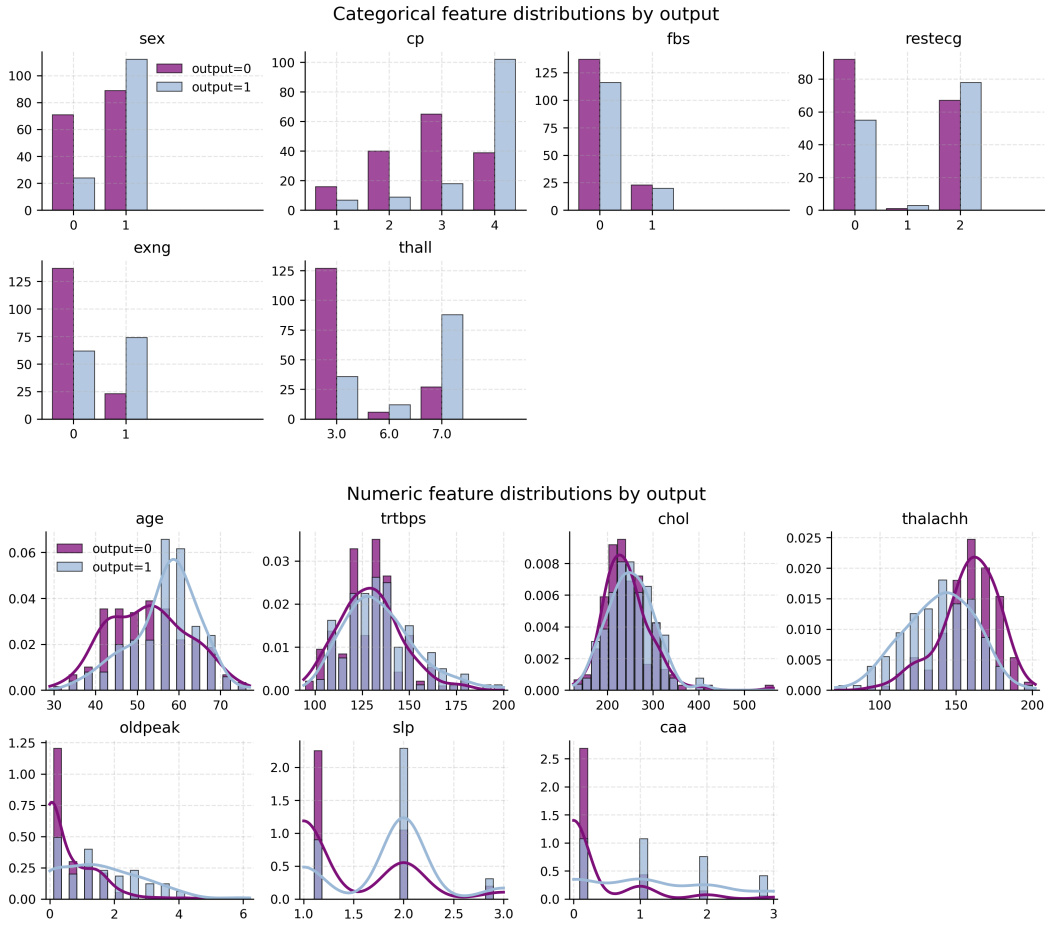


Figure 10: Feature distributions after label correction and data cleaning. Risk-associated features (ST depression, exercise angina, etc.) now concentrate in `output=1`, and protective features (high max heart rate, absence of angina) concentrate in `output=0`, aligning with clinical expectations.

4.2 LLM-as-Judge Evaluation Protocol

To assess the clinical validity of mined rules beyond statistical metrics, we employed a blind evaluation protocol using frontier large language models as domain expert proxies. Three state-of-the-art models were used: *Gemini-3-Pro-Preview*, *GPT-5.2*, and *o4-mini*. Each model evaluated rules from all three mining algorithms (Association Rule Mining, RuleFit, Subgroup Discovery) across two target definitions (binary labels and binned risk strata).

The evaluation followed a randomized blind design. For each algorithm, rules mined under different target definitions were assigned anonymous group labels (A, B, C) with random permutation seeded by algorithm name, model identifier, and round index. This prevents models from inferring target semantics from label ordering. Each model completed three independent evaluation rounds per algorithm, yielding nine judgments per algorithm-target pair across all models.

Models received a system prompt establishing cardiology domain expertise, a feature glossary with clinical interpretations, and a scoring rubric mapping 0-10 scores to clinical plausibility levels. The prompt explicitly instructs models to evaluate rules based on alignment with established cardiovas-

cular risk factors rather than statistical properties. Temperature was set to 1.0 to encourage response diversity across rounds, and maximum output tokens (including reasoning) was set to 32,000.

LLM-as-Judge System Prompt

The following system prompt was used verbatim for all LLM-as-Judge evaluations:

```
You are a cardiology-domain medical reviewer evaluating whether
mined rules align with well-known heart disease risk factors and
clinical intuition.

You will receive multiple anonymous groups (e.g., Group A, Group B,
Group C). Each group contains a list of rules mined from the SAME
dataset, but using different (hidden) target definitions. The group
labels and their order are arbitrary and randomly permuted each time.
Do NOT assume any meaning from the letters or ordering.

Your task:
- Score EACH group on a 0-10 scale for how medically plausible and
  clinically meaningful its rules are.
- Higher score = rules are more consistent with cardiology knowledge
  and have fewer nonsensical/contradictory patterns.
- Evaluate each group independently based only on its content.
- Multi-class labels and continuous outputs are not necessarily
  better, unless you are confident they provide superior medical
  insights.

Dataset feature glossary (Heart Disease style dataset):
- age: age in years (higher often increases risk)
- sex: 0=female, 1=male (male often higher risk)
- cp: chest pain type (1=typical angina, 2=atypical angina,
  3=non-anginal pain, 4=asymptomatic; cp=4 often higher risk)
- trtbps: resting blood pressure (higher often higher risk)
- chol: serum cholesterol in mg/dl (higher can increase risk,
  but noisy)
- fbs: fasting blood sugar > 120 mg/dl (1=true; diabetes risk factor)
- restecg: resting ECG (0=normal, 1=ST-T abnormality,
  2=LV hypertrophy; abnormal values can increase risk)
- thalachh: maximum heart rate achieved (lower can indicate worse
  functional capacity; interpretation depends on context/age)
- exng: exercise-induced angina (1=yes; increases risk)
- oldpeak: ST depression induced by exercise vs rest (higher
  increases risk)
- slp: slope of peak exercise ST segment (1=upsloping, 2=flat,
  3=downsloping; flat/downsloping increases risk)
- caa: number of major vessels (0-3) colored by fluoroscopy (higher
  increases risk)
- thall: thalassemia test (3=normal, 6=fixed defect, 7=reversible
  defect; defects increase risk)

How to interpret rule formats:
- Rules are patterns over the features above. They may use equality
  (e.g., "cp==4") and/or ranges (e.g., "age=(54, 62]").
- Focus on whether the patterns align with established risk factors
  and clinical intuition.

Scoring rubric:
- 9-10: Strongly aligned with established risk factors, coherent,
  few/no contradictions, not dominated by spurious patterns.
- 6-8: Mostly plausible but some weak/noisy or questionable rules.
- 3-5: Mixed plausibility, many unclear/uninformative rules, or
  several questionable patterns.
- 0-2: Largely nonsensical, contradicts medical intuition, or
  essentially empty/unusable.

Output format (STRICT):
- Return ONLY valid JSON (no markdown, no code fences, no extra text).
- Schema:
  {
    "scores": { "A": <number 0-10>, "B": <number 0-10>, ... },
    "brief_reason": { "A": "<short>", "B": "<short>", ... }
  }
- Adapt keys to match the number of groups presented.
- Scores must be numeric and within [0, 10].
```

Listing 1: Complete system prompt used for LLM-as-Judge evaluations.

5 References

[1] N. Hollmann *et al.*, "Accurate predictions on small data with a tabular foundation model," *Nature*, vol. 637, no. 8045, pp. 319–326, Jan. 2025, doi: 10.1038/s41586-024-08328-6.

[2] J. Qu, D. Holzmüller, G. Varoquaux, and M. L. Morvan, "TabICL: A Tabular Foundation Model for In-Context Learning on Large Data." Accessed: Nov. 12, 2025. [Online]. Available: http://arxiv.org/abs/2502.05564

[3] X. Zhang *et al.*, "Mitra: Mixed Synthetic Priors for Enhancing Tabular Foundation Models." Accessed: Nov. 12, 2025. [Online]. Available: http://arxiv.org/abs/2510.21204

[4] X. Zhang *et al.*, "LimiX: Unleashing Structured-Data Modeling Capability for Generalist Intelligence." Accessed: Dec. 18, 2025. [Online]. Available: http://arxiv.org/abs/2509.03505

[5] M. Pavlovic, "Understanding Model Calibration – A gentle introduction and visual exploration of calibration and the expected calibration error (ECE)." Accessed: Dec. 19, 2025. [Online]. Available: http://arxiv.org/abs/2501.19047

[6] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks." Accessed: Dec. 19, 2025. [Online]. Available: http://arxiv.org/abs/1706.04599

[7] T. Popordanoska, R. Sayer, and M. B. Blaschko, "A Consistent and Differentiable Lp Canonical Calibration Error Estimator." Accessed: Dec. 19, 2025. [Online]. Available: http://arxiv.org/abs/2210.07810

[8] M. Atzmueller, "Subgroup discovery," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 5, no. 1, pp. 35–49, Jan. 2015, doi: 10.1002/widm.1144.

[9] P. F. Cohn, K. M. Fox, and C. Daly, "Silent Myocardial Ischemia," *Circulation*, vol. 108, no. 10, pp. 1263–1277, 2003, doi: 10.1161/01.CIR.0000088001.59265.EE.

[10] S. O. Gottlieb, M. L. Weisfeldt, P. Ouyang, E. D. Mellits, and G. Gerstenblith, "Silent Ischemia as a Marker for Early Unfavorable Outcomes in Patients with Unstable Angina," *New England Journal of Medicine*, vol. 314, no. 19, pp. 1214–1219, 1986, doi: 10.1056/NEJM198605083141903.

[11] P. F. Cohn, "Silent Myocardial Ischemia," *Annals of Internal Medicine*, vol. 109, no. 4, pp. 312–317, 1988, doi: 10.7326/0003-4819-109-4-312.

[12] R. J. Gibbons *et al.*, "ACC/AHA 2002 Guideline Update for Exercise Testing: Summary Article," *Circulation*, vol. 106, no. 14, pp. 1883–1892, 2002, doi: 10.1161/01.CIR.0000034670.06526.15.

[13] Y. C. Lim, S.-G. Teo, and K.-K. Poh, "ST-Segment Changes with Exercise Stress," *Singapore Medical Journal*, vol. 57, no. 7, pp. 347–353, 2016, doi: 10.11622/smedj.2016116.

[14] P. Greenland, M. J. Blaha, M. J. Budoff, R. Erbel, and K. E. Watson, "Coronary Calcium Score and Cardiovascular Risk," *Journal of the American College of Cardiology*, vol. 72, no. 4, pp. 434–447, 2018, doi: 10.1016/j.jacc.2018.05.027.

[15] L. Mosca, E. Barrett-Connor, and N. K. Wenger, "Sex/Gender Differences in Cardiovascular Disease Prevention: What a Difference a Decade Makes," *Circulation*, vol. 124, no. 19, pp. 2145–2154, 2011, doi: 10.1161/CIRCULATIONAHA.110.968792.

[16] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes, "Ensemble selection from libraries of models," in *Twenty-first international conference on Machine learning - ICML '04*, Banff, Alberta, Canada: ACM Press, 2004, p. 18. doi: 10.1145/1015330.1015432.

[17] D. Salinas and N. Erickson, "TabRepo: A Large Scale Repository of Tabular Model Evaluations and its AutoML Applications." Accessed: Nov. 12, 2025. [Online]. Available: http://arxiv.org/abs/2311.02971

[18] N. Erickson *et al.*, "TabArena: A Living Benchmark for Machine Learning on Tabular Data." Accessed: Nov. 12, 2025. [Online]. Available: http://arxiv.org/abs/2506.16791

[19] Y. Gorishniy, A. Kotelnikov, and A. Babenko, "TabM: Advancing Tabular Deep Learning with Parameter-Efficient Ensembling." Accessed: Nov. 12, 2025. [Online]. Available: http://arxiv.org/abs/2410.24210

[20] N. Erickson *et al.*, "AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data." Accessed: Nov. 12, 2025. [Online]. Available: http://arxiv.org/abs/2003.06505

[21] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Nov. 12, 2025. [Online].

Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

[22] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2018. Accessed: Nov. 12, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html

[23] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[24] N. Erickson, "autogluon/autogluon." Accessed: Nov. 12, 2025. [Online]. Available: https://github.com/autogluon/autogluon

[25] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[26] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support Vector Regression Machines," in *Advances in Neural Information Processing Systems*, MIT Press, 1996. Accessed: Nov. 12, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/1996/hash/d38901788c533e8286cb6400b40b386d-Abstract.html

[27] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.

[28] I.-K. Yeo and R. A. Johnson, "A new family of power transformations to improve normality or symmetry," *Biometrika*, vol. 87, no. 4, pp. 954–959, Dec. 2000, doi: 10.1093/biomet/87.4.954.

[29] World Health Organization, "Cardiovascular diseases (CVDs)." Accessed: Dec. 18, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[30] H. Ismail *et al.*, "Heart Attack Analysis & Prediction Dataset." Accessed: Dec. 18, 2025. [Online]. Available: https://www.kaggle.com/datasets/sonialikhan/heart-attack-analysis-and-prediction-dataset

[31] F. Lemmerich and M. Becker, "pysubgroup: Easy-to-Use Subgroup Discovery in Python: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part III." pp. 658–662, Jan. 2019. doi: 10.1007/978-3-030-10997-4_46.