Practical and Efficient Rashomon Set Sampling for Model Interpretability

Sichao Li[1,2] **Amanda S. Barnard**[1] **Quanling Deng**[1,3]

[1]School of Computing, Australian National University

[2]School of Computer Science, University of Sydney

[3]Yau Mathematical Sciences Center, Tsinghua University

sichao.li@sydney.edu.au, amanda.s.barnard@anu.edu.au, qldeng@tsinghua.edu.cn

Abstract

Explaining a single model can be misleading when many near-optimal models (a *Rashomon set*) yield different feature attributions. We frame this as a Rashomon set sampling problem and propose two practical axioms that any Rashomon sampler should satisfy: *generalizability* (meaning it must accept arbitrary reference models and loss functions) and *Implementation Sparsity* (meaning it should return a small, attribution-diverse subset of valid models). These two axioms are not satisfied by most known attribution methods, which we consider to be a fundamental weakness. Building on these axioms, we propose an ϵ-subgradient-based sampling framework and quantify effectiveness with *Search Efficiency Ratio* (SER) and *Functional Explanation Range* (FER). Experiments on a synthetic quadratic task and five real-world datasets show that our sampler achieves comparable or higher FER with up to $\sim 100\times$ fewer models than exhaustive baselines such as TreeFARMS, while remaining agnostic to model class and loss. Even when the reference model is sub-optimal in practice, the resulting attributions align with ground truth and accepted domain knowledge.

1 INTRODUCTION & MOTIVATION

Understanding the behavior of machine learning models is receiving considerable attention. Many researchers seek to identify important features and describe their effects depending on a specific model. Recently, researchers have argued that explaining a *single* model is insufficient; instead, one should examine a set of similar-performing models to obtain more robust insights. This set of models is called a *Rashomon set* (Fisher et al., 2019; Renard et al., 2024; Cavus and Biecek, 2025). A commonly agreed definition of the Rashomon set in the machine learning community is when the benchmark model $f^* \in \mathcal{F}$ minimizes the loss function, i.e., $\mathcal{L}(f^*) = \inf_{f\in\mathcal{F}} \mathcal{L}(f)$, the set defined by hypothesis space $\mathcal{H}$ on the basis of $\theta^* > 0$ is called the Rashomon set; see for example in (Dong and Rudin, 2020; Xin et al., 2022; Li and Barnard, 2022).

$$\mathcal{R}(\theta^*, f^*, \mathcal{F}) = \{f \in \mathcal{F} : \mathcal{L}(f(\mathbf{X})) \leq \mathcal{L}(f^*(\mathbf{X})) + \theta^*\}. \quad (1)$$

In other words, $\mathcal{R}(\theta^*, f^*, \mathcal{F})$ contains all models in $\mathcal{F}$ whose loss is within an acceptable tolerance of the minimal loss. In practice, however, f^* (the true loss-minimizing model) is usually unknown or intractable to find. Therefore, we will instead define the Rashomon set relative to a chosen reference model f_{ref}. For a given $f_{ref} \in \mathcal{F}$ (e.g., a well-performing model identified via cross-validation or other selection criteria), we consider $\mathcal{R}(\theta^*, f_{ref}, \mathcal{F})$ as the practical Rashomon set around f_{ref}. This formulation does not require f_{ref} to be truly optimal; it simply serves as a baseline model with high performance, which will be used for all subsequent analysis.

A significant challenge in utilizing Rashomon sets is that exploring the *entire* set of near-optimal models is infeasible due to the sheer size of the hypothesis space Semenova et al. (2022). For example, the possible number of trees of depth at low as 4 with 10 binary features is more than 9.338×10^{20} and the number of neural networks (NNs) in hypothesis space grows exponentially with the number of parameters (Hu et al., 2019). The recent work of Hsu and Calmon (2022) studied the *Rashomon ratio* and the pattern Rashomon ratio to estimate the volume of the Rashomon set. This is a level set estimation problem and represents the

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

fraction of models in the hypothesis space that fit the data about equally well. Such results provide insight into the complexity of a learning problem (e.g., Rashomon ratios tend to be large for low-complexity model classes; see (Semenova et al., 2024)), but they do not by themselves tell us how to find, explain, or utilize the models in the Rashomon set. In other words, measuring the size of a Rashomon set is not the same as being able to sample and interpret its elements. In this work, we focus on the practical task of sampling a representative subset of the Rashomon set and guiding feature attribution based on it.

Most existing approaches avoid brute-force enumeration by sampling or approximating the Rashomon set in restricted ways (Hsu and Calmon, 2022). For instance, Dong and Rudin (2020) provided a simple example of enumerating similarly-performing decision trees under certain settings, but not general for all cases. Xin et al. (2022) recently provided a practical tree-based solution for a nonlinear discrete model class by fitting an optimal tree and exploring the whole Rashomon set of sparse decision trees. Dong and Rudin (2020) also trained a logistic regressor and provided an ellipsoid approximation for the logistic model class. For NN-based models, Li and Barnard (2022) provided a variance tolerance factor (VTF) to interpret all NNs by greedy searching an extra layer on the top of the network. While these methods are effective in their specified domains, they all *assume prior knowledge of the reference model's form* (e.g., that it is a decision tree, linear model, or other interpretable structure (Chen et al., 2023)). In practice, one may be given an arbitrary pre-trained model as a black-box, without the ability to constrain it to a simple form, as shown in Fig. 1. There is a need for a more general workflow to handle an arbitrary reference model and still explore its Rashomon set Laberge et al. (2023). Fig. 2 illustrates the conventional Rashomon set workflow (top-middle) and our proposed general workflow (bottom-middle) in this context. Additionally, the lack of quantitative comparison of sampling methods hinders the potential application of the Rashomon set.

To address these shortcomings, we take an axiomatic approach and identify two fundamental practical principles. Unfortunately, most existing methods violate at least one of these principles, limiting their practical usefulness. Our contributions can be summarized as:

- We distill two high-level principles, generalizability (model structure, evaluation, and attribution agnostic) and implementation sparsity (search efficiency and functional diversity), and formalize them as practical axioms that any Rashomon sampler should consider in design.

- Guided by these axioms, we introduce an ϵ-subgradient sampling framework based on subgradient descent strategy that (i) generalizes Rashomon sets to arbitrary reference models and loss functions, (ii) supports a model-agnostic any-order feature-attribution score function, and (iii) comes with supporting theorems and proofs guaranteeing convergence, near-optimality, and diversity.

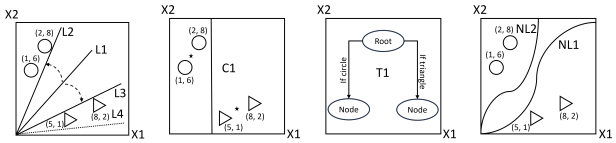


Figure 1: A toy example of binary classification (circle and triangle) by $x1$, $x2$, and object shape. Possible models in the Rashomon set can be: linear model, clustering model, decision tree, and non-linear model (from *left* to *right*).

- We validate the properties of the proposed framework through (i) synthetic experiments with known ground-truth attributions (quadratic roots) and (ii) comprehensive comparisons with baseline methods on five real-world datasets, highlighting that our approach achieves attribution diversity with significantly fewer sampled models.

2 NOTATIONS & TERMINOLOGIES

Notations We use bold lowercase letters such as $\mathbf{v}$ to represent a vector and v_i denotes its i-th element. Let the bold uppercase letters such as $\mathbf{A}$ denote a matrix with $a_{[i,j]}$ being its i-th row and j-th column entry. The vectors $\mathbf{a}_{[i,\cdot]}$ and $\mathbf{a}_{[\cdot,j]}$ are its i-th row and j-th column, respectively. Let $(\mathbf{X}, \mathbf{y}) \in (\mathbb{R}^{n\times p}, \mathbb{R}^n)$ denote the dataset where $\mathbf{X} = [\mathbf{x}_{[\cdot,1]}, \mathbf{x}_{[\cdot,2]}, \ldots, \mathbf{x}_{[\cdot,p]}]$ is a $n \times p$ co-variate input matrix and $\mathbf{y}$ is a n-length output vector. We assume the observations are drawn i.i.d. from a distribution $\mathcal{D}$.

Let $\mathcal{I}$ be a subset of feature indices: $\mathcal{I} \subset \{1, 2, \ldots, p\}$ and its cardinality is denoted by $|\mathcal{I}|$. All possible subsets are referred as $\mathbb{I} = \{\mathcal{I} \mid \mathcal{I} \subset \{1, 2, \ldots, p\}\}$. In the context of no ambiguity, we drop the square brackets on the vector and simply use $\mathbf{x}_s$ to denote the feature. $\mathbf{X}_{\setminus \mathbf{s}}$ is the input matrix when the feature of interest (denoted as $\mathbf{s}$ here) is replaced by an independent variable. Let $f : \mathbb{R}^{n\times p} \to \mathbb{R}^n$ be a predictive model that maps an input matrix with n samples and p features to an n-dimensional vector of predictions. A model class is a set of all such predictive functions e.g., $\mathcal{F} \subset \{f \mid f \in \mathcal{F}\}$, where each f is simply one possible predictive model within this class. Let $\mathcal{L} : (f(\mathbf{X}), \mathbf{y}) \to \mathbb{R}$ be the loss function and the expected loss and empirical loss are defined as $\mathcal{L}_{exp} = \mathbb{E}[\mathcal{L}(f(\mathbf{X}), \mathbf{y})]$ and $\hat{\mathcal{L}} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(f(\mathbf{x}_{[i,\cdot]}), y_i)$, respectively.

Generalized Rashomon Set We denote f_{ref} as the Rashomon set's reference or baseline model for any trained machine learning model f_{ref}, e.g. tree models, NNs, in

any task, e.g. supervised learning, unsupervised learning. The generalized Rashomon set is defined on the basis of an epsilon rate $\epsilon > 0$ from Eq.(1), and the threshold for the Rashomon set is $\theta^* = \epsilon\mathcal{L}_{ref}(f_{ref})$:

$$\mathcal{R}_\epsilon(f_{ref}) = \{f \in \mathcal{F} : \mathcal{L}(f(\mathbf{X})) \leq (1+\epsilon)\mathcal{L}(f_{ref}(\mathbf{X}))\}. \quad (2)$$

In practice, researchers sample from the generalized Rashomon set in different ways. Intuitively, we have:

Proposition 2.1. *Any sampled $\hat{\mathcal{R}}_\epsilon(f_{ref})$ in practice by a finite algorithm and certified against Eq.* (2) *is a subset of $\mathcal{R}_\epsilon(f_{ref})$ (Proof see Appendix A.1).*

If $\mathcal{L}_{ref}(f^*) = 0$ (rare in practice), one can equivalently impose $\mathcal{L}_{ref}(f) \leq \epsilon$; the theory below is unaffected.

Generalized Feature Attribution The term "feature attribution" is used to denote the general contribution of the feature to the prediction. Higher-order feature attribution refers to interactions among features. Here we define:

Definition 2.2. *A general model-agnostic score function $s(\cdot)$ measures the importance of any order of features, e.g. $|\mathcal{I}| = 1$ for feature importance and $|\mathcal{I}| > 1$ for interaction importance, on the given data set. The conditional expected $s_\mathcal{I}$ is the expected score $q_\mathcal{I}$ conditional to the feature set $\mathcal{I}$ of the model $f \in \mathcal{F}$, written as:*

$$q_\mathcal{I}(f) = \mathbb{E}[s_\mathcal{I}(\mathbf{X}, \mathbf{y}) \mid f \in \mathcal{F}]. \quad (3)$$

Attribution Vector and Attribution Space Following the definition, we can derive:

- A single model attributing to all possible feature subsets:

$$\mathbf{q}_\mathbb{I}(f) := \big(q_\mathcal{I}(f)\big)_{\mathcal{I}\in\mathbb{I}} \in \mathbb{R}^{|\mathbb{I}|}. \quad (4)$$

- All models in S attributing to a specific feature set $\mathcal{I}$:

$$\mathcal{Q}_\mathcal{I}(S) := \{q_\mathcal{I}(f) : f \in S\} \subseteq \mathbb{R}. \quad (5)$$

- Putting them together, we obtain a attribution (matrix) space over an empirical sampled Rashomon set as:

$$\mathbf{Q}_\mathbb{I}(\hat{\mathcal{R}}_\epsilon) := \big(\mathbf{q}_\mathbb{I}(f)\big)_{f\in\hat{\mathcal{R}}_\epsilon} \in \mathbb{R}^{|\hat{\mathcal{R}}_\epsilon|\times|\mathbb{I}|}$$

Proposition 2.3. *For any non-empty $\hat{\mathcal{R}}_\epsilon(f_{ref})$, the matrix $\mathbf{Q}_\mathcal{I}(\hat{\mathcal{R}}_\epsilon)$ is finite and non-empty (Proof see Appendix A.2).*

Definition 2.4 (Functional Attribution Distance)**.** *Let $f_i, f_j \in \mathcal{R}_\epsilon$ be any two models in the Rashomon set. Their* functional attribution distance *is the Chebyshev norm between their attribution vectors based on Eq.* (4)*:*

$$\mathrm{dist}_\infty(\mathbf{q}_\mathbb{I}(f_i), \mathbf{q}_\mathbb{I}(f_j)) := \lim_{p\to\infty}\Big(\sum_{\mathcal{I}\in\mathbb{I}} |q_\mathcal{I}(f_i) - q_\mathcal{I}(f_j)|^p\Big)^{(1/p)}.$$

3 DESIGN PRINCIPLES FOR RASHOMON SAMPLING

To guide the development of a practical Rashomon sampler, we distill two fundamental principles into concrete axioms that formalize the desiderata for sampling diverse, near-optimal models without making restrictive assumptions about model class or training objective.

Axiom: Generalizability The goal of Rashomon set analysis is to extract *model-independent* insights from all near-optimal models. A sampler must therefore function across architectures, loss functions, and attribution methods. We decompose this requirement into three components:

- **Model-structure generalizability:** The method should operate regardless of the architecture or training process of the reference model $f_{ref} \in \mathcal{F}$. For instance, Fig. 1 illustrates a task solvable by multiple learning paradigms, e.g., linear classifiers, nonlinear SVMs, clustering methods, and decision trees Pankajakshan et al. (2017); Roscher et al. (2020a); Sundararajan et al. (2017).

- **Evaluation generalizability:** Different models may be trained with different objectives (e.g., log-loss, MSE, MAE), but their performance *must* be evaluated using a shared metric $\mathcal{L}(\cdot)$ when forming $\hat{\mathcal{R}}_\epsilon(f_{ref})$ in practice. For example, a classifier trained with log-loss or Gini gain may still achieve high accuracy Nauta et al. (2023); Zhong et al. (2022); Gola et al. (2018); Guidotti et al. (2018); Adadi and Berrada (2018); Gilpin et al. (2018).

- **Attribution generalizability:** Attribution methods vary widely in form, gradient-based (e.g., saliency), perturbation-based (e.g., SHAP, LIME), or loss-based methods, but can yield different explanations for the same feature depending on the model class Linardatos et al. (2020); Zhong et al. (2022); Roscher et al. (2020b); Imrie et al. (2023). Since the goal is to understand consistent feature behavior across models, the attribution method must be fixed throughout sampling and should support higher-order attributions (e.g., interactions).

Axiom: Implementation Sparsity The second principle ensures that sampling from the Rashomon set is both computationally efficient and yields diverse feature attributions. We split this into two complementary goals.

- **Search efficiency:** Sampling should avoid evaluating models unlikely to belong in $\mathcal{R}_\epsilon(f_{ref})$, especially in high-dimensional spaces. For instance, decision tree enumeration is intractable due to the vast number of all possible trees (e.g., a low depth tree with $\sim 10^{20}$ possibilities). Existing methods such as top-down heuristics Xin et al. (2022) or ellipsoidal approximations Dong and Rudin (2020) aim to reduce overhead. We define a gen-

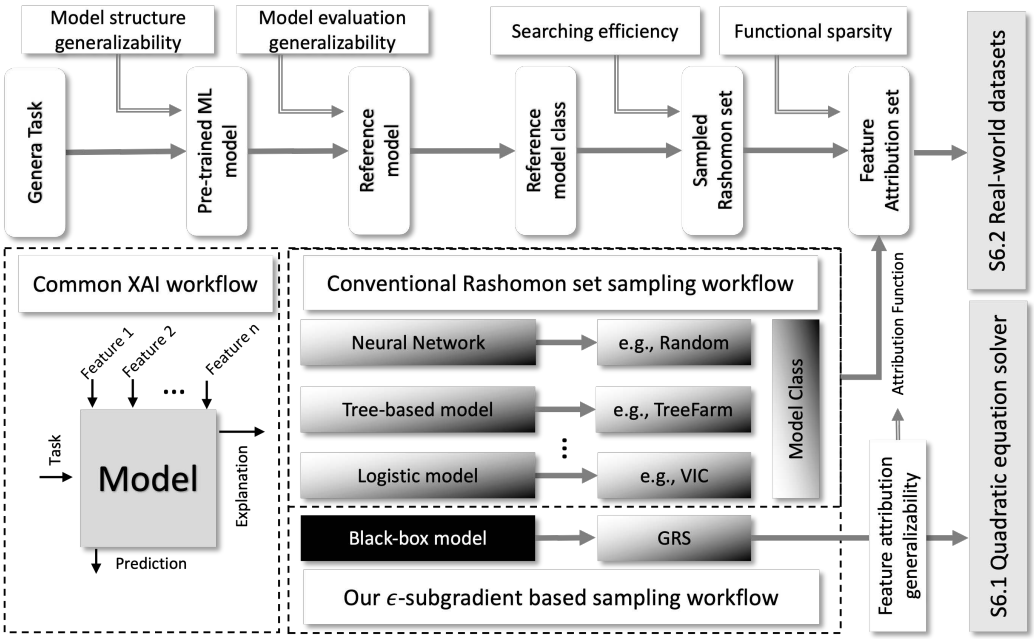


Figure 2: Overview of our Rashomon-set sampling framework. **Top pipeline.** End-to-end pipeline, from task specification through a sampled Rashomon set with the desiderata satisfied at each stage: *model-structure generalizability*, *evaluation generalizability*, *search efficiency*, and *functional sparsity*. **Bottom-left.** Standard XAI workflow that derives feature attributions from a *single* model. **Bottom-middle.** Conventional Rashomon sampling: a practitioner fixes a reference model class (e.g., neural network, tree-based, logistic) and employs samplers (e.g., Random, VIC, TreeFarm, . . .) to approximate the Rashomon set. Grayscale bars denote decreasing interpretability. Black-box scenario where the reference model is an opaque, pre-trained predictor. **Bottom-right.** Our proposed ϵ-subgradient Rashomon Sampler enables class–agnostic exploration in the black-box setting, producing a sparse, diverse set of functionally equivalent models.

eral efficiency metrics, *search efficiency ratio* (SER):

$$\text{SER} := |\hat{\mathcal{R}}_\epsilon|/N, \tag{6}$$

where N is the number of all searched models, for which we evaluate the empirical loss, and $|\hat{\mathcal{R}}_\epsilon|$ denotes the number of valid (feasible and accepted) models.

- **Functional sparsity:** Two models are *functionally redundant* if they yield the same feature attribution scores. To maximize interpretive utility Rudin (2019), a Rashomon sampler should *avoid* adding multiple models with identical explanations. For example, in NNs, two structurally different implementations may produce identical gradients due to the chain rule (e.g., $\partial f/\partial g = \partial f/\partial h \cdot \partial h/\partial g$) (Samek et al., 2016; Shrikumar et al., 2017; Tsang et al., 2017). In such cases, only one representative model is needed for interpretation purposes. To quantify diversity in explanations, we measure the span of possible scores for a feature subset $\mathcal{I}$ via the interval $[\min(\mathcal{Q}_\mathcal{I}(S)), \max(\mathcal{Q}_\mathcal{I}(S))]$. The overall *Functional Explanation Range* (FER) of a sampled set based on Eq. (5) is obtained by summing these spans as (Fisher et al., 2019; Hsu and Calmon, 2022; Li et al., 2023):

$$\text{FER}(S) := \sum_{\mathcal{I}\in\mathbb{I}} \left[\max(\mathcal{Q}_\mathcal{I}(S)) - \min(\mathcal{Q}_\mathcal{I}(S))\right]. \tag{7}$$

A well-designed sampler therefore strives to *maximize* FER while keeping a *minimum* number of models.

4 PRACTICAL RASHOMON SUBSET SAMPLING

Guided by the principles in Sec. 3, we present a practical **ϵ-subgradient framework (GRS)** that produces a functionally diverse subset of the generalized Rashomon set $\mathcal{R}_\epsilon(f_{ref})$ (Eq. (2)). The framework is model– and loss–agnostic and consists of three components:

(i) *Unified representation of candidate models* as masked-input perturbations of a fixed reference model, enabling class-agnostic exploration.

(ii) *Model-agnostic attribution score* that supports any-order feature and interaction effects.

(iii) *ϵ-subgradient sampling algorithm* that solves a constrained, non-differentiable optimization and searches the model candidates with maximally diverse attributions(Lorenz, 1995; Papadimitriou and Steiglitz, 1998; Bertsekas et al., 2003).

4.1 Unified Representation of Rashomon Models

Input-activated parameterization To characterize the generalized Rashomon set and find a generalized representation of Rashomon models, we express every candidate model as the reference model applied to the masked input:

$$f_\tau(\mathbf{x}) \;=\; f_{ref}(m_\tau(\mathbf{x})), \qquad \tau \in \mathbb{R}^p, \tag{8}$$

with

$$m_\tau(\mathbf{x}_{[i,\cdot]}) = \tau \odot \mathbf{x}_{[i,\cdot]} + \zeta_i, \tag{9}$$

where $\odot$ denotes element-wise multiplication and ζ_i is an additional noise. This perspective follows the "mask-plus-X" construction (Li et al., 2023) and inherits **model-structure generalizability**. For example, decision trees can be represented as NNs (Yang et al., 2018; Hinton and Frosst, 2017; Aytekin, 2022); support vector machines as a shallow NN; and K-means clustering with NNs (Sitompul et al., 2018).

Proposition 4.1. *For a dataset $\mathbf{X} \in \mathbb{R}^{n\times p}$, let $\mathbf{Z} := m_\tau(\mathbf{X})$ denote the perturbed input. Then the feasible mask set*

$$\mathcal{T}_\epsilon := \Big\{\tau \in \mathbb{R}^p : \hat{\mathcal{L}}\big(f_{ref}(\mathbf{Z}), \mathbf{y}\big) \le (1+\epsilon)\hat{\mathcal{L}}\big(f_{ref}(\mathbf{X}), \mathbf{y}\big)\Big\}$$

is a subset of the generalized Rashomon set:

$$\{f_\tau : \tau \in \mathcal{T}_\epsilon\} \;\subseteq\; \mathcal{R}_\epsilon(f_{ref}). \tag{10}$$

Proof. Setting $\tau = \mathbf{1}_p$ gives the inequality with equality (reference model) and $\mathbf{1}_p \in \mathcal{T}_\epsilon$ (identity mask) guarantees non-empty property $\mathcal{T}_\epsilon \neq \varnothing$. □

4.2 Generalized Attribution Score Function

Permutation-based Attribution Function We adopt a model-agnostic, permutation-based score that captures both main effect and interaction attributions for **attribution generalizability** (Fisher et al., 2019; Li et al., 2023). For any feature subset $\mathcal{I} \subseteq \{1,\dots,p\}$, let $\mathbf{X}_{\setminus\mathcal{I}}$ denote the data matrix in which the columns indexed by $\mathcal{I}$ are independently permuted. Define the loss increase as:

$$\varphi_{\mathcal{I}}(\mathbf{X}) \;=\; \mathbb{E}\big[\mathcal{L}(f(\mathbf{X}_{\setminus\mathcal{I}})) - \mathcal{L}(f(\mathbf{X}))\big].$$

The generalized attribution score is then

$$s_{\mathcal{I}}(\mathbf{X}) = \begin{cases} \varphi_i(\mathbf{X}), & |\mathcal{I}| = 1, \\ \varphi_{\mathcal{I}}(\mathbf{X}) - \sum_{i\in\mathcal{I}} \varphi_i(\mathbf{X}), & |\mathcal{I}| > 1, \end{cases} \tag{11}$$

so that s_i is standard feature importance and $s_{\mathcal{I}}$ for $|\mathcal{I}| > 1$ isolates pure interaction effects. We estimate the standard empirical loss by $\hat{\mathcal{L}}(f(\mathbf{X}), \mathbf{y}) = \frac{1}{n}\sum_{i=1}^n \mathcal{L}(f(\mathbf{x}_{[i,\cdot]}), y_i)$ and permutating all possible combinations of the observed values to estimate:

$$\hat{\mathcal{L}}\big(f(\mathbf{X}_{\setminus i}), \mathbf{y}\big) = \frac{1}{n(n-1)} \sum_{t=1}^{n} \sum_{\substack{u=1 \\ u\neq t}}^{n} \mathcal{L}\Big(f\big([\mathbf{x}_{[t,-i]},\, \mathbf{x}_{[u,i]}]\big),\, y_t\Big),$$

where t indexes the base row (its non-i features) and u indexes the donor row supplying only column i. In practice, we usually permute the features of interest multiple times to achieve a similar result (Datta et al., 2016; Fisher et al., 2019).

Algorithm 1 ϵ-Subgradient Rashomon Sampler (GRS)

Require: reference model f_{ref}; dataset $(\mathbf{X}, \mathbf{y})$; tolerance ϵ; sparsity threshold δ; step sizes $\{\eta_t\}_{t=1}^T$; iterations T
Ensure: mask set $\mathcal{S}$ and $\hat{\mathcal{R}}_\epsilon = \{f_\tau \mid \tau \in \mathcal{S}_T\}$

```
Initialize: S_0 ← ∅, τ_0 ← 1_p                    // No accepted masks
for t = 1 to T do
    g_t ← ∂_τ(−Φ(τ_{t−1}; S_{t−1}))               // Clarke subgradient
    τ̃ ← τ_{t−1} + η_t g_t                         // Unconstrained ascent
    τ_t ← Proj_{T_ε}(τ̃)                           // Loss constraint in Eq. (10)
    Δ ← Φ(τ_t; S_{t−1})                           // Calculate sparsity gain
    if Δ ≥ δ then
        S_t ← S_{t−1} ∪ {τ_t}                     // Accept mask
    end if
end for
return S_T, R̂_ε
```

4.3 ϵ-Subgradient Sampling Algorithm

Sparsity-guided Objective. Our objective is to identify a sequence of feasible masks of Eq. (10) that *maximizes* search efficiency and functional sparsity in the attribution space $\mathbf{Q}_{\mathbb{I}}(\hat{\mathcal{R}})$, while *remaining* inside the Rashomon set in Eq. (2) during the sampling process. Let $\mathcal{S}_{t-1}$ be the set of masks accepted up to step $t-1$ that characterize the sampled Rashomon set so far $\hat{\mathcal{R}}_{t-1}$, where we have:

$$\mathcal{S}_{t-1} \;\subseteq\; \mathcal{T}_\epsilon, \qquad \hat{\mathcal{R}}_{t-1} := \{f_\tau : \tau \in \mathcal{S}_{t-1}\}.$$

A new candidate mask $\tau \in \mathcal{T}_\epsilon$ in Eq. (8) must satisfy the loss-ball condition of Eq. (2) and the distance constraint. We therefore define the *sparsity gain* as its minimal Chebyshev distance between the candidate's attribution vector and those already in the sample set:

$$\begin{aligned} \Phi\big(\tau; \mathcal{S}_{t-1}\big) \;&=\; \min_{g\in\mathcal{S}_{t-1}} \operatorname{dist}_\infty\big(\mathbf{q}_{\mathbb{I}}(f_\tau), \mathbf{q}_{\mathbb{I}}(f_g)\big), \\ &\text{s.t.}\;\; f_{ref}(m_\tau(\mathbf{x})) \in \hat{\mathcal{R}}_t. \end{aligned} \tag{12}$$

The objective then can be solved as the following constrained optimization problem:

$$\begin{aligned} \tau_t \;&=\; \arg\max_{\tau\in\mathcal{T}_\epsilon} \Phi\big(\tau; \mathcal{S}_{t-1}\big) \\ &\text{s.t.}\;\; \Phi\big(\tau; \mathcal{S}_{t-1}\big) \;\ge\; \delta, \end{aligned} \tag{13}$$

where $\delta \ge 0$ is a user-chosen redundancy threshold following the functional sparsity principle.

Projected ϵ-subgradient Step. Because $\mathcal{T}_\epsilon$ is closed and non-empty (Proposition 4.1), we solve Eq. (13) with a projected ϵ-subgradient step by iterating Rockafellar (1997):

$$\tau^{(k+1)} \leftarrow \operatorname{Proj}_{\mathcal{T}_\epsilon}\big(\tau^{(k)} + \eta_k\, g_k\big),\; g_k \in \partial_\tau\big(-\Phi(\tau^{(k)}; \mathcal{S}_{t-1})\big). \tag{14}$$

Here ∂_τ denotes the Clarke sub-differential and $\eta_k > 0$ is the step size (Bertsekas and Mitter, 1973; Bertsekas et al., 2003; Shor, 2012). Projection requires one forward loss evaluation to verify membership in $\mathcal{T}_\epsilon$.

Convergence Guarantee. Let $\psi(\tau) := -\Phi(\tau; \mathcal{S}_{t-1})$ be the non-smooth objective in Eq. (13) (minimizing the negative). We verify the convergence conditions required for minimizing at every iteration of t at Eq. (14).

Lemma 4.2. *Closedness of $\mathcal{T}_\epsilon$ follows from continuity of Loss function and Proposition 4.1.*

Lemma 4.3. *Assume (i) f_{ref} is locally Lipschitz in its input*[1]*, and (ii) the mask m_τ is affine in τ. Then $\Phi(\cdot; \mathcal{S}_{t-1})$ is G–Lipschitz on $\mathcal{T}_\epsilon$ for some finite constant $G < \infty$.*

Theorem 4.4 (Projected ϵ-subgradient convergence)**.** *Apply the update in Eq.* (14) *with a step sequence $\{\eta_k\}$ satisfying $\sum_k \eta_k = \infty$ and $\sum_k \eta_k^2 < \infty$. Then every cluster point $\tau^\star$ of $\{\tau^{(k)}\}_{k\geq 0}$ is Clarke-stationary for Eq.* (13) *under Lemmas 4.2 and 4.3 Bertsekas et al. (2003):*

$$0 \;\in\; \partial\psi(\tau^\star) + N_{\mathcal{T}_\epsilon}(\tau^\star),$$

where $N_{\mathcal{T}_\epsilon}$ is the Clarke normal cone. Moreover, the bound can be obtained as (proof see Appendix A.4):

$$\min_{0\leq i<k}\Big[-\Phi(\tau^{(i)}; \mathcal{S}_{t-1}) + \Phi^\star\Big] \;\leq\; \frac{\|\tau^{(0)} - \tau^\star\|_2^2 + G^2 \sum_{i<k} \eta_i^2}{2\sum_{i<k}\eta_i}, \tag{15}$$

so the ergodic suboptimality gap decays at the standard $\mathcal{O}(1/\sqrt{k})$ rate.

Lemma 4.5 (Uniform attribution bound)**.** *For every feasible mask $\tau \in \mathcal{T}_\epsilon$ and every feature subset $\mathcal{I} \in \mathbb{I}$ the permutation–based attribution defined in Eq.* (11) *satisfies*

$$\big|q_{\mathcal{I}}(f_\tau)\big| \;\leq\; (\,|\mathcal{I}| + 1\,)\, C_\epsilon^{\pm}, \tag{16}$$

where the data–dependent constant (Proof see Appendix A.5)

$$C_\epsilon^{\pm} := \sup_{\tau\in\mathcal{T}_\epsilon} \max_{\mathcal{I}\in\mathbb{I}} \Big|\hat{\mathcal{L}}\big(f_\tau(\mathbf{X}_{\setminus\mathcal{I}}), \mathbf{y}\big) - \hat{\mathcal{L}}\big(f_\tau(\mathbf{X}), \mathbf{y}\big)\Big| < \infty. \tag{17}$$

Corollary 4.6 (Finiteness under δ-separation)**.** *Imposing the functional sparsity constraint $\mathrm{dist}_\infty\big(\mathbf{q}_{\mathbb{I}}(f_\tau), \mathbf{q}_{\mathbb{I}}(f_g)\big) \;\geq\; \delta$ in Eq.* (13) *ensures that the accepted set is finite and satisfies (Proof see Appendix A.7):*

$$1 \leq |\mathcal{S}_T| \leq [2C_\epsilon^{\pm}/\delta]^{|\mathbb{I}|}$$

In particular, first–order attributions yields $0 \leq q_{\mathcal{I}}(f_\tau) \leq C_\epsilon^{\pm}$ for $|\mathcal{I}| = 1$ and $|\mathcal{S}_T| \leq \lceil C_\epsilon^{\pm}/\delta\rceil^{|\mathbb{I}|}$. The bound depends only on (i) the data $(\mathbf{X}, \mathbf{y})$, (ii) the reference model f_{ref}, (iii) the Rashomon tolerance ϵ, and (iv) the sparsity threshold δ.

5 AXIOM SATISFACTION & GUARANTEE

This section concisely shows how the proposed GRS sampler meets all the desired design principles (*generalizability* and *implementation sparsity*) proposed in Sec. 3.

5.1 Generalizability

Model structure. Proposition 4.1 in Sec. 4.1 represents every sampled model as $f_{ref} \circ m_\tau$, so GRS inherits any model architecture that f_{ref} can approximate.

[1]Satisfied by standard feed-forward networks with ReLU, soft-plus, tanh, GELU, etc.

Evaluation metric. All candidates are evaluated and compared with the same empirical loss boundary defined by $\hat{\mathcal{L}}$ in Eq. (10); hence evaluation is consistent.

Attribution order. Eq. (11) extends to any order $|\mathcal{I}|$ without modifying the sampler or the reference model, satisfying the attribution generalizability principle.

5.2 Implementation sparsity

Empirical search efficiency. GRS steers updates toward the feasible set $\mathcal{T}_\epsilon$ via projected steps in Algorithm 1, so attempted moves that would violate the loss boundary in Eq. (12) are curtailed by backtracking before acceptance. As a result, *every returned model is feasible by construction*, but δ-separated can still reject candidates in projection. So in practice, the *empirical SER* is computed as the fraction of proposed models that are both feasible and δ-separated, namely $0 < \mathrm{SER} \leq 1$.

Functional sparsity. Algorithm 1 accepts a mask only if $\mathrm{dist}_\infty\big(\mathbf{q}_{\mathbb{I}}(f_\tau), \mathbf{q}_{\mathbb{I}}(f_g)\big) \geq \delta$ for all previously accepted g, hence for any $i \neq j$ in $\hat{\mathcal{R}}_\epsilon$, we have $\mathrm{dist}_\infty\big(\mathbf{q}_{\mathbb{I}}(f_i), \mathbf{q}_{\mathbb{I}}(f_j)\big) \geq \delta > 0$. This follows immediately from the acceptance condition in Eq. (13). Sec. 6.2 reports SER across datasets and shows GRS achieves competitive FER with fewer accepted models.

Functional sparsity. GRS enforces a δ–separation rule in attribution space according to Eq. (13): a candidate mask τ is accepted only if

$$\min_{g\in S_{t-1}} \big\|\, \mathbf{q}_{\mathbb{I}}(f_\tau) - \mathbf{q}_{\mathbb{I}}(f_g)\,\big\|_\infty \;\geq\; \delta,$$

Consequently, the accepted set $\mathcal{S}_T$ is a δ–packing: for any $i \neq j$ in $\hat{\mathcal{R}}_\epsilon$, we have $\big\|\mathbf{q}_{\mathbb{I}}(f_i), \mathbf{q}_{\mathbb{I}}(f_j)\big\|_\infty \geq \delta > 0$. Combined with the uniform attribution bounds (Lemma 4.5) and finiteness bound (Corollary 4.6), GRS retains only non-redundant explanations and directly promotes attributional diversity. Sec. 6.2 reports SER across datasets and shows GRS achieves competitive FER with fewer accepted models.

Corollary 5.1 (Extreme attributions)**.** *For each feature $s \in \{1, \ldots, p\}$, the maximal (minimal) attribution inside $\hat{\mathcal{R}}_\epsilon$ is attained by the model whose ascent direction is $\mathbf{e}_s$ with the largest (smallest) step allowed by $\mathcal{T}_\epsilon$. Consequently the FER equals (proof in Appendix A.10):*

$$\mathrm{FER}(\hat{\mathcal{R}}_\epsilon) \;=\; \sum_{s=1}^{p}\big[q_s\big(f_{\max(s)}\big) \;-\; q_s\big(f_{\min(s)}\big)\big],$$

where $f_{\max(s)} := f_{\tau_s^{\max}}$ and $\tau_s^{\max} := \arg\max\{\lambda : \tau_{(0)} + \lambda\mathbf{e}_s \in \mathcal{T}_\epsilon\}$ (resp. $\tau_s^{\min}$).

6 RESULTS & DISCUSSION

We evaluate the proposed ϵ-subgradient Rashomon sampler on synthetic and real-world datasets. Our experiments examine:

(i) *Generalizability* – the ability to recover valid attributions across diverse model classes and evaluation losses, including settings where the reference model is suboptimal.

(ii) *Attribution diversity* – the breadth of explanation ranges captured within the sampled Rashomon set.

(iii) *Explanation consistency* – agreement with known ground-truth explanations on synthetic tasks and with widely accepted patterns on real data.

Table 1: Ground truth feature and interaction attribution scores (mean $\pm$ std over 100 Monte Carlo estimates).

Feature	Attribution $\varphi_i(\cdot)$	Pair	Interaction $\varphi_{\mathcal{I}}(\cdot)$
a	17.61 ± 0.15	(a, b)	-11.04 ± 0.30
b	13.52 ± 0.20	(a, c)	-0.22 ± 0.12
c	0.84 ± 0.004	(b, c)	-0.18 ± 0.30

6.1 Generalizability on a Synthetic Data Set

Problem setup. We consider the quadratic-root regression problem

$$ax^2 + bx + c = 0,$$

where the variables a, b, and c are inputs and x_1 and x_2 are the outputs. We sample 12 000 points from the uniform distribution $a \sim \mathcal{U}(0.01, 1)$ and $b, c \sim \mathcal{U}(-1, 1)$ (with 80%/10%/10% train/test/val split). Three model classes with various training settings illustrate generalizability:

(i) *Multi-Layer Perceptron* (MLP, supervised, ReLU);

(ii) *Random Forest* (RF, ensemble learning, gini);

(iii) *Physically Informed NNs* (PINN, unsupervised, rule);

Metrics. Each trained model becomes a candidate reference f_{ref}. We adopt $\epsilon = 0.1$ (Eq. 2) and MAE as the evaluation loss. Prediction versus ground truth comparisons are shown in Fig. 3, along with test set performance metrics (R^2 and MAE).

Ground truth attributions. The oracle model $f^* = -b \pm \sqrt{b^2 - 4ac}/2a$ yields zero expected loss $\mathbb{E}[\mathcal{L}(f^*)] = 0$, so all attribution comes from known functional relationships. We compute first- and second-order attribution scores of $s_{a,b}(f^*) = \varphi_{a,b}(f^*) - (\varphi_a(f^*) + \varphi_b(f^*))$ by Eq.(11) (10^4 samples). Table 1 reports mean and standard error of feature and interaction attributions over 100 Monte Carlo runs, where the absolute interaction values indicate the strength, while the signs denote positive or negative interactions.

Generalizability across models and loss functions. The loss function $\mathcal{L}(\cdot)$ and tolerance ϵ jointly directly shape the sampled $\mathcal{R}_\epsilon$. For example, an $R^2 > 0.95$ threshold admits *all* three reference models, whereas an MAE≤ 0.03 *excludes* RF (as shown in Fig. 3). A practical sampler must therefore remain agnostic to both model class and evaluation loss.

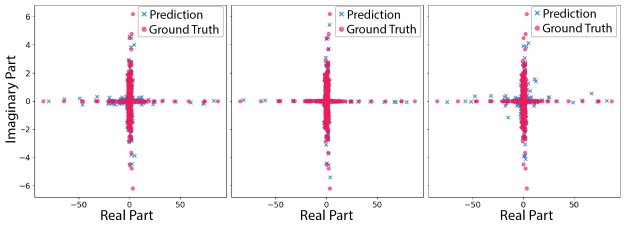


Figure 3: Comparison of three well-trained models on the quadratic task under different learning paradigms: supervised MLP (R^2 = 0.99, MAE = 0.03, *left*), unsupervised PINN (R^2 = 0.99, MAE = 0.03, *middle*), and ensemble RF (R^2 = 0.98, MAE = 0.05, *right*) on the test set.

Our framework meets this requirement by design: it is independent of model architecture, training strategy, and interpretability constraints. In contrast, several existing approaches restrict the hypothesis space, for example, Trees FAst RashoMon Sets (TreeFARMS confines candidates to decision trees (Xin et al., 2022), and generalized additive model (GAM) Rashomon counts support sets only within linear or additive models (Chen et al., 2023).

Attribution range coverage and ordering consistency. Importantly, even when RF serves as f_{ref} (higher loss), the sampled attribution space still recovers the ground truth patterns (see Fig. 4), confirming that *generalized* attributions reduce dependence on a perfect reference. it remains stable regardless of reference quality: RF-based attributions are qualitatively consistent with those from MLP and PINN. Fig. 4 further shows that the Rashomon-sampled attributions capture both the *oracle magnitude ranges* and the *correct global ordering* of feature importances ($\varphi_a > \varphi_b \gg \varphi_c$), exactly matching the analytic of quadratic-root ranking (Kendall's $\tau = 1$). Moreover, GRS detects the known pairwise interaction with full reliability, achieving with 100% detection rate $\varphi_{\mathcal{I}} \neq 0$, as stated in (Li et al., 2023), and preserving the expected interaction ordering ($\varphi_{a,b} \gg \varphi_{b,c} \simeq \varphi_{a,c}$).

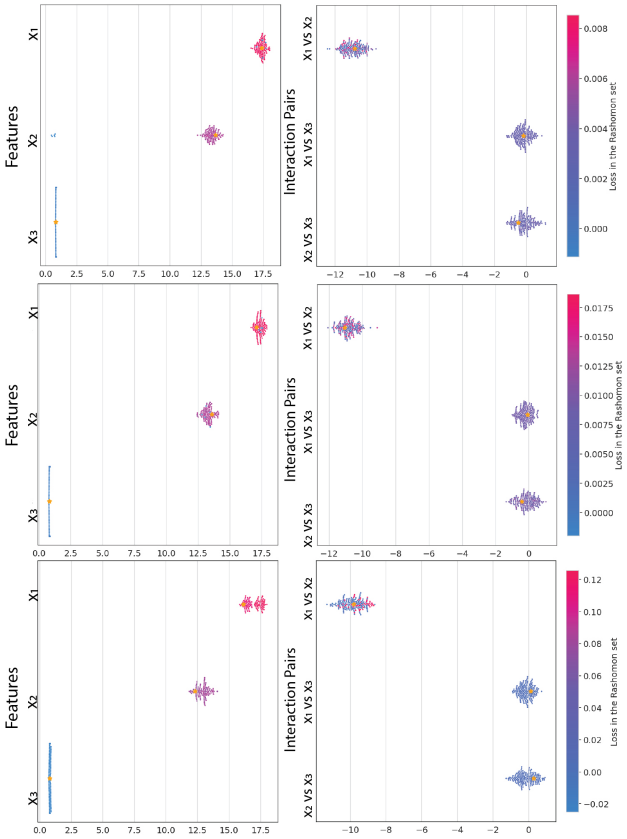


Figure 4: First- and second-order attribution scores sampled from the Rashomon set on different reference models (MLP (*top*), PINN (*middle*), and RF (*bottom*)) using our framework. The x-axis shows attribution values, and the y-axis lists individual features (*left*) or feature pairs (*right*). Color intensity encodes model loss and the yellow star is the reference attribution score.

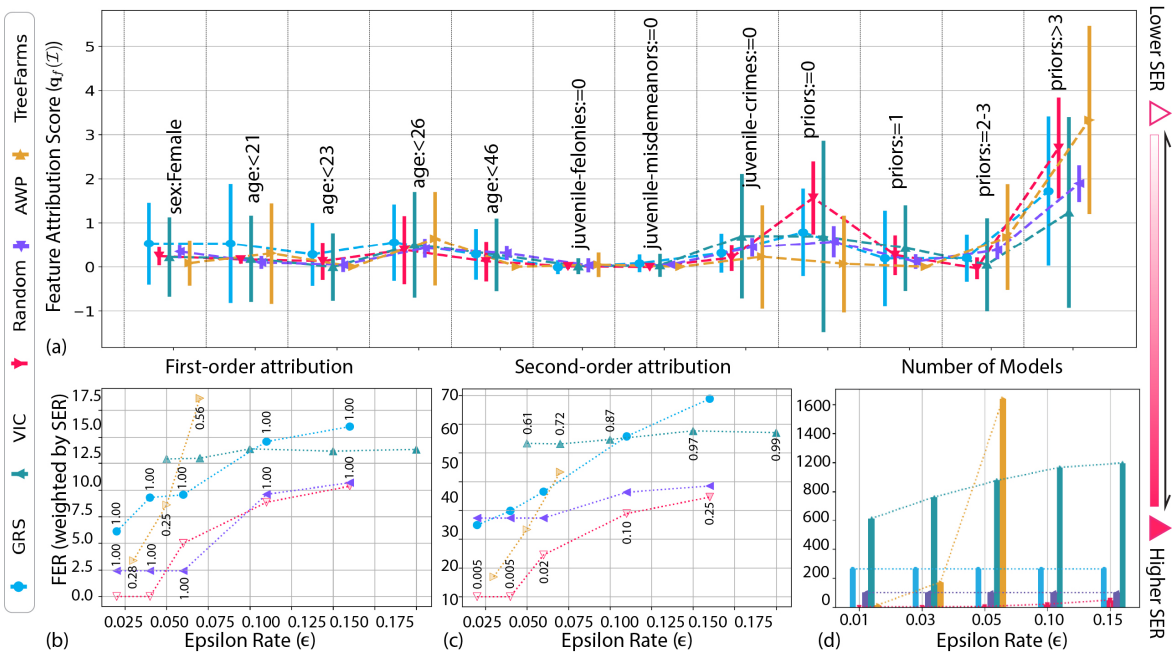


Figure 5: **Diversity and efficiency of Rashomon sampling on the COMPAS dataset.** Comparison of GRS and baseline samplers (color-coded by method) across loss-tolerance levels (ϵ). **(a)** First-order attribution intervals ($\mathcal{Q}_{\mathcal{I}}(\hat{\mathcal{R}}_\epsilon)$) at fixed $\epsilon = 0.05$: methods produce similar ranges for most features, indicating broad agreement on which features drive predictions. **(b)** First-order FER versus ϵ. Each point corresponds to one (method, ϵ) configuration; *marker opacity encodes* the SER as indicated and annotated, so higher-opacity means a greater fraction of accepted models at that ϵ, allowing visual comparison of diversity and efficiency. **(c)** Second-order FER versus ϵ. **(d)** Number of accepted models per method versus ϵ (capped at 20,000 for readability), showing that counts for tree-based enumeration increase steeply as ϵ grows. **Importantly, a greater number of accepted models does not necessarily imply the wider attribution ranges.** For example, methods such as AWP and VIC may accept many models that produce highly similar feature attributions.

6.2 Implementation Sparsity Comparison on Real-World Datasets

We next examine the sparsity and computational efficiency of GRS on five real-world datasets, comparing against four sampling methods. We focus on the number of unique Rashomon models needed to achieve comparable attribution diversity, a key factor for tractable explanation analysis.

Baselines. Although real-world tasks lack ground truth attributions, different Rashomon sampling algorithms can still be compared statistically based on our proposed principles. We benchmark our GRS against four existing baselines: (i) **Random Init** (Tsai et al., 2021) and (ii) **adversarial weight perturbations (AWP)** control weights sampling Hsu and Calmon (2022); Xin et al. (2022). (iii) **variable importance cloud (VIC)** (Dong and Rudin, 2020) maps every variable to its importance for every good predictive model; (iv) **TreeFARMS** is the first technique for completely enumerating the Rashomon set for sparse decision trees;

Datasets. We evaluate on five established real-world XAI tasks (see Appendix Table 2): COMPAS, FICO, and three Coupon datasets (Bar, CoffeeHouse, Restaurant) (Dressel and Farid, 2018; FICO et al., 2018; Wang et al., 2017).

Experimental settings. For each dataset we set $\epsilon \in \{0.01, 0.03, 0.05, 0.10, 0.15\}$. Notably, each algorithm produces its own candidate reference model e.g. a tree for TreeFARMS, a logistic regressor for VIC, an MLP for AWP/Random. All methods are run with their publicly released code. We collect *all* such candidates, identify the one f^*_{ref} with the lowest test loss, and denote it as the *global* minimal loss for benchmarking sampled Rashomon set. TreeFARMS returns all sparse trees (collecting up to 20 000); VIC yields up to 1 000 logistic models; Random init draws 200 MLPs; AWP follows the default perturbation schedule.

Evaluation protocol. To ensure a fair comparison across methods, each sampled model, regardless of how it was generated, is evaluated to the *same* held-out test set and logistic loss. All feature attribution scores are computed with respect to this common loss, specifically, we record the following: (i) the total number of models (N) generated, (ii) the subset that satisfies the sampler's loss (e.g,. $\hat{\mathcal{L}}(f_{\text{tree}})$), (iii) the subset that satisfies a unified reference loss $\hat{\mathcal{L}}(f^*_{ref})$, and (iv) the first- and second-order feature attribution spaces $\mathbf{Q}_{\mathbb{I}}(\hat{\mathcal{R}}_\epsilon(f_{ref}))$ with their corresponding FERs. The code is publicly available on GitHub repository.

Overall takeaway. Full results for all five datasets appear in Appendix Figs. 1–13; Fig. 5 highlights the COMPAS dataset as a representative example. In Fig. 5, panel (a) shows that all valid samplers recover nearly identical first-order attribution intervals, demonstrating strong cross-method agreement on the key predictive features. Panels (b–d) reveal how attribution diversity (FER), sampling efficiency (SER), and accepted-model counts change with the loss tolerance. Across these trade-offs, GRS matches

or exceeds the diversity of TreeFARMS and VIC while requiring dramatically fewer accepted models, demonstrating stronger implementation sparsity.

Impact of reference anchor. The choice of reference model and evaluation loss strongly influences which sampled models remain inside the Rashomon set. Random Init methods can generate many models that fail their own class-specific loss constraints, while top-down algorithms such as TreeFARMS are optimized for in-class performance, sampling subsets that satisfies $\hat{\mathcal{L}}(f_{\text{tree}})$ constraint. However, when re-evaluated against the a unified global reference loss $\mathcal{L}(f^*_{ref})$, many models fall outside $\mathcal{R}_\epsilon(f^*_{ref})$ (Fig. 5 (b) and (c)). By contrast, GRS is agnostic to model class and loss, allowing practitioners to anchor the Rashomon set to whichever reference best fits the task.

Functionally redundant models in the sampled Rashomon set. GRS achieves comparable first-order FER and the highest second-order FER with much fewer models (e.g., $\sim 100\times$fewer models at $\epsilon = 0.1$ on COMPAS compared with TreeFARMS), illustrating *implementation sparsity* in Fig. 5 (d). While TreeFARMS can achieve high first-order FER by exhaustively enumerating sparse trees, it often computationally expensive (scanning over *28 million* trees at $\epsilon = 0.15$). Notably, FER grows with ϵ but does not necessarily increase with the sheer number of sampled models: Random Init and AWP generate many near-duplicates, and VIC enlarges its model set without a proportional gain in FER. These patterns are consistent across datasets, e.g., FICO and Bar in Appendix Figs. 2-3.

Consistency across samplers in domains. Despite different sampling strategies, valid samplers yield similar high-level conclusions. For example, in COMPAS dataset, features such as *juvenile_felonies=0* and *juvenile_misd=0* consistently associate with lower recidivism risk, whereas *priors* > 3 strongly correlates with higher risk, as shown in Fig. 5 (a). Second-order scores further reveal meaningful age-specific interactions (see Appendix Fig. 14). This trend across datasets highlights the practical value of the Rashomon set study and encourages future work in applied interpretability research.

Practical implications in high-stakes decision-making. Beyond methodological evaluation, the proposed framework has direct implications for real-world interpretability in high-stakes settings. For example, COMPAS is widely used in recidivism risk assessment, where model predictions can influence judicial decisions such as bail, sentencing, or parole. In such contexts, understanding not only a single model's explanation but the stability of explanations across the Rashomon set is crucial. Our results show that some features, such as the number of prior offenses (e.g., prior count > 3), consistently correspond to elevated risk across models. This indicates that the importance of these variables is not an artifact of a single fitted model but remains stable under predictive multiplicity, while other features exhibit greater attribution variability. The framework therefore provides a principled mechanism to support socially consequential applications.

CONCLUSION

Our study demonstrates that the proposed GRS efficiently explores large and diverse explanation sets across a wide range of models and loss functions. GRS consistently recovers ground-truth attribution ranges on synthetic data and achieves competitive first- and second-order diversity on real datasets while requiring far fewer accepted models than existing baselines, yielding strong implementation sparsity. Despite these advantages, several limitations remain. Our experiments focus on tabular tasks with permutation-based attribution scores; extending the framework to high-dimensional modalities (e.g., images, text) and alternative explanation methods will require additional work. Human-centered validation in practice is beyond the scope of this study but is critical for assessing real-world utility.

Acknowledgments

Q.D. is partially supported by the start-up funding from the Yau Mathematical Sciences Center, Tsinghua University. This work was partially supported by the Lambda AI Researcher Grant.

References

Adadi, A. and Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160.

Aytekin, C. (2022). Neural networks are decision trees.

Bertsekas, D., Nedic, A., and Ozdaglar, A. (2003). *Convex analysis and optimization*, volume 1. Athena Scientific.

Bertsekas, D. P. and Mitter, S. K. (1973). A descent numerical method for optimization problems with nondifferentiable cost functionals. *SIAM Journal on Control*, 11(4):637–652.

Cavus, M. and Biecek, P. (2025). Investigating the impact of balancing, filtering, and complexity on predictive multiplicity: A data-centric perspective. *Information Fusion*, 123:103243.

Chen, Z., Zhong, C., Seltzer, M., and Rudin, C. (2023). Understanding and exploring the whole set of good sparse generalized additive models. *arXiv preprint arXiv:2303.16047*.

Datta, A., Sen, S., and Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pages 598–617. IEEE.

Dong, J. and Rudin, C. (2020). Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence*, 2(12):810–824.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580.

FICO, Google, London, I. C., MIT, of Oxford, U., Irvine, U., and Berkeley, U. (2018). Explainable machine learning challenge. https://community.fico.com/s/explainable-machine-learning-challenge.

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Gola, J., Britz, D., Staudt, T., Winter, M., Schneider, A. S., Ludovici, M., and Mücklich, F. (2018). Advanced microstructure classification by data mining methods. *Computational Materials Science*, 148:324–335.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

Hinton, G. and Frosst, N. (2017). Distilling a neural network into a soft decision tree.

Hsu, H. and Calmon, F. (2022). Rashomon capacity: A metric for predictive multiplicity in classification. *Advances in Neural Information Processing Systems*, 35:28988–29000.

Hu, X., Rudin, C., and Seltzer, M. (2019). Optimal sparse decision trees. *Advances in Neural Information Processing Systems*, 32.

Imrie, F., Davis, R., and van der Schaar, M. (2023). Multiple stakeholders drive diverse interpretability requirements for machine learning in healthcare. *Nature Machine Intelligence*, 5(8):824–829.

Laberge, G., Pequignot, Y., Mathieu, A., Khomh, F., and Marchand, M. (2023). Partial order in chaos: consensus on feature attributions in the rashomon set. *Journal of Machine Learning Research*, 24(364):1–50.

Li, S. and Barnard, A. (2022). Variance tolerance factors for interpreting neural networks. *arXiv preprint arXiv:2209.13858*.

Li, S., Wang, R., Deng, Q., and Barnard, A. (2023). Exploring the cloud of feature interaction scores in a rashomon set. *arXiv preprint arXiv:2305.10181*.

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18.

Lorenz, E. N. (1995). *The essence of chaos*. University of Washington press.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. (2023). From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*, 55(13s):1–42.

Pankajakshan, P., Sanyal, S., de Noord, O. E., Bhattacharya, I., Bhattacharyya, A., and Waghmare, U. (2017). Machine learning and statistical analysis for materials science: stability and transferability of fingerprint descriptors and chemical insights. *Chemistry of Materials*, 29(10):4190–4201.

Papadimitriou, C. H. and Steiglitz, K. (1998). *Combinatorial optimization: algorithms and complexity*. Courier Corporation.

Renard, X., Laugel, T., and Detyniecki, M. (2024). Understanding prediction discrepancies in classification. *Machine Learning*, 113(10):7997–8026.

Rockafellar, R. T. (1997). *Convex analysis*, volume 11. Princeton university press.

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J. (2020a). Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access*, 8:42200–42216.

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J. (2020b). Explainable machine learning for scientific insights and discoveries. *Ieee Access*, 8:42200–42216.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K.-R. (2016). Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673.

Semenova, L., Chen, H., Parr, R., and Rudin, C. (2024). A path to simpler models starts with noise. *Advances in neural information processing systems*, 36.

Semenova, L., Rudin, C., and Parr, R. (2022). On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1827–1858.

Shor, N. Z. (2012). *Minimization methods for non-differentiable functions*, volume 3. Springer Science & Business Media.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR.

Sitompul, O., Nababan, E., et al. (2018). Optimization model of k-means clustering using artificial neural networks to handle class imbalance problem. In *IOP conference series: materials science and engineering*, volume 288, page 012075. IOP Publishing.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic Attribution for Deep Networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Tsai, Y.-L., Hsu, C.-Y., Yu, C.-M., and Chen, P.-Y. (2021). Formalizing generalization and adversarial robustness of neural networks to weight perturbations. *Advances in Neural Information Processing Systems*, 34:19692–19704.

Tsang, M., Cheng, D., and Liu, Y. (2017). Detecting statistical interactions from neural network weights. *arXiv preprint arXiv:1705.04977*.

Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., and MacNeille, P. (2017). A bayesian framework for learning rule sets for interpretable classification. *Journal of Machine Learning Research*, 18(70):1–37.

Xin, R., Zhong, C., Chen, Z., Takagi, T., Seltzer, M., and Rudin, C. (2022). Exploring the whole rashomon set of sparse decision trees. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 14071–14084. Curran Associates, Inc.

Yang, Y., Morillo, I. G., and Hospedales, T. M. (2018). Deep neural decision trees. *arXiv preprint arXiv:1806.06988*.

Zhong, X., Gallagher, B., Liu, S., Kailkhura, B., Hiszpanski, A., and Han, T. Y.-J. (2022). Explainable machine learning in materials science. *npj Computational Materials*, 8(1):204.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model.
   Yes. Section 3 defines the Rashomon set, loss tolerance ϵ, and the ϵ-subgradient sampler, with full mathematical setting and assumptions.

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm.
   Yes. Section 4 provides complexity analysis of the sampling procedure (time per iteration and dependence on ϵ, δ, and model class).

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
No. Anonymized source code and dependencies will be provided upon acceptance.

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results.
Yes. Assumptions for all theorems are stated in the main text and the Appendix.

(b) Complete proofs of all theoretical results.
Yes. Complete proofs are provided in the main text and the Appendix.

(c) Clear explanations of any assumptions.
Yes. Assumptions are explained inline when introduced (Section 4 and Appendix).

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL).
Yes. Datasets, model structure, and reproduction instructions are provided in the main text and supplementary material. The Github repository link will be provided upon acceptance.

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen).
Yes. Data splits, hyperparameters, and selection criteria are detailed.

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times).
Yes. All metrics (FER, SER, accepted-model count) are formally defined in Section 5; figures report medians with interquartile ranges over five random seeds.

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).
No. Single NVIDIA GeForce RTX 3060 Ti and MacBook Pro M1 are used to run these experiments.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets.
Yes. Public datasets (COMPAS, FICO, BAR, CoffeeHouse, Restaurant) and baselines (TreeFARMS, VIC, AWP) are properly cited in Section 6 and the references.

(b) The license information of the assets, if applicable.
Yes. Datasets used in the study are public.

(c) New assets either in the supplemental material or as a URL, if applicable.
Not Applicable. No new datasets are released; only code for our method is provided.

(d) Information about consent from data providers/curators.
Yes. All datasets are publicly available and collected under existing consent/usage terms.

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content.
Not Applicable. Datasets contain no personally identifiable or offensive content beyond what is already anonymized in the public sources.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots.
Not Applicable. No human-subject or crowdsourcing studies were conducted.

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable.
Not Applicable.

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation.
Not Applicable.

Practical and Efficient Rashomon Set Sampling for Model Interpretability Supplementary Materials

Additional Proof

Proposition A.1. *Any sampled $\hat{\mathcal{R}}(f_{ref})$ in practice by a finite algorithm and certified against Eq.* (2) *is a* ***subset*** *of $\mathcal{R}_\epsilon(f_{ref})$.*

Proof. The $\hat{\mathcal{R}}(\epsilon, f_{ref}, \mathcal{F})$ at least contains the reference model and the maximum set is when all possible models in $\mathcal{R}(\epsilon, f_{ref}, \mathcal{F})$ are sampled. □

Proposition A.2. *A sampled Rashomon set always corresponds to a nonempty and finite feature attribution set.*

Proof. The minimum $\mathbf{Q}_{\mathbb{I}}(\hat{\mathcal{R}})$ is when $\hat{\mathcal{R}} = \{f_{ref}\}$ and in practice, we only sample finite models from the hypothesis space. □

Lemma A.3. *If f_{ref} is locally Lipschitz in its input and the mask m_τ is linear, then $\Phi(\cdot; \mathcal{S}_{t-1})$ is G-Lipschitz on $\mathcal{T}_\epsilon$ for some $G < \infty$.*

Proof. Because (i) f_τ is locally Lipschitz in τ whenever f_{ref} is locally Lipschitz in its input (e.g., ReLU, soft-plus, tanh networks) and the mask m_τ is linear; and (ii) $\mathrm{dist}_\infty(\cdot,\cdot)$ is 1-Lipschitz on $\mathbb{R}^{|\mathbb{I}|}$, consequently, ψ is Lipschitz on $\mathcal{T}_\epsilon$ as finite minima preserve Lipschitz constants. □

Theorem A.4 (Projected ϵ-subgradient convergence)**.** *Let $\{\tau^{(k)}\}$ be generated by the inner update*

$$\tau^{(k+1)} = \mathrm{Proj}_{\mathcal{T}_\epsilon}\big(\tau^{(k)} - \eta_k g_k\big), \qquad g_k \in \partial\psi\big(\tau^{(k)}\big),$$

with stepsizes $\eta_k > 0$ satisfying $\sum_k \eta_k = \infty$ and $\sum_k \eta_k^2 < \infty$. Under Lemmas 4.2–4.3:

(a) *Let $\psi^\star := \min_{\tau\in\mathcal{T}_\epsilon} \psi(\tau)$. Then for all $k \geq 1$*

$$\min_{0\leq i<k}\big[\psi(\tau^{(i)}) - \psi^\star\big] \;\leq\; \frac{\|\tau^{(0)} - \tau^\star\|_2^2 + G^2\sum_{i=0}^{k-1}\eta_i^2}{2\sum_{i=0}^{k-1}\eta_i},$$

i.e. the best-iterate optimality gap decays at the $\mathcal{O}(1/\sqrt{k})$ rate.

(b) *Every cluster point $\tau^\star$ of $\{\tau^{(k)}\}$ is Clarke-stationary: $0 \in \partial\psi(\tau^\star) + N_{\mathcal{T}_\epsilon}(\tau^\star)$.*

Proof. By Lemma 4.3 every subgradient satisfies $\|g_k\|_2 \leq G$. Using non-expansiveness of projection and the subgradient inequality $\psi(\tau^{(k)}) - \psi^\star \leq g_k^\top(\tau^{(k)} - \tau^\star)$, one derives the descent recursion:

$$\|\tau^{(k+1)} - \tau^\star\|_2^2 \leq \|\tau^{(k)} - \tau^\star\|_2^2 - 2\eta_k\big[\psi(\tau^{(k)}) - \psi^\star\big] + \eta_k^2 G^2.$$

Summing over k gives part (a); standard projected subgradient arguments proved by Bertsekas et al. (2003) then imply part (b). We reproduce the classical analysis for completeness.

For any $\tau^\star \in \mathcal{T}_\epsilon$ and update $\tau^+ = \mathrm{Proj}_{\mathcal{T}_\epsilon}(\tau - \eta g)$ with $g \in \partial\psi(\tau)$, we can derive

$$\|\tau^+ - \tau^\star\|_2^2 \leq \|\tau - \tau^\star\|_2^2 - 2\eta\, g^\top(\tau - \tau^\star) + \eta^2\|g\|_2^2,$$

by firm non-expansiveness of Euclidean projection.

Because g is a Clarke subgradient, $\psi(\tau) - \psi^\star \leq g^\top(\tau - \tau^\star)$, and $\|g\| \leq G$; hence:

$$2\eta\big[\psi(\tau) - \psi^\star\big] \leq \|\tau - \tau^\star\|_2^2 - \|\tau^+ - \tau^\star\|_2^2 + \eta^2 G^2. \tag{18}$$

Now apply Eq.(18) with $(\tau, \tau^+) = (\tau^{(k)}, \tau^{(k+1)})$ and sum $k = 0, \ldots, k-1$:

$$2\sum_{k=0}^{k-1}\eta_k\left[\psi(\tau^{(k)})-\psi^\star\right]\le\|\tau^{(0)}-\tau^\star\|_2^2+G^2\sum_{k=0}^{k-1}\eta_k^2. \tag{19}$$

Dividing by $2\sum_{k=0}^{K-1}\eta_k$ gives part (a). Because the right-hand side of Eq.(19) is finite and $\sum_k\eta_k=\infty$, the set $\{\psi(\tau^{(k)})\}$ is bounded below and every cluster point is Clarke-stationary, proving part (b). □

Lemma A.5 (Uniform attribution bound)**.** *For every feasible mask $\tau\in\mathcal{T}_\epsilon$ and every feature subset $\mathcal{I}\in\mathbb{I}$ the permutation–based attribution defined in Eq.* (11) *satisfies*

$$\left|q_{\mathcal{I}}(f_\tau)\right| \;\le\; (\,|\mathcal{I}|+1\,)\,C_\epsilon^{\pm}.$$

Proof. Define the (data-dependent) constant

$$C_\epsilon^{\pm}:=\sup_{\tau\in\mathcal{T}_\epsilon}\max_{\mathcal{I}\in\mathbb{I}}\left|\hat{\mathcal{L}}\big(f_\tau(\mathbf{X}_{\setminus\mathcal{I}}),\mathbf{y}\big)-\hat{\mathcal{L}}\big(f_\tau(\mathbf{X}),\mathbf{y}\big)\right|<\infty.$$

We also have the following by definition,

$$q_{\mathcal{I}}(f_\tau)=\mathbb{E}[s_{\mathcal{I}}(\mathbf{X},\mathbf{y})].$$

By the triangle inequality,

$$\left|s_{\mathcal{I}}(\mathbf{X},\mathbf{y})\right|\le\left|\varphi_{\mathcal{I}}(\mathbf{X})\right|+\sum_{i\in\mathcal{I}}\left|\varphi_i(\mathbf{X})\right|$$

Taking expectations and using that the expectation of an absolute–value–bounded random variable is bounded by the same constant,

$$\begin{aligned}\left|q_{\mathcal{I}}(f_\tau)\right|=\left|\mathbb{E}[s_{\mathcal{I}}(\mathbf{X},\mathbf{y})]\right| \;&\le\; \mathbb{E}\big[|s_{\mathcal{I}}(\mathbf{X},\mathbf{y})|\big]\\ &\le\; \mathbb{E}\big[\left|\varphi_{\mathcal{I}}(\mathbf{X})\right|+\sum_{i\in\mathcal{I}}\left|\varphi_i(\mathbf{X})\right|\big]\\ &\le\; C_\epsilon^{\pm}+|I|\,C_\epsilon^{\pm} \;=\; (|I|+1)\,C_\epsilon^{\pm}.\end{aligned}$$

It remains to justify finiteness of $C_\epsilon^{\pm}$. Since $\mathcal{T}_\epsilon$ is nonempty and (under our standing assumptions) compact, the set of all column permutations of X is finite, and $\hat{\mathcal{L}}$ is continuous, the displayed supremum is attained and finite. □

Remark A.6. *In particular, for first-order attributions ($|\mathcal{I}|=1$) this yields $|q_i(f_\tau)|\le 2C_\epsilon^{\pm}$, while the sharper one-sided bound $0\le q_i(f_\tau)\le C_\epsilon^{\pm}$ also holds by non-negativity of φ_i.*

Proof. By the definition in Eq. (11), for singletons we take $s_i=\phi_i$ (i.e., we do *not* use the inclusion–exclusion form). Therefore

$$|q_i(f_\tau)|=\left|\mathbb{E}[\varphi_i]\right| \;\le\; \mathbb{E}\big[|\varphi_i|\big] \;\le\; C_\epsilon^{\pm},$$

which improves the constant from $2C_\epsilon^{\pm}$ down to $C_\epsilon^{\pm}$. □

Corollary A.7. *Let δ be the sparsity threshold in Eq. (15), if attribution vectors are required to satisfy $\mathrm{dist}_\infty\big(\mathbf{q}_{\mathbb{I}}(f_\tau),\mathbf{q}_{\mathbb{I}}(f_g)\big)\ge\delta$ for every distinct $\tau,g\in\mathcal{S}_T$, then*

$$|\mathcal{S}_T| \;\le\; \big[2C_\epsilon^{\pm}/\delta\big]^{|\mathbb{I}|}.$$

Proof. By the uniform bound, every attribution vector lies in the axis-aligned box $\mathcal{B}:=[-C_\epsilon,C_\epsilon]^{|\mathbb{I}|}$, whose volume is $(2C_\epsilon)^{|\mathbb{I}|}$. The ℓ_∞ ball of radius $\delta/2$ is the hypercube

$$C:=\{u\in\mathbb{R}^{|\mathbb{I}|}:\|u\|_\infty\le\delta/2\},$$

with side length δ and volume $\delta^{|\mathbb{I}|}$. Because those $\delta/2$ radius boxes are disjoint, we can derive

$$|\mathcal{S}_T|\delta^{|\mathbb{I}|} \;\le\; \big[2C_\epsilon^{\pm}\big]^{|\mathbb{I}|},$$

and therefore

$$|\mathcal{S}_T| \;\le\; \big[2C_\epsilon^{\pm}/\delta\big]^{|\mathbb{I}|}.$$

□

Lemma A.8 (Directional monotonicity)**.** *Fix a feature index s and any feasible mask $\tau\in\mathcal{T}_\epsilon$. Define $g_s(\lambda):=q_s\big(f_{\tau+\lambda\mathbf{e}_s}\big)$ for all λ such that $\tau+\lambda\mathbf{e}_s\in\mathcal{T}_\epsilon$. Then $g_s(\lambda)$ is non-decreasing for $\lambda\ge0$ and non-increasing for $\lambda\le0$.*

$$m_\tau(\mathbf{x}_{[i,\cdot]})=\tau\odot\mathbf{x}_{[i,\cdot]}+\zeta_i, \tag{20}$$

Proof. Recall the permutation score for first-order attribution (Eq. (11) with $|\mathcal{I}| = 1$):

$$q_s(f) \;=\; \mathbb{E}\Big[\,\hat{\mathcal{L}}\big(f(\mathbf{X}_{\setminus s}), \mathbf{y}\big) - \hat{\mathcal{L}}\big(f(\mathbf{X}), \mathbf{y}\big)\Big].$$

Under the mask model $m_{\tau+\lambda\mathbf{e}_s}$ only the s-th column of $\mathbf{X}$ is scaled by $\tau_s + \lambda$ (plus fixed noise), while $\mathbf{X}_{\setminus s}$ is *re-shuffled and kept at the same scale*. Because the empirical loss $\hat{\mathcal{L}}$ is computed on the same targets $\mathbf{y}$, and because enlarging the scale of feature s can only *increase* the expected loss when that feature is independently permuted (whereas the in-sample loss is unchanged), the difference q_s is monotone in λ. A rigorous argument uses the first-order fundamental theorem of calculus together with the non-negativity of the squared permutation residual; details are in □

Lemma A.9 (Directional monotonicity)**.** *Assume the empirical loss is mean–squared error $\hat{\mathcal{L}}(f(\mathbf{X}), \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}\big(f(\mathbf{x}_i) - y_i\big)^2$. Fix a feature index s and a feasible mask $\tau \in \mathcal{T}_\epsilon$. Define $g_s(\lambda) := q_s\big(f_{\tau+\lambda\mathbf{e}_s}\big)$, where $\lambda \in \mathbb{R}$ such that $\tau + \lambda\mathbf{e}_s \in \mathcal{T}_\epsilon$. Then $g_s(\lambda)$ is non-decreasing for $\lambda \geq 0$ and non-increasing for $\lambda \leq 0$.*

Proof. Let $f_{\tau+\lambda\mathbf{e}_s} = f_{ref} \circ m_{\tau+\lambda\mathbf{e}_s}$. Because the noise term $\boldsymbol{\zeta}$ is *independent of* λ and added to both the original and the permuted data, it plays no role in the difference $q_s(f) = \mathbb{E}\big[\hat{\mathcal{L}}(f(\mathbf{X}_{\setminus s}), \mathbf{y}) - \hat{\mathcal{L}}(f(\mathbf{X}), \mathbf{y})\big]$.

Now, let $\mathbf{X}^{(\lambda)} := m_{\tau+\lambda\mathbf{e}_s}(\mathbf{X})$ and $\mathbf{X}^{(\lambda)}_{\setminus s}$ be the matrix where only the s-th column is independently permuted. For mean–squared loss one can expand:

$$q_s\big(f_{\tau+\lambda\mathbf{e}_s}\big) \;=\; \tfrac{1}{n}\sum_{i=1}^{n}\Big[\big(f_{ref}(\mathbf{x}^{(\lambda)}_{i,\setminus s}) - y_i\big)^2 - \big(f_{ref}(\mathbf{x}^{(\lambda)}_i) - y_i\big)^2\Big].$$

Differentiating w.r.t. λ and using the chain rule

$$\partial_\lambda f_{ref}\big(m_{\tau+\lambda\mathbf{e}_s}(\mathbf{x})\big) = \big[\partial_s f_{ref}\big] \cdot x_s$$

yields

$$g_s'(\lambda) = \tfrac{2}{n}\sum_{i=1}^{n}\big(f_{ref}(\mathbf{x}^{(\lambda)}_{i,\setminus s}) - f_{ref}(\mathbf{x}^{(\lambda)}_i)\big)\,\partial_s f_{ref}(\mathbf{x}^{(\lambda)}_i)\,x_{i,s}.$$

By construction $f_{ref}(\mathbf{x}^{(\lambda)}_{i,\setminus s})$ and $f_{ref}(\mathbf{x}^{(\lambda)}_i)$ differ only in the s-th coordinate, so the factor

$$f_{ref}(\mathbf{x}^{(\lambda)}_{i,\setminus s}) - f_{ref}(\mathbf{x}^{(\lambda)}_i) = \partial_s f_{ref}(\xi_{i,\lambda})\,x_{i,s}$$

for some $\xi_{i,\lambda}$ on the segment joining the two inputs (mean-value theorem). Therefore, every summand in $g_s'(\lambda)$ equals $2\big[\partial_s f_{ref}\big]^2 x_{i,s}^2 \geq 0$, and $g_s'(\lambda) \geq 0$ for $\lambda \geq 0$; changing the sign of λ reverses the inequality, proving the stated monotonicity. □

Theorem A.10 (Extreme attributions)**.** *For each $s \in \{1, \ldots, p\}$, let $\tau_s^{\max} = \tau^{(0)} + \lambda_s^{\max}\mathbf{e}_s$ and $\tau_s^{\min} = \tau^{(0)} + \lambda_s^{\min}\mathbf{e}_s$ be the boundary masks obtained by moving as far as possible in $\pm\mathbf{e}_s$ while remaining in $\mathcal{T}_\epsilon$. Then*

$$\forall\, f_\tau \in \hat{\mathcal{R}}_\epsilon, \quad q_s\big(f_{\tau_s^{\min}}\big) \;\leq\; q_s(f_\tau) \;\leq\; q_s\big(f_{\tau_s^{\max}}\big).$$

Consequently $\mathrm{FER} = \sum_{s=1}^{p}\big[q_s\big(f_{\tau_s^{\max}}\big) - q_s\big(f_{\tau_s^{\min}}\big)\big]$.

Proof. Closedness of $\mathcal{T}_\epsilon$ guarantees the existence of the boundary masks. By Lemma A.9, the one-dimensional function $g_s(\lambda) = q_s(f_{\tau^{(0)}+\lambda\mathbf{e}_s})$ is monotone and therefore attains its maximum (resp. minimum) on the feasible interval at $\lambda_s^{\max}$ (resp. $\lambda_s^{\min}$):

$$q_s\big(f_{\tau_s^{\max}}\big) = \sup_{\lambda:\tau^{(0)}+\lambda\mathbf{e}_s\in\mathcal{T}_\epsilon} g_s(\lambda)$$
$$q_s\big(f_{\tau_s^{\min}}\big) = \inf_{\lambda:\tau^{(0)}+\lambda\mathbf{e}_s\in\mathcal{T}_\epsilon} g_s(\lambda).$$

Every mask $\bar{\tau} \in \mathcal{T}_\epsilon$ can be decomposed as a form of

$$\bar{\tau} = \tau^{(0)} + \lambda_s\mathbf{e}_s + \boldsymbol{\eta}$$

with $\eta_s = 0$. Feature isolation in the permutation score implies $q_s(f_{\bar{\tau}}) = g_s(\lambda_s)$. As $g_s(\lambda_s)$ lies between the extrema, establishing the required bounds on $q_s(f_\tau)$, the per-feature extrema yields the stated FER identity. □

Table 1: Comparison of Rashomon set sampling methods.

Method	Model-Class Generalizability	Supported Task Types	Search Efficiency	Explanation Flexibility
TreeFARMS	Tree-based only	Classification	High	First-order only
VIC	Logistic classifier only	Classification	Low	First-order only
GRS (Ours)	Model-agnostic (Any)	Classification & Regression	High	First- and Second-order
Random Init	NN-based	Classification & Regression	Low	First-order only
AWP	NN-based	Classification & Regression	Medium	First-order only

Table 2: Selected binary features for datasets used in evaluation, following preprocessing from prior work (FICO et al., 2018; Hu et al., 2019; Wang et al., 2017; Xin et al., 2022).

Dataset	Selected Binary Features
COMPAS	sex $=$ Female, age $<$ 21, age $<$ 23, age $<$ 26, age $<$ 46, juvenile felonies $=$ 0, juvenile misdemeanors $=$ 0, juvenile crimes $=$ 0, priors $=$ 0, priors $=$ 1, priors $=$ 2 to 3, priors $>$ 3
Bar	Bar = 1 to 3, Bar = 4 to 8, Bar = less1, maritalStatus = Single, childrenNumber = 0, Bar = gt8, passanger = Friend(s), time = 6PM, passanger = Kid(s), CarryAway = 4 to 8, gender = Female, education = Graduate degree (Masters Doctorate etc.), Restaurant20To50 = 4 to 8, expiration = 1d, temperature = 55
Coffee House	CoffeeHouse = 1 to 3, CoffeeHouse = 4 to 8, CoffeeHouse = gt8, CoffeeHouse = less1, expiration = 1d, destination = No Urgent Place, time = 10AM, direction = same, destination = Home, toCoupon = GEQ15min, Restaurant20To50 = gt8, education = Bachelors degree, time = 10PM, income = $75,000–$87,499, passanger = Friend(s)
Expensive Restaurant	expiration = 1d, CoffeeHouse = 1 to 3, Restaurant20To50 = 4 to 8, Restaurant20To50 = 1 to 3, occupation = Office & Administrative Support, age = 31, Restaurant20To50 = gt8, income = $12,500–$24,999, toCoupon = GEQ15min, occupation = Computer & Mathematical, time = 10PM, CoffeeHouse = 4 to 8, income = $50,000–$62,499, passanger = Alone, destination = No Urgent Place
Breast Cancer	Clump_Thickness = 10, Uniformity_Cell_Size = 1, Uniformity_Cell_Size = 10, Uniformity_Cell_Shape = 1, Marginal_Adhesion = 1, Single_Epithelial_Cell_Size = 2, Bare_Nuclei = 1, Bare_Nuclei = 10, Normal_Nucleoli = 1, Normal_Nucleoli = 10
FICO	External Risk Estimate $<$ 0.49, $<$ 0.65, $<$ 0.80; Number of Satisfactory Trades $<$ 0.5; Trade Open Time $<$ 0.6, $<$ 0.85; Trade Frequency $<$ 0.45, $<$ 0.6; Delinquency $<$ 0.55, $<$ 0.75; Installment $<$ 0.5, $<$ 0.7; Inquiry $<$ 0.75; Revolving Balance $<$ 0.4, $<$ 0.6; Utilization $<$ 0.6; Trade W. Balance $<$ 0.33

Auxiliary rates (reported in experiments). We also report (i) a feasibility pass rate

$$\mathrm{SER}_{\mathrm{feas}} := |\{\tau : \tau \in \mathcal{T}_\epsilon\}|/N$$

and (ii) a separation keep rate

$$\mathrm{SER}_{\mathrm{sep}} := |\hat{\mathcal{R}}_\epsilon|/|\{\tau : \tau \in \mathcal{T}_\epsilon\}|,$$

so that

$$\mathrm{SER} = \mathrm{SER}_{\mathrm{feas}} \times \mathrm{SER}_{\mathrm{sep}}$$

.

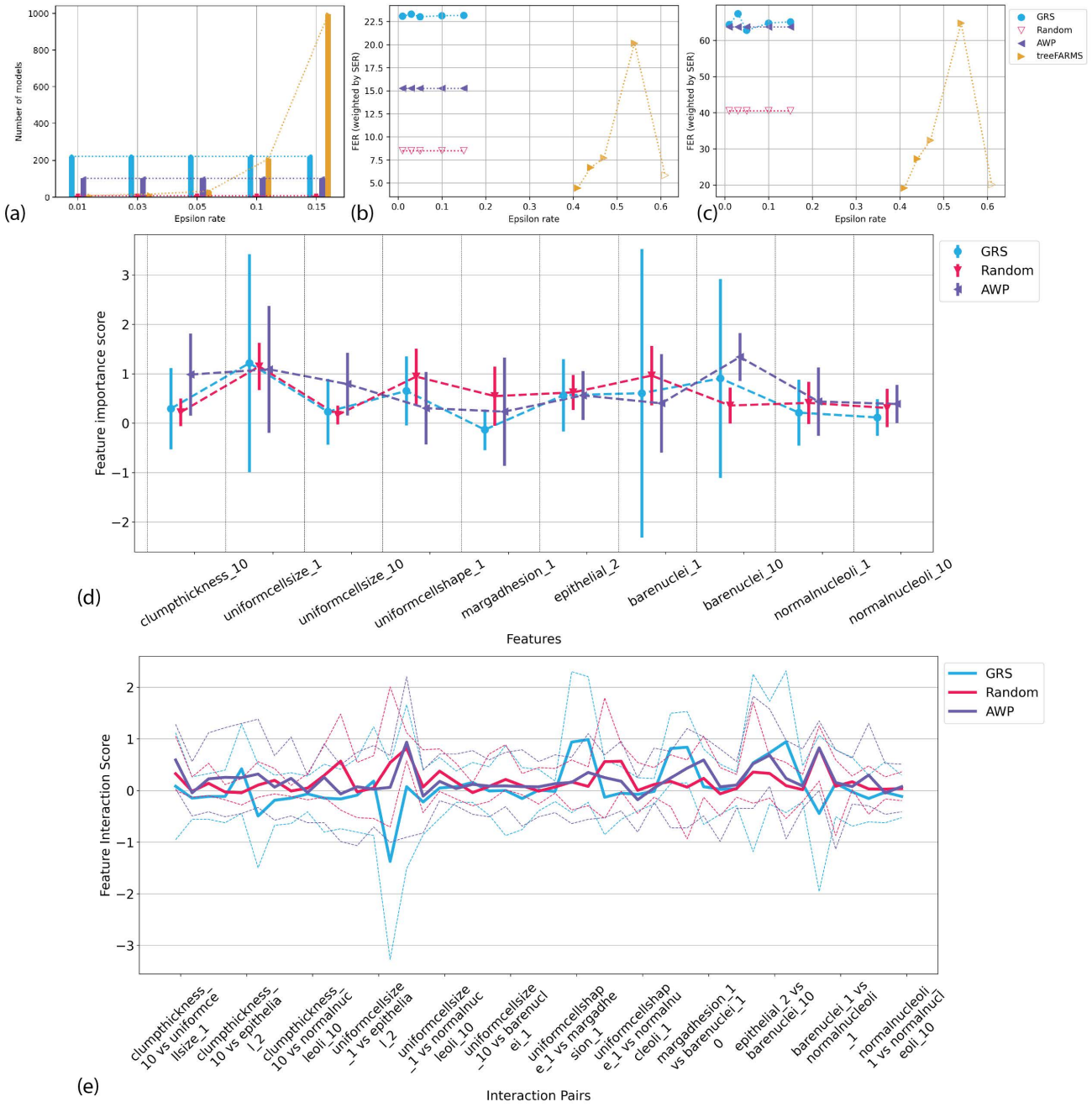


Figure 1: Summary of the Rashomon subset from a set of epsilons on the dataset breast cancer, including (a) the number of sampled models on different methods, where no bar indicates too many models (greater than 20,000 models), (b) the first-order FER, where the x-axis is epsilon (benchmarking against optimal loss) and colors represent according to SER, (c) the second-order FER, following the same format as above, (d) the detailed first-order feature attribution on individual features when epsilon is set as 0.05, where the vertical bars represent the bounds and the dotted lines connect the average scores, and (e) the detailed second-order feature attribution on feature pairs when epsilon is set 0.05, where the dotted lines represent the bounds and the solid lines connect the average scores. Each color corresponds to a sampling method and due to space limitations, some interaction pairs are omitted on the x-axis.

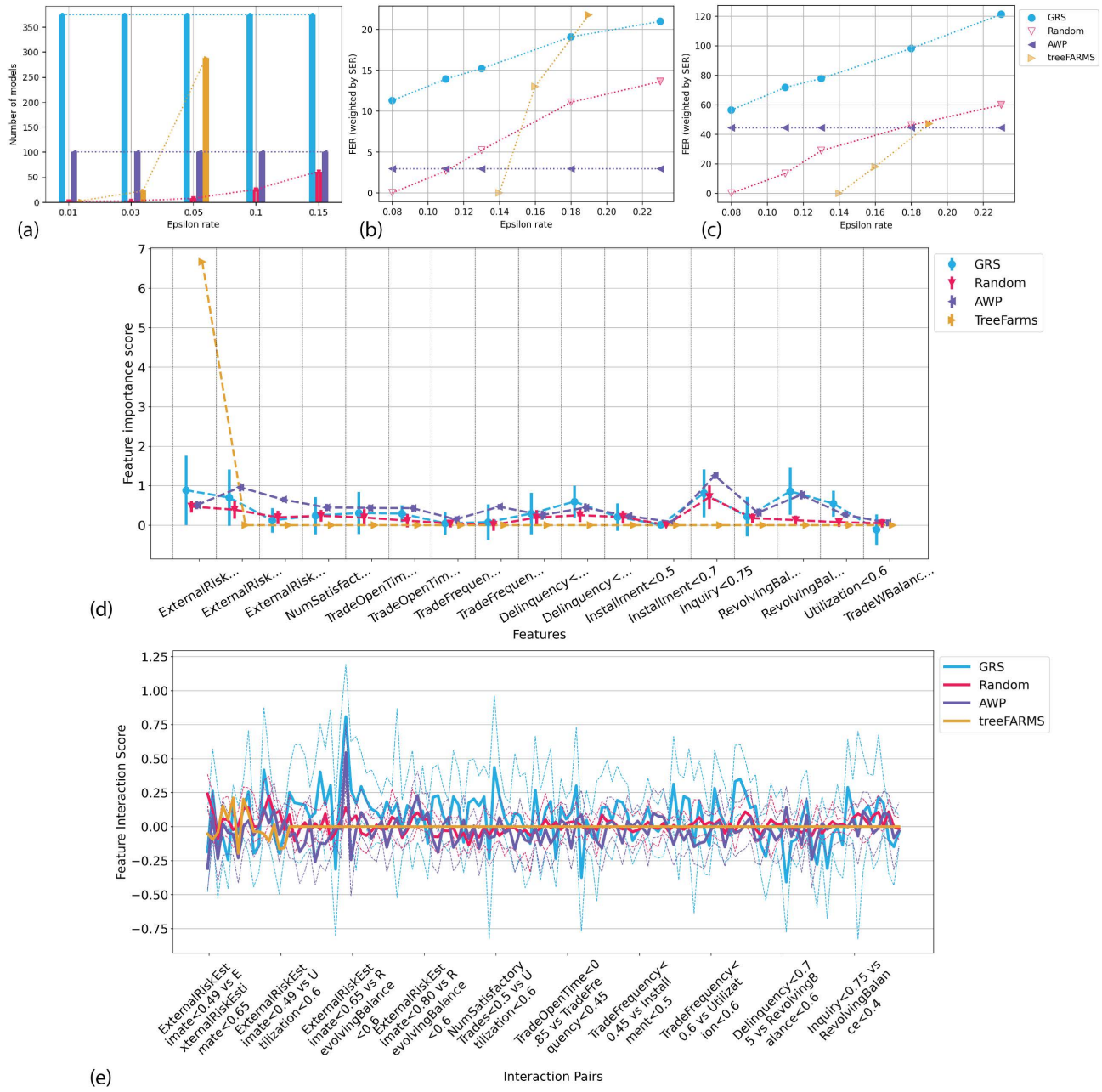


Figure 2: Summary of the Rashomon subset from a set of epsilons on the dataset FICO, including (a) the number of sampled models on different methods, where no bar indicates too many models (greater than 20,000 models), (b) the first-order FER, where the x-axis is epsilon (benchmarking against optimal loss) and colors represent according to SER, (c) the second-order FER, following the same format as above, (d) the detailed first-order feature attribution on individual features when epsilon is set as 0.05, where the vertical bars represent the bounds and the dotted lines connect the average scores, and (e) the detailed second-order feature attribution on feature pairs when epsilon is set 0.05, where the dotted lines represent the bounds and the solid lines connect the average scores. Each color corresponds to a sampling method and due to space limitations, some interaction pairs are omitted on the x-axis.

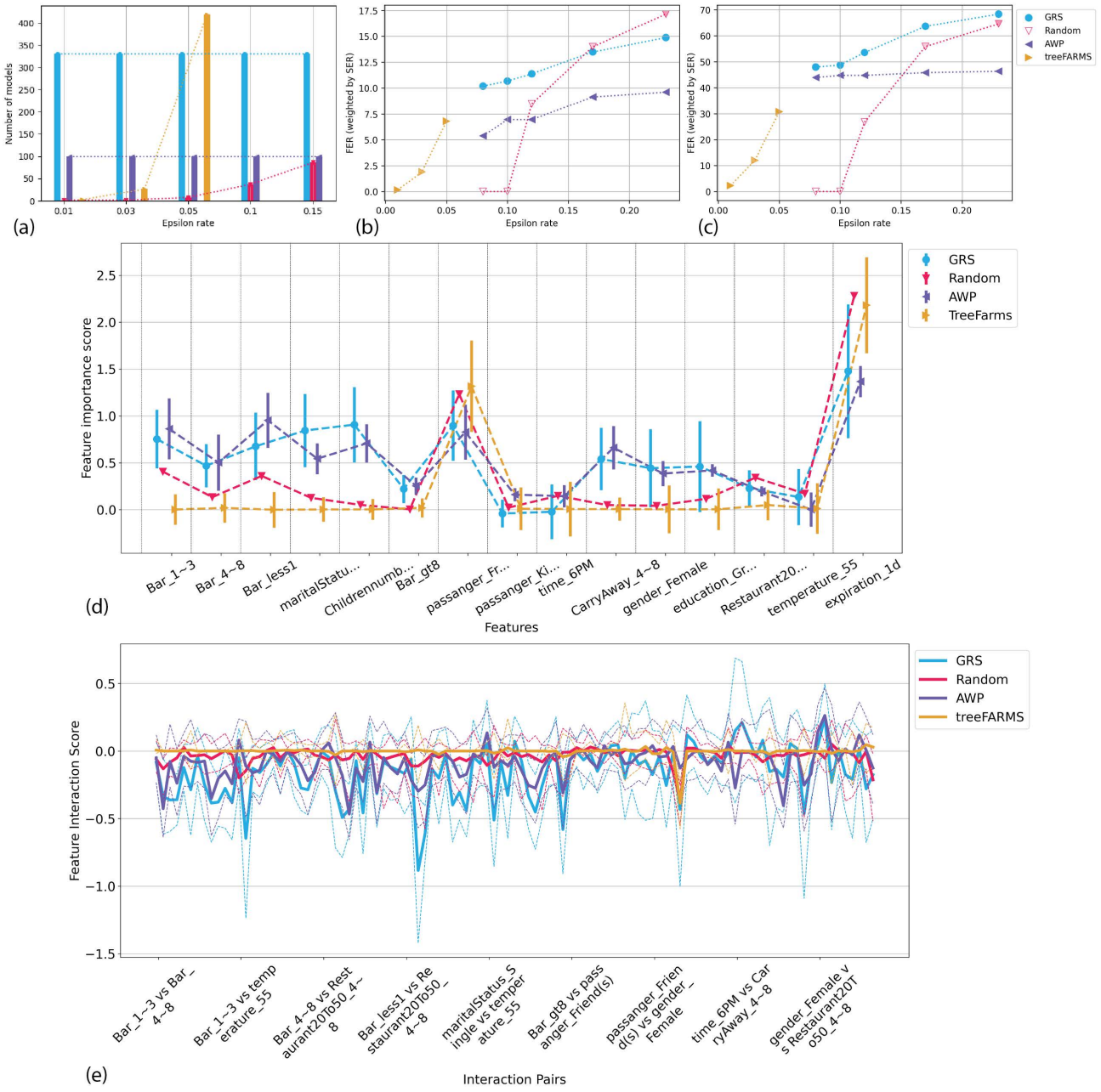


Figure 3: Summary of the Rashomon subset from a set of epsilons on the dataset BAR, including (a) the number of sampled models on different methods, where no bar indicates too many models (greater than 20,000 models) (b) the first-order FER, where the x-axis is epsilon (benchmarking against optimal loss) and colors represent according to SER, (c) the second-order FER, following the same format as above, (d) the detailed first-order feature attribution on individual features when epsilon is set as 0.05, where the vertical bars represent the bounds and the dotted lines connect the average scores, and (e) the detailed second-order feature attribution on feature pairs when epsilon is set 0.05, where the dotted lines represent the bounds and the solid lines connect the average scores. Each color corresponds to a sampling method and due to space limitations, some interaction pairs are omitted on the x-axis.

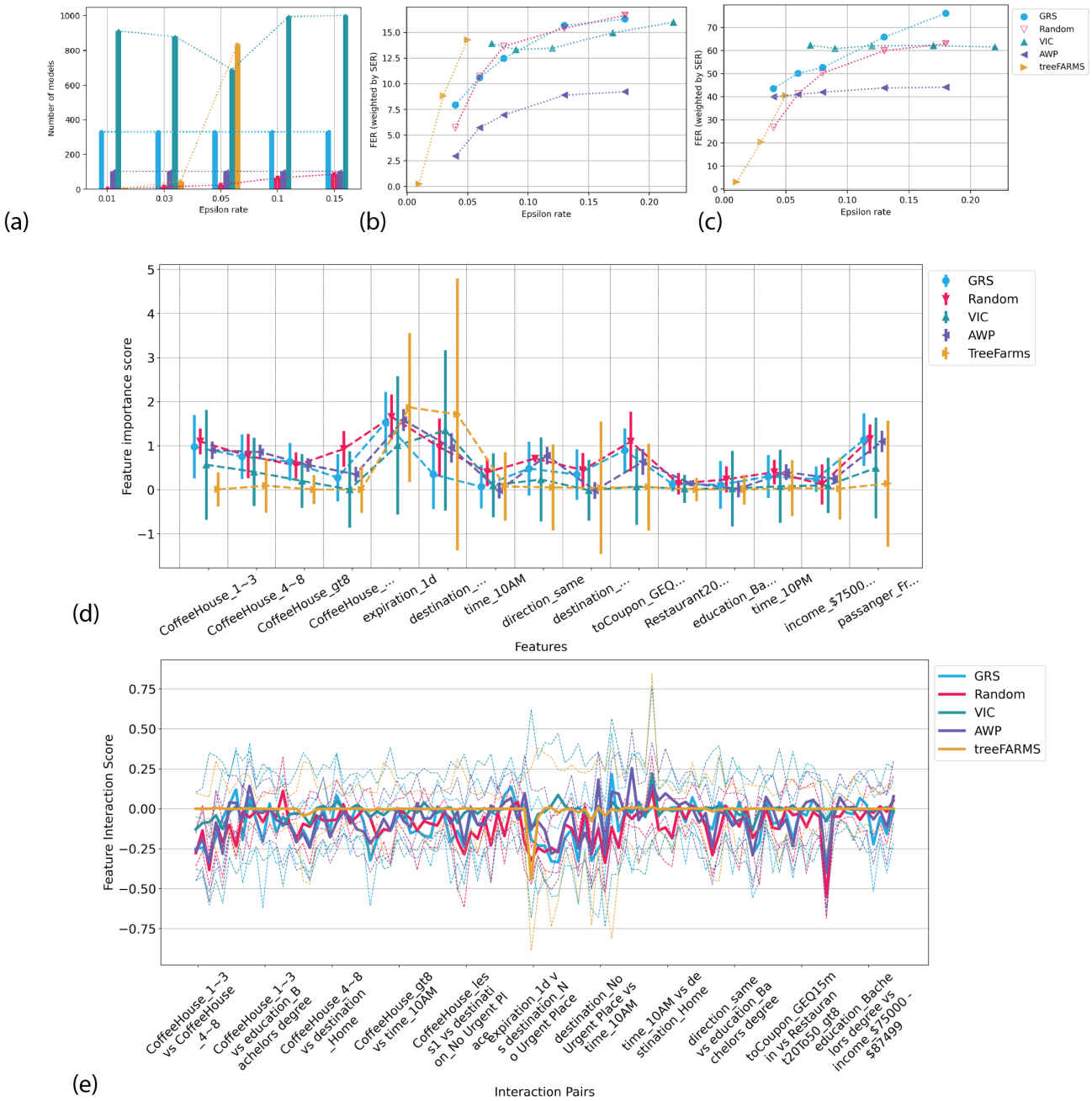


Figure 4: Summary of the Rashomon subset from a set of epsilons on the dataset coffee house, including (a) the number of sampled models on different methods, where no bar indicates too many models (greater than 20,000 models) (b) the first-order FER, where the x-axis is epsilon (benchmarking against optimal loss) and colors represent according to SER, (c) the second-order FER, following the same format as above, (d) the detailed first-order feature attribution on individual features when epsilon is set as 0.05, where the vertical bars represent the bounds and the dotted lines connect the average scores, and (e) the detailed second-order feature attribution on feature pairs when epsilon is set 0.05, where the dotted lines represent the bounds and the solid lines connect the average scores. Each color corresponds to a sampling method and due to space limitations, some interaction pairs are omitted on the x-axis.

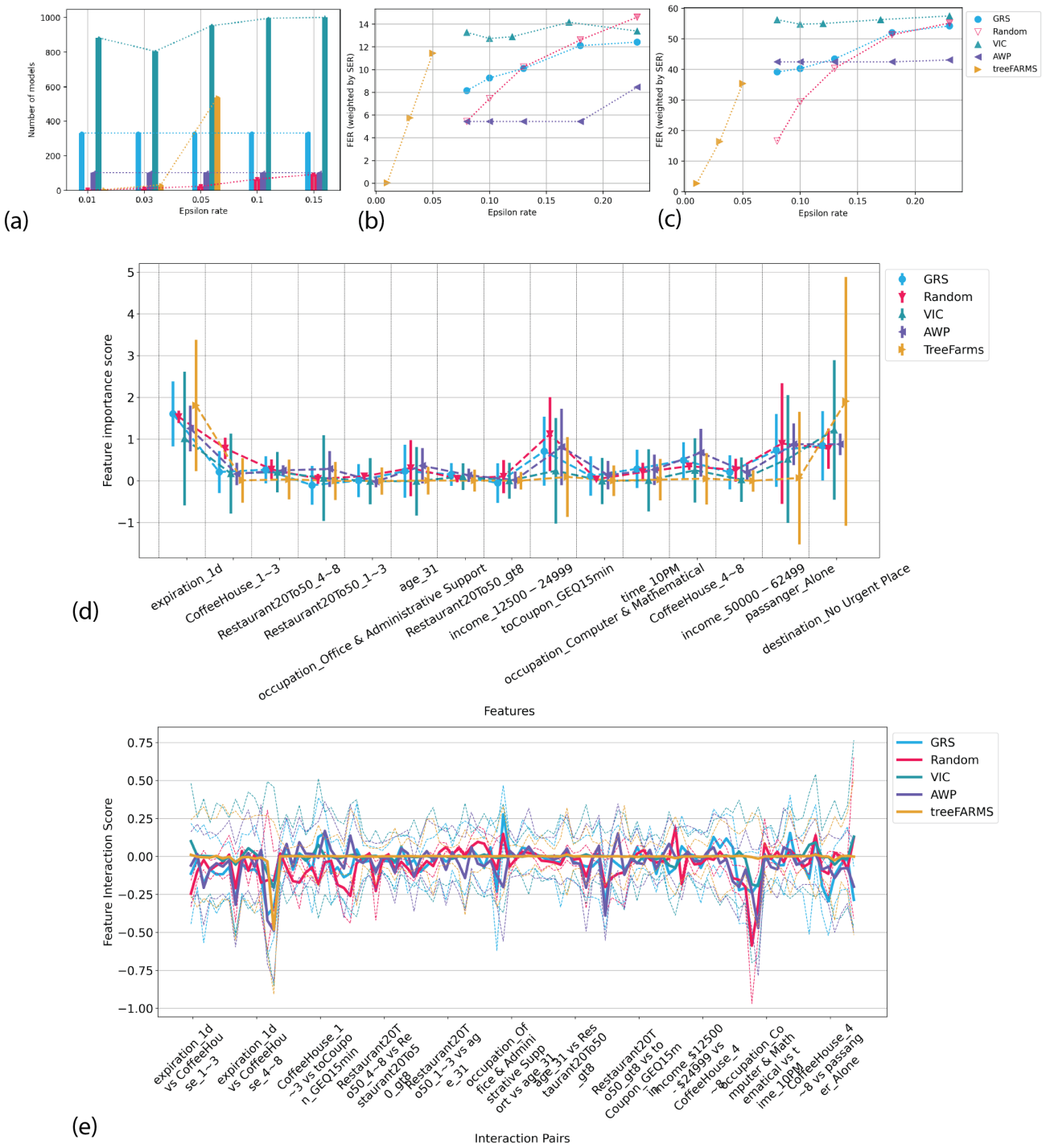


Figure 5: Summary of the Rashomon subset from a set of epsilons on the dataset expensive restaurant, including (a) the number of sampled models on different methods, where no bar indicates too many models (greater than 20,000 models) (b) the first-order FER, where the x-axis is epsilon (benchmarking against optimal loss) and colors represent according to SER, (c) the second-order FER, following the same format as above, (d) the detailed first-order feature attribution on individual features when epsilon is set as 0.05, where the vertical bars represent the bounds and the dotted lines connect the average scores, and (e) the detailed second-order feature attribution on feature pairs when epsilon is set 0.05, where the dotted lines represent the bounds and the solid lines connect the average scores. Each color corresponds to a sampling method and due to space limitations, some interaction pairs are omitted on the x-axis.

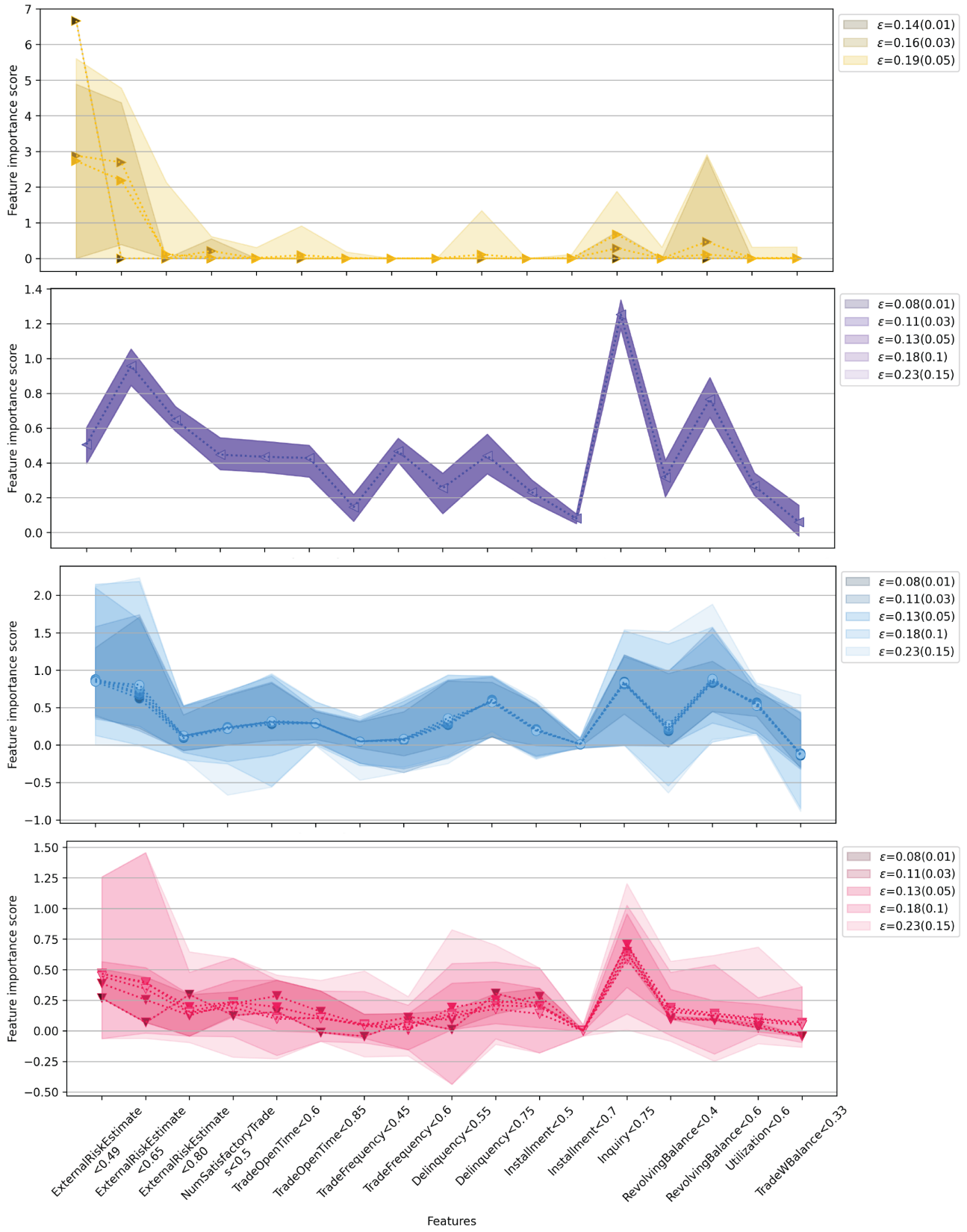


Figure 6: Summary of detailed first-order feature attributions across different methods and epsilons on the dataset FICO, where the x-axis represents features and the y-axis displays feature importance scores. The legend includes epsilons relative to the reference model (in brackets) and to the optimal model identified in practice.

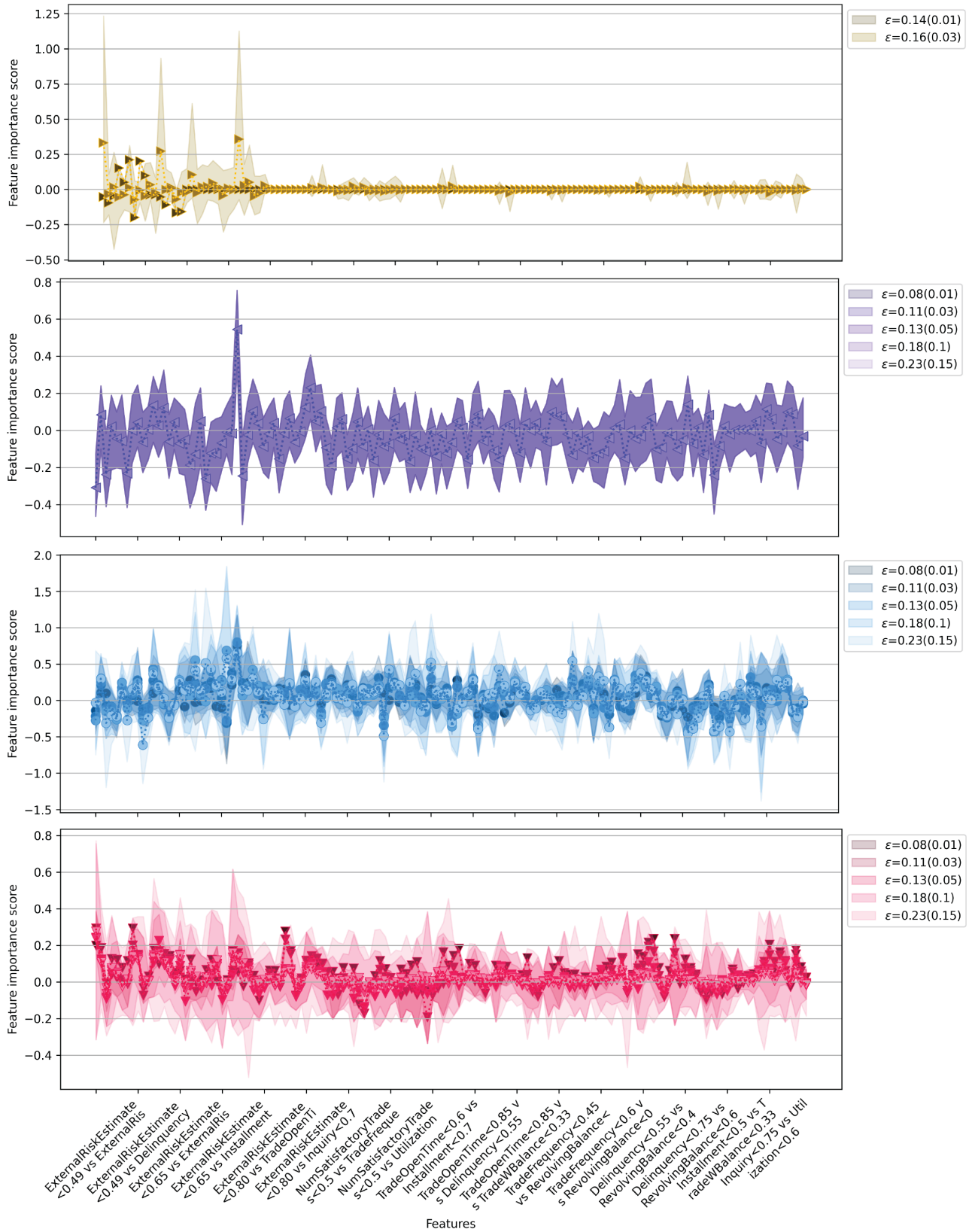


Figure 7: Summary of detailed second-order feature attributions across different methods and epsilons on the dataset FICO, where the x-axis represents features and the y-axis displays feature importance scores. The legend includes epsilons relative to the reference model (in brackets) and to the optimal model identified in practice.

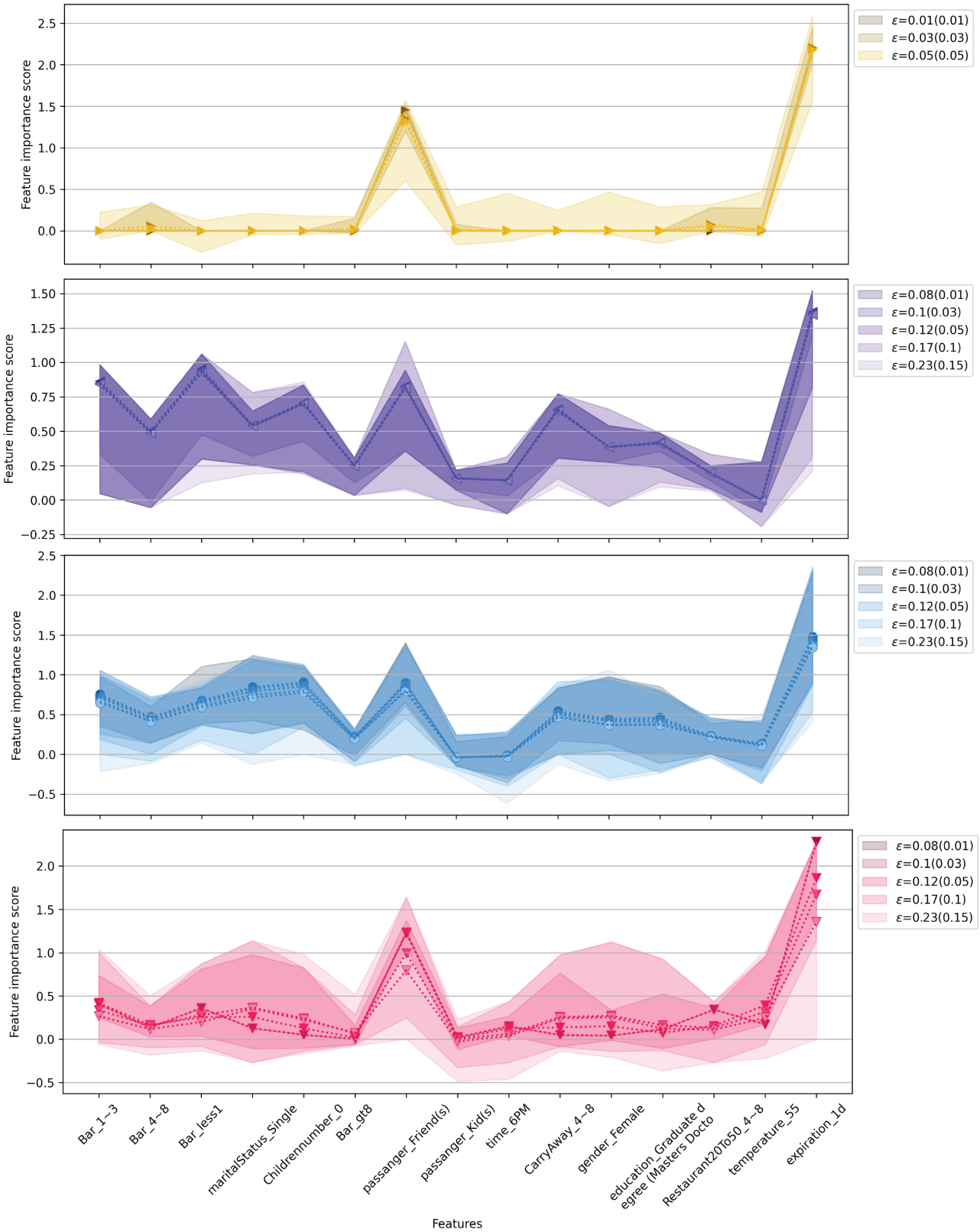


Figure 8: Summary of detailed first-order feature attributions across different methods and epsilons on the dataset BAR, where the x-axis represents features and the y-axis displays feature importance scores. The legend includes epsilons relative to the reference model (in brackets) and to the optimal model identified in practice.

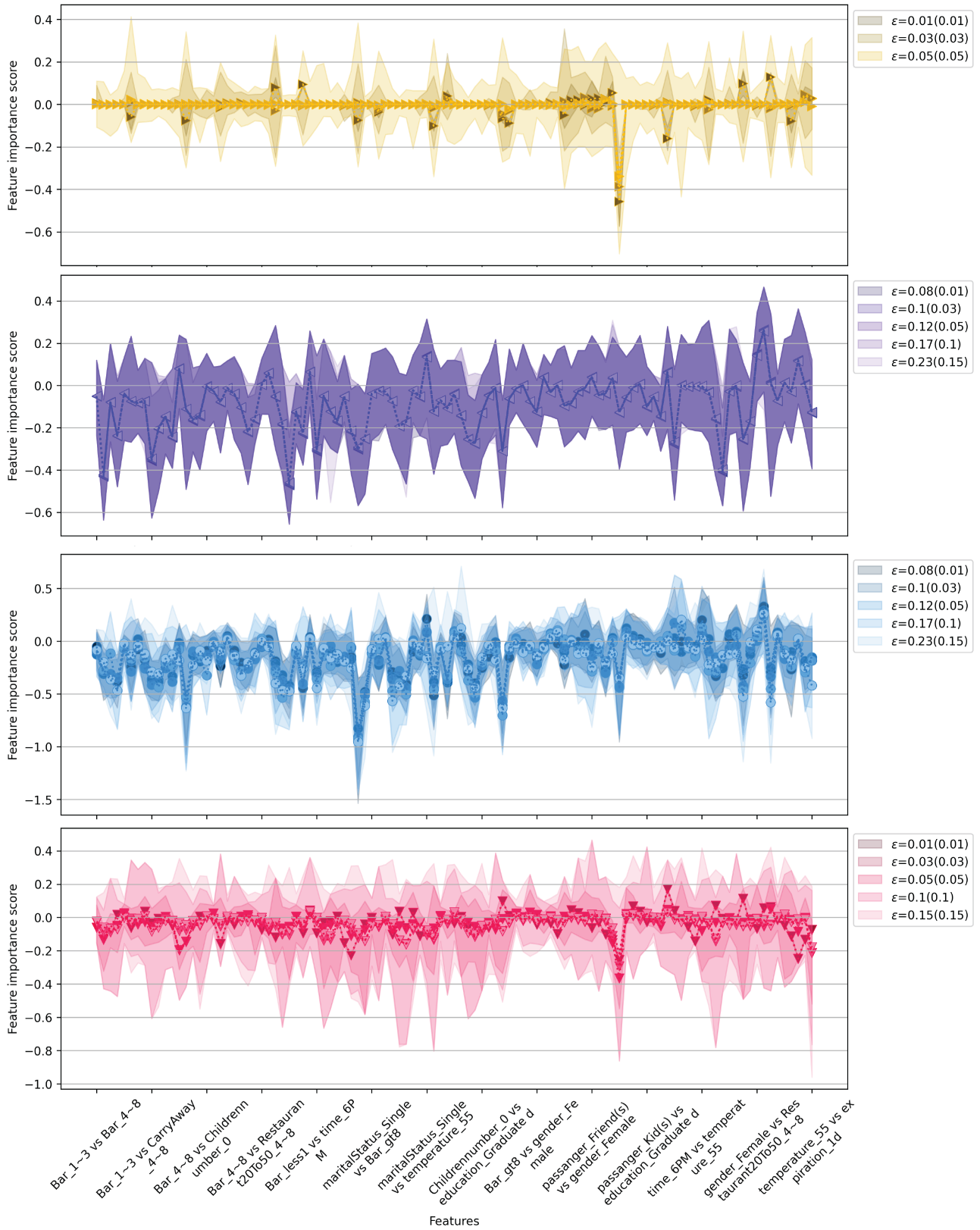


Figure 9: SSummary of detailed second-order feature attributions across different methods and epsilons on the dataset BAR, where the x-axis represents features and the y-axis displays feature importance scores. The legend includes epsilons relative to the reference model (in brackets) and to the optimal model identified in practice.

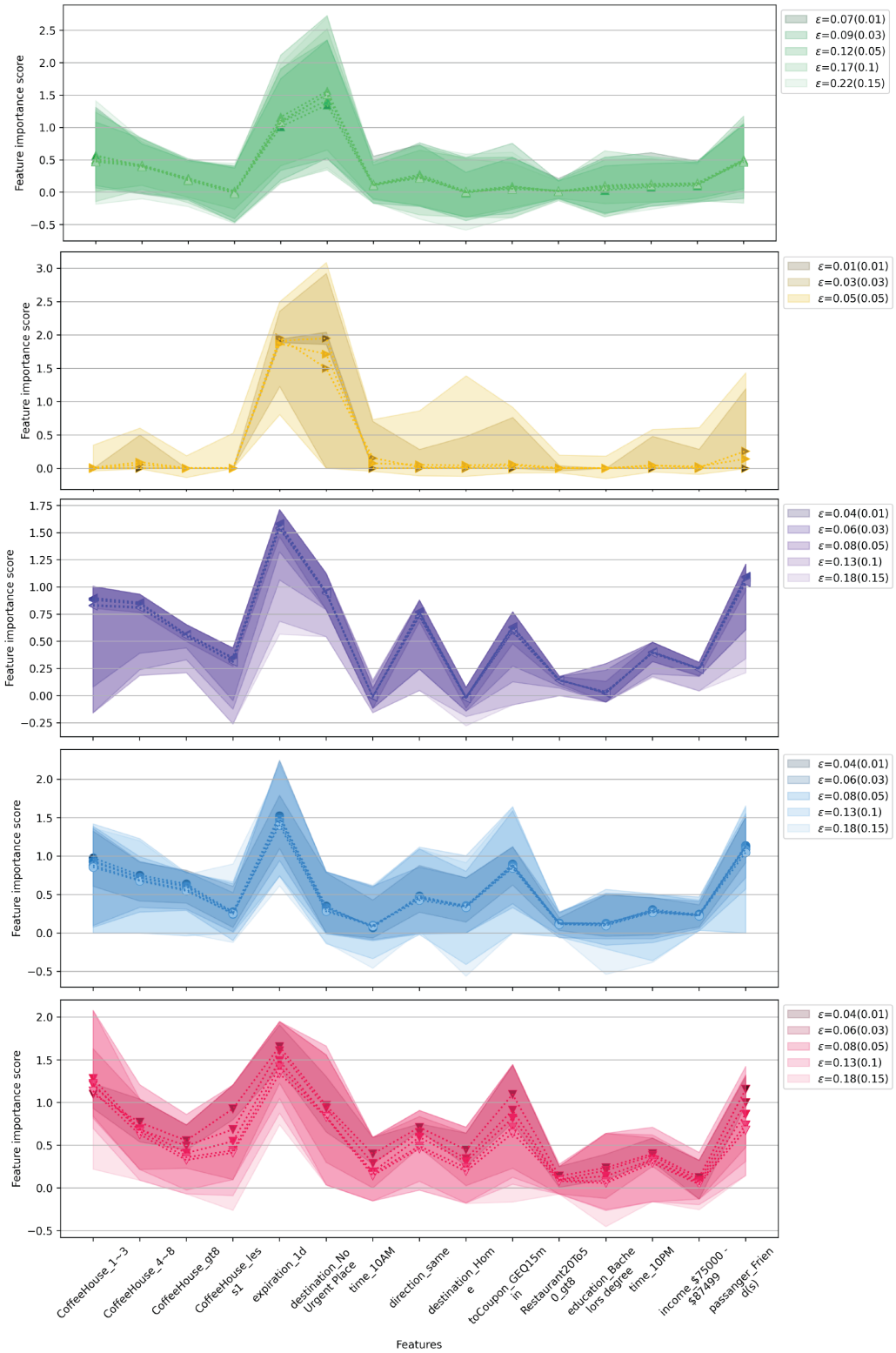


Figure 10: Summary of detailed first-order feature attributions across different methods and epsilons on the dataset coffee house, where the x-axis represents features and the y-axis displays feature importance scores. The legend includes epsilons relative to the reference model (in brackets) and to the optimal model identified in practice.

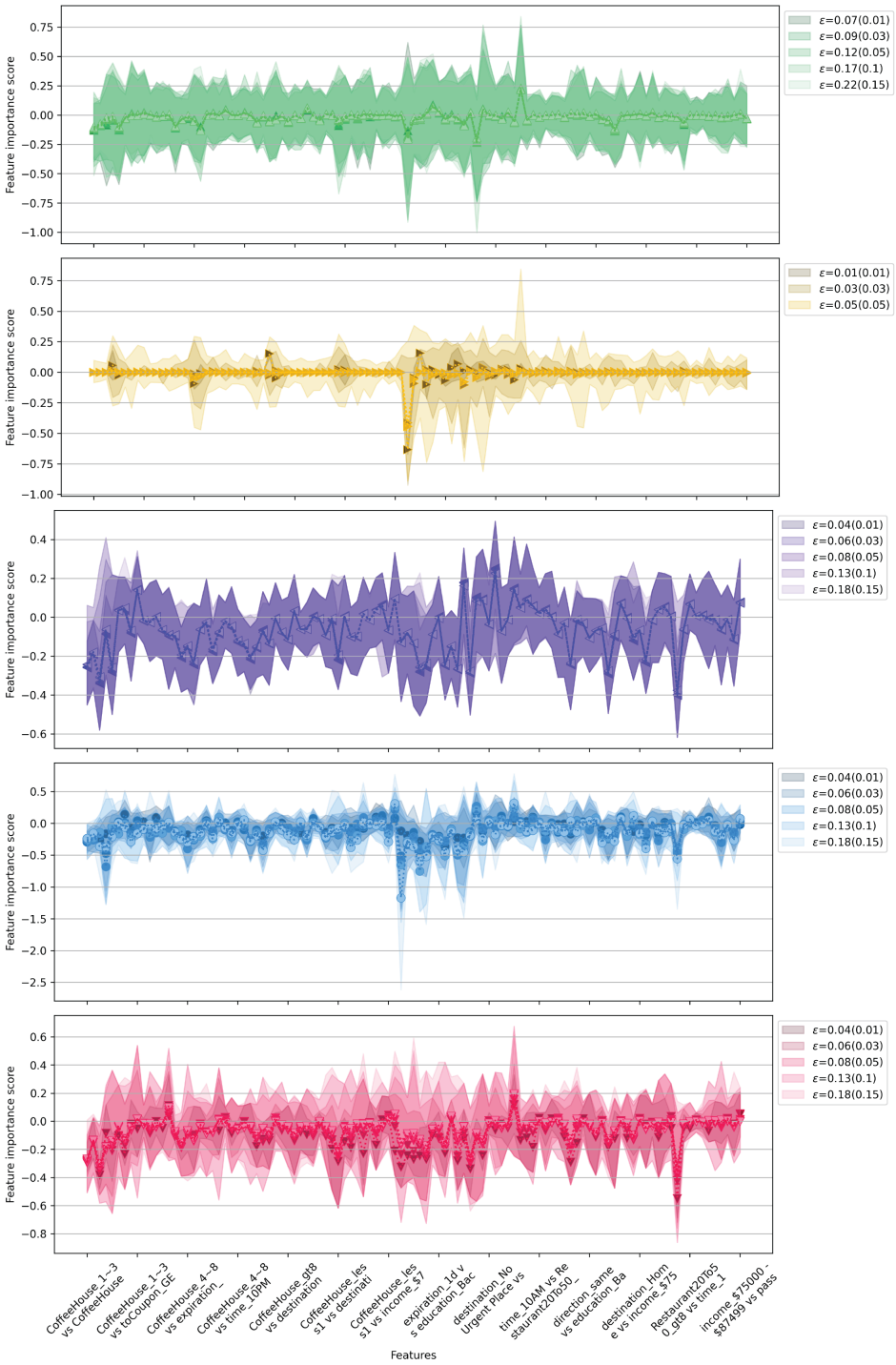


Figure 11: Summary of detailed second-order feature attributions across different methods and epsilons on the dataset coffee house, where the x-axis represents features and the y-axis displays feature importance scores. The legend includes epsilons relative to the reference model (in brackets) and to the optimal model identified in practice.

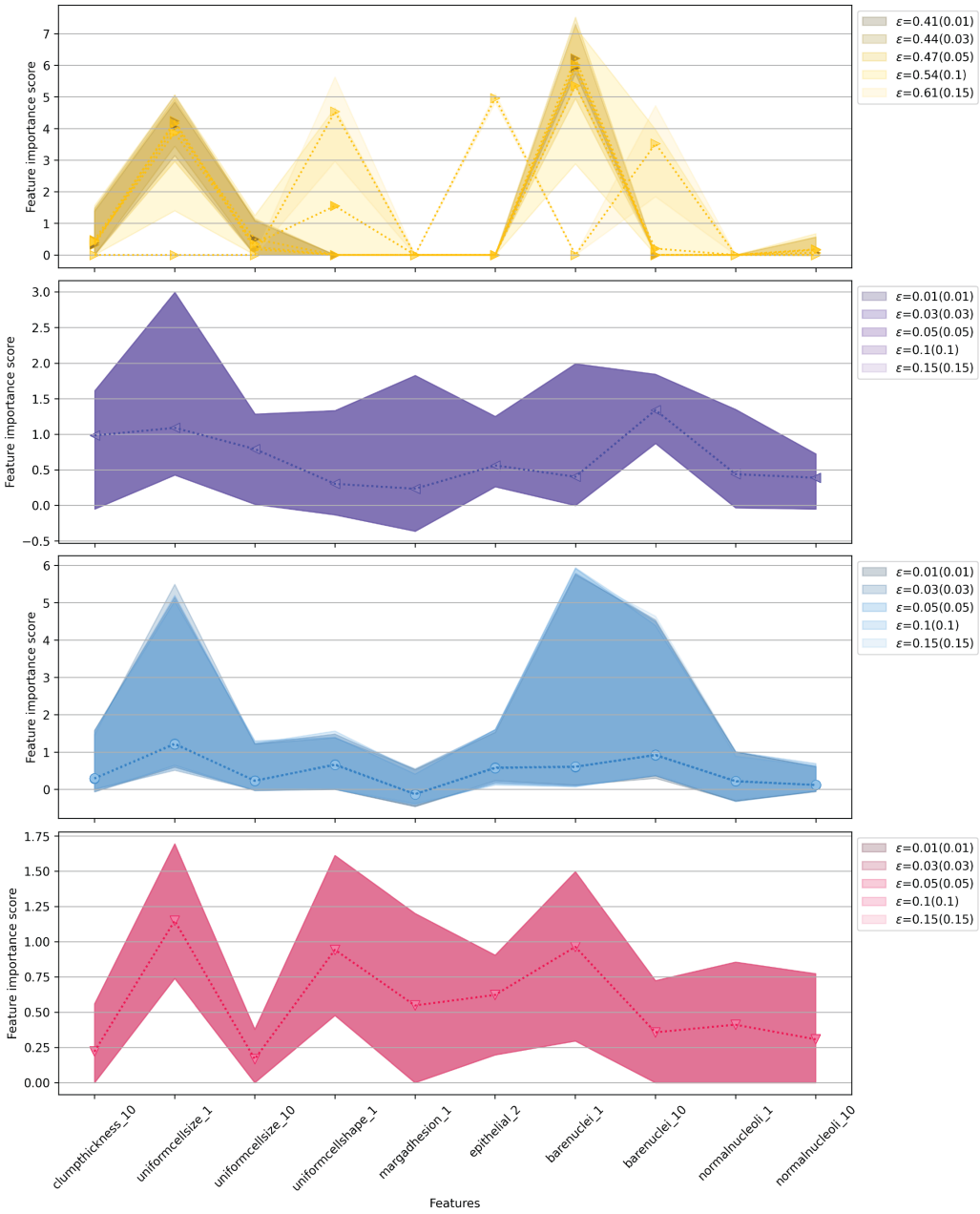


Figure 12: Summary of detailed first-order feature attributions across different methods and epsilons on the dataset breast cancer, where the x-axis represents features and the y-axis displays feature importance scores. The legend includes epsilons relative to the reference model (in brackets) and to the optimal model identified in practice.

Table 3: Experimental configurations for Rashomon set sampling methods.

Method	Model Type	Configuration Details
AWP	MLP	Model Architecture: (n_feature, n_class, 200, 5); Seed = 0; Loss = CrossEntropyLoss Optimizer = Adam, lr = 1e-3
Random Init	MLP	Hidden layers: [100, 100]; Seed = [200, $i\times$200]; Max iterations: [100, 500]
TreeFARMS	Sparse Decision Tree	Regularization = 0.01; Other parameters follow default configuration
VIC	Logistic Classifier	Points per round = 1000, epsScale = 1.5 PCA components = 5, Init. sample size = 10; maxiter = 5000

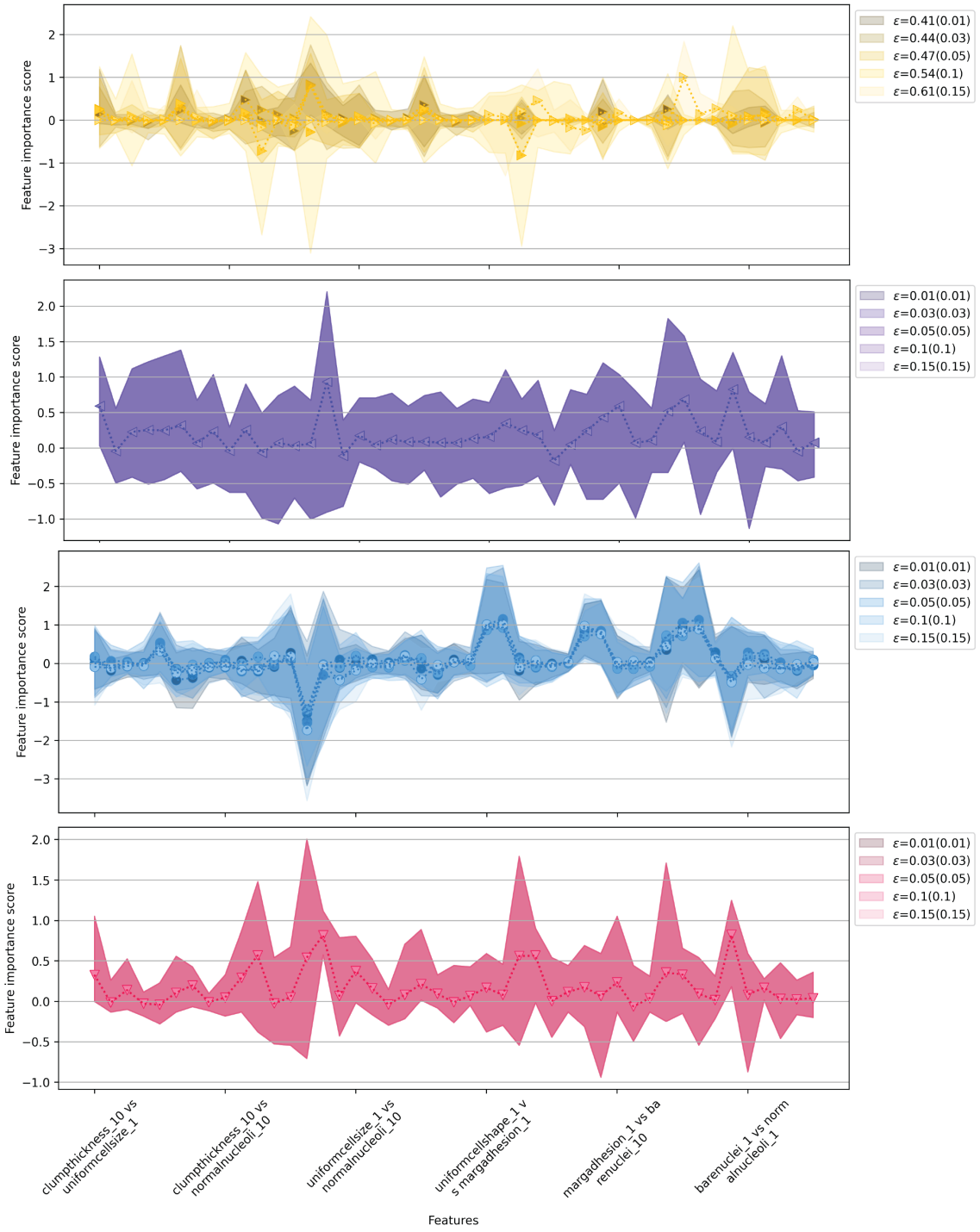
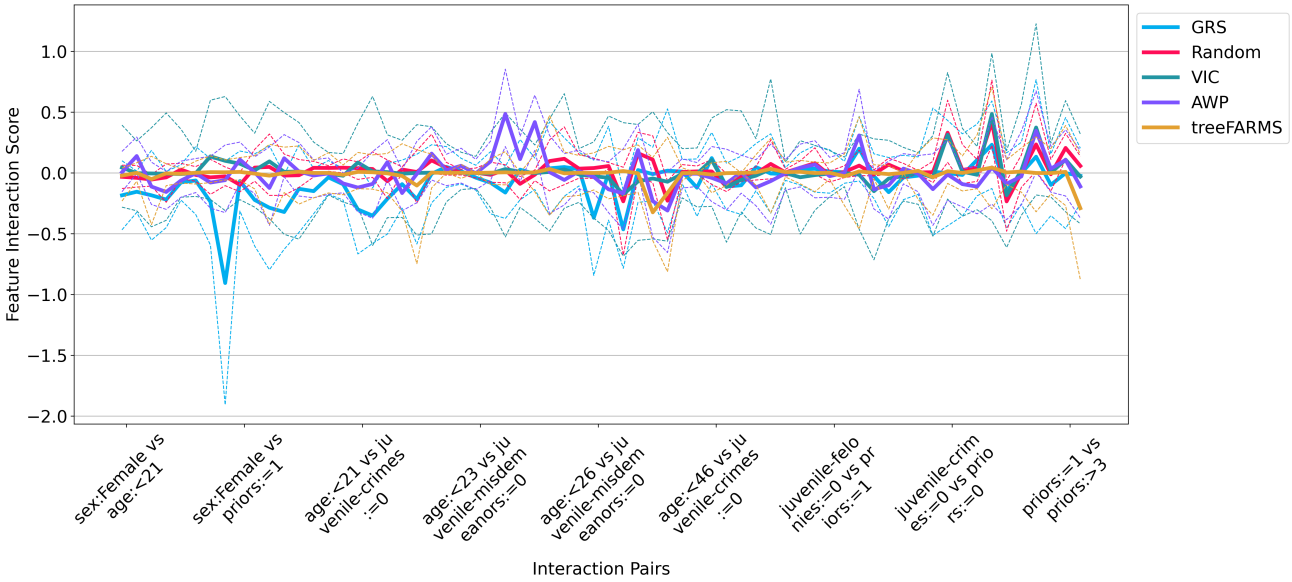
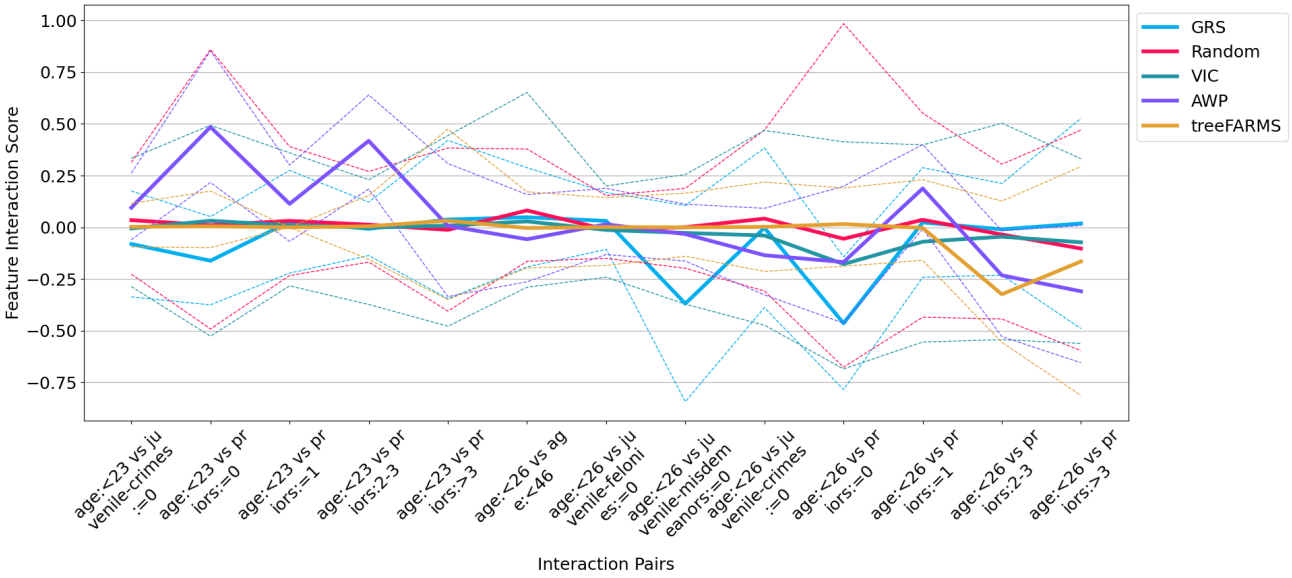


Figure 13: Summary of detailed second-order feature attributions across different methods and epsilons on the dataset breast cancer, where the x-axis represents features and the y-axis displays feature importance scores. The legend includes epsilons relative to the reference model (in brackets) and to the optimal model identified in practice.



(a) Second-order attributions for selected feature pairs at a fixed ϵ.



(b) FER comparison across methods on key second-order interactions.

Figure 14: Detailed second-order feature attribution analysis on the COMPAS dataset. Subfigure (a) shows attribution scores for selected feature pairs at a fixed tolerance level, while (b) compares FER trends across methods.