Fairness and Sparsity within Rashomon Sets: Enumeration-Free Exploration and Characterization

Lucas Langlade[1], Julien Ferry[2], Gabriel Laberge[3], Thibaut Vidal[2]
[1]Ecole nationale des ponts et chaussées, Paris, France
[2]CIRRELT & SCALE-AI Chair in Data-Driven Supply Chains, Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Canada
[3]Department of Computer Engineering and Software Engineering, Polytechnique Montréal, Canada
julien.ferry@polymtl.ca,thibaut.vidal@polymtl.ca

Abstract

We introduce an enumeration-free method based on mathematical programming to precisely characterize various properties such as fairness or sparsity within the set of "good models", known as *Rashomon set*. This approach is generically applicable to any hypothesis class, provided that a mathematical formulation of the model learning task exists. It offers a structured framework to define the notion of business necessity and evaluate how fairness can be improved or degraded towards a specific protected group, while remaining within the Rashomon set and maintaining any desired sparsity level.

We apply our approach to two hypothesis classes: scoring systems and decision diagrams, leveraging recent mathematical programming formulations for training such models. As seen in our experiments, the method comprehensively and certifiably quantifies tradeoffs between predictive performance, sparsity, and fairness. We observe that a wide range of fairness values are attainable, ranging from highly favorable to significantly unfavorable for a protected group, while staying within less than 1% of the best possible training accuracy for the hypothesis class. Additionally, we observe that sparsity constraints limit these tradeoffs and may disproportionately harm specific subgroups. As we evidenced, thoroughly characterizing the tensions between these key aspects is critical for an informed and accountable selection of models.

Code — github.com/vidalt/Rashomon-Explorer
Paper with appendices — arxiv.org/abs/2502.05286

1 Introduction

The increasing reliance on machine learning models for high-stakes decision-support tasks, such as predictive justice (Angwin et al. 2016), hiring (Langenkamp, Costa, and Cheung 2020), and medicine (Aziz et al. 2021) raises important ethical questions and is subject to legal requirements. For instance, article 13 of the recent EU AI Act mandates transparency for AI-based systems classified as "high-risk", a category that encompasses a broad spectrum of applications. This legal and ethical context highlights the importance of developing predictive models that are inherently *interpretable* and *sparse*. Fairness is another critical consideration, reinforced by legal frameworks such as the "80 percent rule" for statistical parity (Feldman et al. 2015) established by the US Equal Employment Opportunity Commission (EEOC. March 2, 1979) in the context of hiring.

When training a machine learning model, the primary objective is typically to maximize its utility. However, multiple models with equivalent performance can produce substantially different predictions, a phenomenon known as predictive multiplicity. This observation gave rise to the concept of ϵ-Rashomon sets (Breiman 2001; Fisher, Rudin, and Dominici 2019), which encompass all models within a given hypothesis class whose utility deviates by no more than ϵ from the optimal value. Due to predictive multiplicity, models within an ϵ-Rashomon set can exhibit markedly different —sparsity or fairness— properties.

Different methods have been proposed to explore Rashomon sets for specific hypothesis classes. On one hand, enumeration-based methods (Xin et al. 2022; Mata, Kanamori, and Arimura 2022; Ciaperoni, Xiao, and Gionis 2024) sample all models within the Rashomon set of interest. This can be computationally prohibitive since the Rashomon sets might contain an untractable number of models. On the other hand, enumeration-free approaches (Coker, Rudin, and King 2021; Watson-Daniels, Parkes, and Ustun 2023; Fisher, Rudin, and Dominici 2019; Zhong et al. 2024; Coston, Rambachan, and Chouldechova 2021) can characterize properties across all models in the Rashomon set without explicitly enumerating them. Despite their efficiency, existing enumeration-free approaches tend to underestimate the size of Rashomon sets and their corresponding fairness-utility tradeoffs. This limitation arises from their reliance on convex upper bounds (e.g., logistic loss, hinge loss) as approximations of the actual model error ($0/1$ loss). Moreover, they do not explore the effects of sparsity requirements, which further influence these tradeoffs.

To overcome these limitations, we introduce an enumeration-free method based on mathematical programming to precisely explore various properties within Rashomon sets, such as fairness or sparsity. Our approach provides a structured, quantitative framework for evaluating the concept of "business necessity" (Grover 1995), a legal argument often used by companies with unbalanced employment outcomes among protected groups. A key aspect of proving "business necessity" is demonstrating that no alternative employment policy could achieve the

same objectives with less discriminatory impact. Within our framework, this translates to asserting that all high-performing models within a given hypothesis class exhibit a disparate impact toward a specific group. Equivalently, the search for less discriminatory alternative models is becoming a legal requirement (Black et al. 2024), and our framework automates this task. More precisely, it can be used to precisely and provably bound the achievable fairness values within given performance and sparsity levels. Our main contributions are:

- We propose an enumeration-free framework for exploring the properties of models within the Rashomon set, focusing on key aspects such as fairness and sparsity. This framework is broadly applicable to any hypothesis class where the learning process can be formulated as a mathematical program.
- To illustrate the versatility of our framework, we apply it to two different hypothesis classes: scoring systems and decision diagrams. The associated source code is openly accessible on our repository, under a MIT license.
- We conduct extensive experiments validating the effectiveness of our approach in certifiably quantifying the tradeoffs (and tensions) between predictive performance, fairness, and sparsity for a given hypothesis class. Our results precisely characterize the range of fairness values achievable under specified sparsity and performance constraints. Additionally, we observe that sparsity does not come for free —imposing stringent sparsity requirements significantly limits the achievable tradeoffs between fairness and performance.

2 Technical Background

Supervised Machine Learning. Let $S := (\boldsymbol{x}_i, y_i)_{i=1}^N$ be a dataset in which each example $i \in \{1..N\}$ is characterized by a feature vector $\boldsymbol{x}_i \in \mathbb{R}^M$ with M attributes and a binary label $y_i \in \{\text{-}1, 1\}$. The objective of a supervised learning algorithm is to produce a predictive model $h : \mathbb{R}^M \to \{\text{-}1, 1\}$ from a given hypothesis space $\mathcal{H}$ that minimizes a loss function $\ell : \{\text{-}1, 1\} \times \{\text{-}1, 1\} \to \mathbb{R}_+$ encoding the error between a predicted and actual outcome. In practice, the empirical loss $\widehat{\mathcal{L}}_S(h) := \frac{1}{N} \sum_{i=1}^N \ell(h(\boldsymbol{x}_i), y_i)$ is minimized to get a good predictor h_S:

$$h_S \in \operatorname*{arg\,min}_{h \in \mathcal{H}} \quad \widehat{\mathcal{L}}_S(h) + \text{Regularization}(h). \tag{1}$$

The regularization term steers an *a priori* preference toward certain hypotheses in $\mathcal{H}$. For this study, we adopt the commonly used $0/1$ loss, defined as $\ell_{0/1}(\hat{y}, y) := \mathbb{1}[\hat{y} = y]$, with $\hat{y} = h(\boldsymbol{x})$.

Fairness. Undesirable biases —specifically harming some individuals or demographic groups— can be embedded in the dataset S, introduced or amplified by the learning algorithm, or arise at any other step of the machine learning pipeline (Mehrabi et al. 2021). Because learning such spurious correlations and using them for decision-making raises ethical questions and is often legally prohibited, different notions of fairness have been proposed (Verma and Rubin 2018). In particular, *statistical fairness* metrics are widely adopted due to their quantifiability and their ability to align with legal standards, such as the "80 percent rule" for statistical parity (Feldman et al. 2015), as outlined by the US Equal Employment Opportunity Commission (EEOC. March 2, 1979) in the context of hiring practices.

Statistical fairness metrics assess disparities in specific statistical measures between different *protected groups*, defined by the values of *sensitive features* (e.g., race, gender). The goal of these metrics is to ensure that such features do not influence individual outcomes. Typically, these measures are derived from the confusion matrix of the predictor h. Let $G_1 \subset S$ and $G_2 \subset S$ represent two protected groups differentiated by a given sensitive feature. For example, in a hiring context, G_1 might represent the set of male applicants, while G_2 represents the set of female applicants. We consider two widely used fairness metrics:

• *Statistical parity* (Dwork et al. 2012) quantifies the difference in positive prediction rates (e.g., acceptance rates for job applicants) between the two protected groups:

$$d_{SP}(h, S) := \frac{\sum_{i \in G_1} \mathbb{1}[h(\boldsymbol{x}_i) = 1]}{|G_1|} - \frac{\sum_{i \in G_2} \mathbb{1}[h(\boldsymbol{x}_i) = 1]}{|G_2|} \tag{2}$$

• *Equal opportunity* (Hardt, Price, and Srebro 2016) measures the difference in true positive rates (e.g., acceptance rates for genuinely qualified applicants) between the two protected groups:

$$d_{EO}(h, S) := \frac{\sum_{i \in G_1^+} \mathbb{1}[h(\boldsymbol{x}_i) = 1]}{|G_1^+|} - \frac{\sum_{i \in G_2^+} \mathbb{1}[h(\boldsymbol{x}_i) = 1]}{|G_2^+|} \tag{3}$$

with $G_1^+ = \{i \in G_1 | y_i = 1\}$ and $G_2^+ = \{i \in G_2 | y_i = 1\}$. For both metrics, values closer to 0 indicate better fairness in the model. Positive values suggest a bias favoring G_1, while negative values indicate a bias toward G_2.

Interpretability. It can be defined as "the ability to explain or to present something in understandable terms to a human" (Doshi-Velez and Kim 2017). It is a critical property for ensuring the trustworthiness of machine learning systems and is often a legal requirement in real-world applications. One possible approach to achieving interpretability is through *post-hoc* explanations (Guidotti et al. 2018) of black-box models, which aim to clarify either individual decisions or the model's overall behavior. However, such methods can be unreliable in certain contexts and are vulnerable to manipulation (Aïvodji et al. 2019; Slack et al. 2020). An alternative is to develop inherently *interpretable* models, such as decision trees or rule lists, which do not share these weaknesses (Rudin 2019). While interpretability lacks a universal definition, *sparsity* (such as the number of nodes in a decision tree) is often used as a proxy (Rudin et al. 2022). Enforcing sparsity constraints effectively restricts the hypothesis space to a more interpretable subset, $\mathcal{H}_I \subset \mathcal{H}$ (Dziugaite, Ben-David, and Roy 2020).

Mathematical Programming. Mathematical programming involves defining a set of *decision variables*, each constrained to a specific domain, and specifying *constraints* that

describe relationships between these variables. A general-purpose solver is used to find *feasible* assignments of the decision variables that satisfy the given constraints. When an *objective function* is provided, the solver seeks a feasible solution that either maximizes or minimizes the function. The types of domains and constraints that can be expressed depend on the chosen paradigm. For example, *Mixed-Integer Linear Programming* (MILP) solvers can accommodate both continuous and discrete variables but are restricted to linear constraints and objective functions.

3 Related Works: Exploring Rashomon Sets

Beyond predictive performance, other model properties, such as fairness and sparsity, are often desirable. Since these properties are typically not aligned with maximizing predictive performance, it is necessary to tolerate a small drop in performance, quantified as ϵ, to search for alternative models h_{alt} satisfying $\widehat{\mathcal{L}}_S(h_{\text{alt}}) \leq \widehat{\mathcal{L}}_S(h_S) + \epsilon$. The set of such alternative models is referred to as the *Rashomon set* (Breiman 2001; Fisher, Rudin, and Dominici 2019), defined as:

$$\mathcal{R}(\mathcal{H}, \epsilon, S) := \{h \in \mathcal{H} : \widehat{\mathcal{L}}_S(h) \leq \widehat{\mathcal{L}}_S(h_S) + \epsilon\}. \quad (4)$$

Rashomon sets have been studied in the context of *predictive multiplicity*, demonstrating that different models can have conflicting predictions on a substantial subset of data (Marx, Calmon, and Ustun 2020; Hsu and Calmon 2022; Watson-Daniels, Parkes, and Ustun 2023). A key result from Marx, Calmon, and Ustun (2020) establishes that for any alternative $h_{\text{alt}} \in \mathcal{R}(\mathcal{H}, \epsilon, S)$, the following tight bound holds: $\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[h_{\text{alt}}(\boldsymbol{x}_i) \neq h_S(\boldsymbol{x}_i)] \leq 2\widehat{\mathcal{L}}_S(h_S) + \epsilon$. This implies that even with a small ϵ, models within the Rashomon set can differ significantly in their predictions whenever the empirical loss $\widehat{\mathcal{L}}_S(h_S)$ is non-zero. For instance, if h_S has an empirical loss of 10%, alternative models in the Rashomon set could disagree with h_S on up to 20% of the dataset. Similarly, for a predictor h_S with a 25% empirical loss, disagreements with alternative models could extend to 50% of the dataset. These substantial differences emphasize that models within the Rashomon set, despite achieving nearly equivalent predictive performance, can vary markedly in their predictions. This variability has important implications for fairness, as fairness metrics such as statistical parity (Equation (2)) and equal opportunity (Equation (3)) are aggregates of predictions across demographic subgroups. Consequently, the Rashomon set may contain alternative models with more desirable fairness or sparsity properties. However, identifying such models efficiently remains a significant challenge. Existing methods for exploring the Rashomon set can be categorized into two main approaches: *enumeration-based* and *enumeration-free*.

Enumeration-based methods sample all models within the Rashomon set (or a substantial subset thereof). Existing approaches have applied enumeration to the Rashomon sets of rule lists (Mata, Kanamori, and Arimura 2022), rule sets (Ciaperoni, Xiao, and Gionis 2024), and decision trees (Xin et al. 2022) using branch-and-bound techniques. These methods explore the combinatorial hypothesis space $\mathcal{H}$ while leveraging error lower bounds to prune the search space efficiently. These past works have shown that competing models exhibit different fairness properties. However, a key limitation of these methods is their reliance on enumerating (and storing) a large number of models. While methods were proposed to sample only a subset of models from the Rashomon set – e.g., through bootstrapping (Cooper et al. 2024) or shuffling (Ganesh et al. 2023) – they do not aim at exploring the whole Rashomon set and cannot provide bounds on the extrema of a functional.

Alternatively, enumeration-free methods focus on the identification of models within the Rashomon set that achieve the extreme values of a specific functional $\phi : \mathcal{H} \to \mathbb{R}$. This approach allows targeted exploration of the Rashomon set by optimizing for particular properties without exhaustive enumeration. Previous work has investigated the min-max range of the functional $\phi(h) = h(\boldsymbol{x})$ for linear models with the hinge-loss (Coker, Rudin, and King 2021) and the logistic loss (Watson-Daniels, Parkes, and Ustun 2023). Other studies have explored the extreme values of feature importance scores for linear models under the squared or hinge loss (Fisher, Rudin, and Dominici 2019), or Generalized Additive Models (GAMs) under the logistic loss (Zhong et al. 2024). Finally, the min-max range of the functional ϕ underlying fairness metrics (cf. Equations (2) & (3)) has been characterized for linear models with logistic loss (Coston, Rambachan, and Chouldechova 2021). Table 1 summarizes these previous works.

Considering the "true" $0/1$ loss is desirable, since it exactly quantifies the utility of the trained model (the proportion of individuals whose outcome was incorrectly predicted). However, it makes the problem more difficult to solve since it is not continuous nor convex. Using convex upper bounds such as logistic or hinge losses makes the problem more tractable, but leads to underestimating the Rashomon set, hence arbitrarily limiting the tradeoffs between other desiderata (e.g., fairness and sparsity). In turn, discrete optimization tools (such as mathematical programming) are appropriate to directly handle the 0-1 loss and exactly characterize the Rashomon set. They are also particularly well-suited to learn models that are inherently interpretable (which is the focus of this paper), such as rule-based or tree-based ones – which intrinsically have a combinatorial structure. Based on these observations, our study:

- explores the range of unfairness values within the Rashomon sets using the "true" $0/1$ loss, whereas previous works relied on convex upper bounds;
- characterizes the effect of sparsity constraints on the range of possible disparities within the Rashomon set;
- provides a framework applicable to many hypothesis classes $\mathcal{H}$ provided their learning process can be formulated as a mathematical optimization problem.

We now discuss two closely related works that also explore the fairness properties of model classes, though from different perspectives. Considering models within all possible input-output mappings for a finite dataset, Dai et al. (2025) define the notion of *largest possible Rashomon Set*, and propose methods to efficiently bound the achievable fairness values within them. While these bounds hold for

Source	Hypothesize class $\mathcal{H}$	Loss ℓ	Functional ϕ
Coker, Rudin, and King (2021)	Linear	hinge loss	prediction
Watson-Daniels, Parkes, and Ustun (2023)	Linear	logistic loss	prediction
Fisher, Rudin, and Dominici (2019)	Linear/Kernels	hinge loss	feature importance
Zhong et al. (2024)	Additive (GAM)	logistic loss	feature importance
Coston, Rambachan, and Chouldechova (2021)	Linear	logistic loss	fairness metrics
Ours	**Linear/Decision-Diagrams**	**0-1 loss**	**fairness metrics**

Table 1: Summary of enumeration-free Rashomon set exploration methods for binary classification tasks.

any hypothesis class, there might not exist any model within a given hypothesis class implementing the chosen input-output mapping, hence they can be arbitrarily loose. On the contrary, our work aims at providing tight bounds for a given hypothesis class, along with their associated models for a chosen sparsity level. Simson, Pfisterer, and Kern (2024) rather characterize the set of fairness values reachable depending on the design of the ML pipeline (e.g., preprocessing and evaluation choices), while we precisely bound fairness for all possible models within the Rashomon set for a fixed dataset and hypothesis class.

4 Exploring Rashomon Sets Through Mathematical Programming

We first introduce our generic framework for exploring the Rashomon set of a given hypothesis class whose learning is formulated as a mathematical program. We then instantiate it for two widely used classes of interpretable models, namely scoring systems and decision diagrams.

4.1 Generic Framework

As stated in Equation (1), the goal of a machine learning algorithm is to explore the hypothesis space $\mathcal{H}$ to identify a model h_S that minimizes (on a training dataset S) a given objective function, which consists of its empirical loss $\widehat{\mathcal{L}}_S(h_S)$ and, optionally, a regularization term. We focus on the common scenario where the regularization term measures the model's sparsity, with the tradeoff between sparsity and predictive performance governed by a coefficient C. The general mathematical formulation of this learning process is:

$$\min_{h \in \mathcal{H}} \ \widehat{\mathcal{L}}_S(h) + C \cdot \text{Sparsity}(h). \tag{5}$$

The model's structure and parameters are encoded through *decision variables*, while its internal predictions and adherence to the hypothesis space are enforced through a set of *constraints*.

The objective of our proposed framework is to provably determine the maximum and minimum values of a given fairness metric (along with the corresponding models) subject to a desired sparsity constraint while remaining within an ϵ-Rashomon set of the hypothesis space $\mathcal{H}$. To achieve this, we first solve Problem (5) with $C = 0$ to obtain the optimal empirical loss value $\widehat{\mathcal{L}}_S(h_S)$, which by definition constitutes the reference value for the Rashomon set computation. We then formulate and solve the following problem:

$$\begin{aligned} \min_{h} \ & d_{SP}(h, S) \\ \text{s.t.} \ & h \in \mathcal{H} \\ & \text{Sparsity}(h) \leq \alpha \\ & \widehat{\mathcal{L}}_S(h) \leq \widehat{\mathcal{L}}_S(h_S) + \epsilon \end{aligned} \tag{6}$$

Here, α represents the desired sparsity, which sets an upper bound on the model's size, and $d_{SP}(h, S)$ is the statistical parity metric (Equation (2)), although any other fairness measure can replace it. Observe that minimizing $d_{SP}(h, S)$ amounts to maximally favor the protected group G_2 over G_1 (in terms of positive prediction rate). By reversing the sign of the objective, the full range of fairness values within the ϵ-Rashomon set can thereby be characterized. Additionally, the impact of sparsity constraints on the accuracy-fairness tradeoff can be further explored by varying α.

4.2 Instantiation for Scoring Systems

Scoring systems are sparse linear classification models with integer coefficients, widely used in fields like medicine and criminal justice due to their interpretability (Rudin et al. 2022). To make a prediction with such a model on a given example $\boldsymbol{x}$, one multiplies each feature's value x_j by its corresponding coefficient λ_j selected within an acceptable range of values $\Omega_j \subset \mathbb{N}$, sums the results, and compares the total to a fixed threshold. The hypothesis space of scoring systems is then:

$$\mathcal{H} := \left\{ \boldsymbol{x} \mapsto \ \text{sign}(\boldsymbol{x}^T \boldsymbol{\lambda}) \mid \lambda_j \in \Omega_j, j = 1..M \right\}. \tag{7}$$

Table 2 presents an example scoring system trained on the Default of Credit Card Clients dataset (Yeh and hui Lien 2009). The classification task involves predicting whether a person will default on payment based on demographic information and payment histories. In addition to the coefficients associated with the M features (only the non-zero ones are shown), a **threshold** term is also included. This threshold is usually handled by concatenating an additional feature with a value of 1 to all examples before training or inference—so in (7), M actually refers to the number of features plus one. As visible in the table, the model's interpretability allows for straightforward identification of the features influencing predictions. For instance, features indicating delays in previous payments or high payment amounts are associated with an increased likelihood of predicting a default on the next payment. However, the model also exhibits a bias against

females, as the attribute "SEX_Female" increases the computed score, thereby increasing the probability of predicting a default for females. This aligns with the measured statistical parity value (Equation (2)) of -0.046, whose negativity indicates a higher positive prediction rate for group G_2 (females) over group G_1. In this example, interpretability facilitates the detection of such discriminations.

SLIM *(Supersparse Linear Integer Model)* (Ustun, Tracà, and Rudin 2014) is a MILP formulation designed to learn optimal scoring systems. We use it within our framework to instantiate the learning problem defined in Problem (5). The original SLIM formulation aims at finding the coefficients $\boldsymbol{\lambda}_S$ minimizing the following objective:

$$\min_{\boldsymbol{\lambda}} \sum_{i=1}^{N} \mathbb{1}[y_i \boldsymbol{x}_i^T \boldsymbol{\lambda} \leq 0] + C\|\boldsymbol{\lambda}\|_0 \tag{8}$$

where $\boldsymbol{\lambda}$ is the vector of coefficients within the scoring system, $\boldsymbol{x}_i^T \boldsymbol{\lambda}$ is the scoring system's total score for example i (whose sign determines the output label), and C is a regularization coefficient. Then, $\sum_{i=1}^{N} \mathbb{1}[y_i \boldsymbol{x}_i^T \boldsymbol{\lambda} \leq 0]$ computes the $0/1$ empirical loss of the model, while $\|\boldsymbol{\lambda}\|_0$ is a sparsity regularizer, penalizing the number of non-zero coefficients.

We now present our modified formulation, which instantiates Problem (6) to characterize fairness and sparsity within the Rashomon set of scoring systems. Recall that the optimal loss value $\widehat{\mathcal{L}}_S(\boldsymbol{\lambda}_S)$ is first obtained by solving the original SLIM formulation with $C = 0$ (i.e., ensuring that objective (8) focuses solely on predictive performance).

$$\min_{\boldsymbol{\lambda}} \quad \frac{\sum_{i \in G_1} \hat{y}_i}{|G_1|} - \frac{\sum_{i \in G_2} \hat{y}_i}{|G_2|} \tag{9}$$

$$\text{s.t.} \quad \lambda_j = \sum_{\omega \in \Omega_j} \omega \cdot u_{j\omega} \qquad j \in \{1, .., M\} \tag{10}$$

$$\sum_{\omega \in \Omega_j} u_{j\omega} \leq 1 \qquad j \in \{1, .., M\} \tag{11}$$

$$\sum_{j=1}^{M} \sum_{\omega=1}^{\Omega_j} u_{j\omega} \leq \alpha \qquad \text{(Sparsity)} \tag{12}$$

$$\frac{1}{N} \sum_{i=1}^{N} z_i \leq \widehat{\mathcal{L}}_S(\boldsymbol{\lambda}_S) + \epsilon \qquad \text{(Performance)} \tag{13}$$

$$O'_i z_i \geq \gamma - y_i \boldsymbol{x}_i^T \boldsymbol{\lambda} \qquad i \in \{1, .., N\} \tag{14}$$

$$O_i (1 - z_i) \geq y_i \boldsymbol{x}_i^T \boldsymbol{\lambda} \qquad i \in \{1, .., N\} \tag{15}$$

$$\hat{y}_i = (1 - z_i)\mathbb{1}[y_i = 1] + z_i \mathbb{1}[y_i = -1] \qquad i \in \{1, .., N\} \tag{16}$$

$$\lambda_j \in \Omega_j \qquad j \in \{1, .., M\}$$

$$z_i \in \{0, 1\} \qquad i \in \{1, .., N\}$$

$$\hat{y}_i \in \{0, 1\} \qquad i \in \{1, .., N\}$$

$$u_{j\omega} \in \{0, 1\} \qquad j \in \{1, .., M\},\ \omega \in \Omega_j$$

Each coefficient λ_j associated to feature j within the scoring system must take a value within a user-defined domain Ω_j. Specifically, Constraint (10) ensures that the coefficient λ_j takes value $\omega \in \Omega_j$ if and only if $u_{j\omega} = 1$. Constraint (11) guarantees that at most one value $\omega \in \Omega_j$ is set to 1. Note that $\lambda_j = 0$ if all the variables $u_{j\omega}$ equal 0.

Feature	Coefficient
EDUCATION_University	2
PAY_0_Pay_delay≥1	5
PAY_2_Pay_delay≥1	5
PAY_6_Pay_delay≥1	2
PAY_AMT6_high	2
SEX_Female	2
Threshold	**-10**
Predict $+1$ ***if total is*** $> 0, -1$ ***otherwise***	

Table 2: Example scoring system trained on the Default of Credit Card Clients dataset, belonging to the 20%-Rashomon set, exhibiting 0.842 training accuracy and 0.80 test accuracy, as well as -0.046 training statistical parity.

Objective (9) represents the statistical parity metric introduced in Equation (2). By minimizing it, we aim to find the scoring system that maximally favors the protected group G_2 over G_1. Reversing the sign of this difference allows us to optimize the fairness value in the opposite direction. Constraint (13) limits the hypothesis space to the ϵ-Rashomon set, leveraging the previously computed optimal loss $\widehat{\mathcal{L}}_S(\boldsymbol{\lambda}_S)$ (as defined in Equation (8)). Constraint (12) restricts the number of non-zero coefficients in $\boldsymbol{\lambda}$ to at most α, thereby enforcing sparsity.

The remaining constraints handle the intermediate computations of the scoring system's predictive performance and predictions. Specifically, the loss variables z indicate whether each example i is incorrectly classified: $z_i = \mathbb{1}[y_i \boldsymbol{x}_i^T \boldsymbol{\lambda} \leq 0]$. These variables are determined by Constraints (14–15), which compare the sign of each example i's predictions $(\boldsymbol{x}_i^T \boldsymbol{\lambda})$ with its true label y_i. Note that $O'_{i \in \{1..N\}}$ and $O_{i \in \{1..N\}}$ are pre-computed constants large enough to enforce the constraints, and γ is a small constant representing a margin, ensuring that $(y_i \boldsymbol{x}_i^T \boldsymbol{\lambda})$ for all examples i is lower-bounded.

In the original SLIM formulation, because the sum of the loss variables was minimized in the objective, Constraint (14) alone was sufficient to set z_i to 1 if and only if example i is misclassified. As this is no longer the case here, we must additionally include Constraint (15) to force z_i to 0 in case of correct classification.

Finally, the predictions $\hat{y}$ are computed leveraging the loss variables and the actual labels (given as input constants to the model) through Constraint (16). For each example i, we then have: $\hat{y}_i = \mathbb{1}[\boldsymbol{x}_i^T \boldsymbol{\lambda} > 0]$.

This formulation precisely determines the extent to which each protected group can be favored over the other, given a specified sparsity level α (maximum number of non-zero coefficients) and predictive performance threshold (defined by the ϵ-Rashomon set). By varying α and ϵ, one can explore the tradeoffs between these different desiderata.

4.3 Instantiation for Decision Diagrams

Decision diagrams are popular interpretable models exhibiting a top-down hierarchical structure similar to trees. Yet,

unlike decision trees, their branches can be merged. This fundamental property avoids the replication and fragmentation problems of decision trees (Oliver 1993; Kohavi 1994; Florio et al. 2023), hence enhancing interpretability. Formally, a decision diagram is a rooted directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{A})$, where each internal node $v \in \mathcal{V}^I$ represents a splitting hyperplane and each terminal node $v \in \mathcal{V}^C$ is uniquely associated with a prediction c_v. This hypothesis class generalizes rule-lists ($\mathcal{H}_{\text{rule-list}} \subset \mathcal{H}_{\text{diagrams}}$), so investigating its ϵ-Rashomon set is an enumeration-free alternative to the approach of Mata, Kanamori, and Arimura (2022). As with `SLIM`, the objective $d_{SP}(h, S)$, and Sparsity/Performance constraints of Problem (6) are easily expressed as linear functions of the decision variables, allowing for a MILP formulation.

We build upon the original MILP formulation by Florio et al. (2023) for learning optimal decision diagrams for classification. In essence, given a user-specified maximum structure, the formulation aims to determine which nodes and edges should be utilized within this structure and how their splitting hyperplanes should be defined. Sparsity is then computed as the number α of active (utilized) internal nodes. The objective is as follows:

$$\min_{(d,\boldsymbol{a},b)_{v\in\mathcal{V}^I}} \quad \sum_{i=1}^{N} z_i + C\|\boldsymbol{d}\|_0. \tag{17}$$

Here, for each example i, the loss variable z_i indicates whether it is misclassified, so $\sum_{i=1}^{N} z_i$ computes the $0/1$ loss. For each internal node within the predefined structure $v \in \mathcal{V}^I$, the variable $d_v \in \{0, 1\}$ indicates whether it is used in the final structure. The term $\|\boldsymbol{d}\|_0$ quantifies the sparsity of the resulting decision diagram by counting the number of nodes used in the trained structure. Finally, for each internal node v where $d_v = 1$, variables $(\boldsymbol{a}_v, b_v)$ define the hyperplane corresponding to the multivariate split performed by this node. This objective effectively instantiates Problem (5).

We hereafter provide our modified formulation, which instantiates Problem (6) to characterize fairness and sparsity within the Rashomon set of decision diagrams. Recall that the optimal loss value $\widehat{\mathcal{L}}_S((d_S, \boldsymbol{a}_S, b_S)_{v\in\mathcal{V}^I})$ is first obtained by solving the original MILP formulation with $C = 0$ (i.e., ensuring that Objective (17) focuses solely on predictive performance). The following objective represents the statistical parity metric introduced in Equation (2):

$$\min_{(d,\boldsymbol{a},b)_{v\in\mathcal{V}^I}} \quad \frac{\sum_{i\in G_1} \hat{y}_i}{|G_1|} - \frac{\sum_{i\in G_2} \hat{y}_i}{|G_2|} \tag{18}$$

By minimizing it, we aim to find the decision diagram that maximally favors the protected group G_2 over G_1 in terms of positive prediction rate. Reversing the sign of this difference allows us to constrain the fairness value in the opposite direction. We hereafter introduce the different constraints that must be satisfied while optimizing Objective (18). Constraint (19) restricts the number of active nodes in $\boldsymbol{d}$ to at most α, thereby enforcing sparsity. Constraint (20) limits the hypothesis space to the ϵ-Rashomon set, utilizing the previously computed optimal loss $\widehat{\mathcal{L}}_S((d_S, \boldsymbol{a}_S, b_S)_{v\in\mathcal{V}^I})$:

$$\sum_{v\in\mathcal{V}^I} d_v \leq \alpha \qquad \text{(Sparsity)} \tag{19}$$

$$\frac{1}{N}\sum_{i=1}^{N} z_i \leq \widehat{\mathcal{L}}_S((d_S, \boldsymbol{a}_S, b_S)_{v\in\mathcal{V}^I}) + \epsilon \qquad \text{(Performance)} \tag{20}$$

The $0/1$ loss associated to each example $i \in \{1, .., N\}$ is then computed as follows:

$$z_i = \sum_{v\in\mathcal{V}^C} \mathbb{1}[y_i \neq c_v] w_{iv} \tag{21}$$

$$\hat{y}_i = (1 - z_i)\mathbb{1}[y_i = 1] + z_i \mathbb{1}[y_i = -1] \tag{22}$$

Constraint (21) sets the loss variable $z_i = 1$ if and only if example i is assigned to a terminal node $v \in \mathcal{V}^C$ (i.e., $w_{iv} = 1$) whose predicted class c_v differs from the example's true label y_i. Constraint (22) then uses the loss variables z to determine the decision diagram's predictions $\hat{y}$. The remaining constraints are unchanged and model the structure of the constructed decision diagram. Below, we briefly discuss the role of each constraint, and we refer to Florio et al. (2023) for a more comprehensive explanation. For each example $i \in \{1, .., N\}$, Constraints (23–25) model the flow of each example through the nodes of the decision diagram:

$$w_{iv}^+ + w_{iv}^- = \begin{cases} 1 \text{ if } v = 0 \\ \sum_{u\in\delta^-(v)} (f_{iuv}^+ + f_{iuv}^-) \end{cases} \qquad v \in \mathcal{V}^I \tag{23}$$

$$w_{iu}^- = \sum_{v\in\delta^+(u)} f_{iuv}^- \qquad u \in \mathcal{V}^I \tag{24}$$

$$w_{iu}^+ = \sum_{v\in\delta^+(u)} f_{iuv}^+ \qquad u \in \mathcal{V}^I \tag{25}$$

$$\sum_{u\in\mathcal{V}_l^I} w_{iu}^- \leq 1 - g_{il} \qquad l \in \{0, .., D-1\} \tag{26}$$

$$\sum_{u\in\mathcal{V}_l^I} w_{iu}^+ \leq g_{il} \qquad l \in \{0, .., D-1\} \tag{27}$$

Specifically, $\delta^-(u)$ (respectively, $\delta^+(u)$) represents the set of possible predecessors (respectively, successors) of node u in the user-provided decision diagram structure. The variable w_{iu}^- (respectively, w_{iu}^+) takes a non-zero value when example i passes through node u on the negative (respectively, positive) side of the separating hyperplane. Additionally, the variable f_{iuv}^- (respectively, f_{iuv}^+) models the flow from the negative (respectively, positive) side of u to other nodes v. Constraints (26–27) ensure flow integrity using the binary variable g_{il}, which determines, for each example i, whether it follows the negative or positive side at each level $l \in \{0, .., D-1\}$, D being the depth of the decision diagram. For each node $u \in \mathcal{V}^I$, Constraints (28–30) specify that it is used in the decision diagram ($d_u = 1$) if and only if it is connected to or from another node:

$$d_u = \sum_{v\in\delta^+(u)} t_{uv}^+ = \sum_{v\in\delta^+(u)} t_{uv}^- \tag{28}$$

$$d_v \leq \sum_{u\in\delta^-(v)} (t_{uv}^+ + t_{uv}^-) \qquad v \in \mathcal{V}^I \setminus \{0\} \tag{29}$$

$$t_{uv}^+ + t_{uv}^- \leq d_v \qquad v \in \delta^+(u) \tag{30}$$

The binary variable t^-_{uv} (respectively, t^+_{uv}) indicates that node $u \in \mathcal{V}^I$ links to node v on the negative (respectively, positive) side. Note that both the root and terminal nodes are excluded from these constraints, as they are always used. Constraints (31–32) connect the linking variables to the examples' flows, for each node $u \in \mathcal{V}^I$:

$$f^+_{iuv} \leq t^+_{uv} \qquad v \in \delta^+(u),\ i \in \{1,..,N\} \quad (31)$$

$$f^-_{iuv} \leq t^-_{uv} \qquad v \in \delta^+(u),\ i \in \{1,..,N\} \quad (32)$$

Constraints (33–34) implement symmetry breaking, as many equivalent topologies could result from the previously defined constraints and variables:

$$t^-_{uv} + \sum_{w \in \delta^+(u), w \leq v} t^+_{uw} \leq 1 \qquad u \in \mathcal{V}^I,\ v \in \delta^+(u) \quad (33)$$

For each level $l \in \{2,..,D-1\}$, a weak in-degree ordering is enforced for each pair of nodes $u, v \in \mathcal{V}^I_l$ such that $u < v$:

$$\sum_{w \in \delta^-(u)} (t^+_{wu} + t^-_{wu}) \geq \sum_{w \in \delta^-(v)} (t^+_{wv} + t^-_{wv}) \quad (34)$$

Finally, for each example $i \in \{1,..,N\}$, Constraints (35–36) ensure consistency between the hyperplane variables and the example's flow, while Constraint (37) determines the terminal node v to which it is assigned by setting w_{iv} based on the previously computed flows:

$$(w^-_{iv} = 1) \implies (\boldsymbol{a}_v^T \boldsymbol{x}_i + \gamma \leq b_v) \qquad v \in \mathcal{V}^I \quad (35)$$

$$(w^+_{iv} = 1) \implies (\boldsymbol{a}_v^T \boldsymbol{x}_i > b_v) \qquad v \in \mathcal{V}^I \quad (36)$$

$$w_{iv} = \sum_{u \in \delta^-(v)} (f^+_{iuv} + f^-_{iuv}) \qquad v \in \mathcal{V}^C \quad (37)$$

5 Experimental Study

Our numerical experiments serve two main objectives. First, we demonstrate the applicability and effectiveness of our framework in characterizing fairness and sparsity within Rashomon sets, through an instantiation for two hypothesis classes. Second, we explore the interplays between the three desiderata, highlighting the main trends.

5.1 Experimental Setup

Datasets. We consider three datasets widely used in the fair and interpretable machine learning literature. First, the UCI Adult Income dataset (Dua and Graff 2017) contains records on $32,561$ individuals from the 1994 U.S. census, described by 36 binary attributes. The classification task is to predict whether an individual earns more than \$50K per year. In our experiments, G_1 represents males and G_2 represents females. Second, the Default of Credit Card Clients dataset (Yeh and hui Lien 2009) includes demographic information and payment histories for $29,986$ individuals in Taiwan, each described by 21 attributes. The task is to predict whether a person will default on payment, with G_1 as males and G_2 as females. Third, the COMPAS dataset (Angwin et al. 2016) contains data on $7,214$ criminal offenders in Broward County, Florida, described by 27 binary attributes. The task is to predict whether an individual will re-offend within two years. Here, G_1 represents African-Americans, and G_2 represents the rest of the population.

Learning Procedure. For each dataset, we randomly subsample training sets S of size $N = 500$, with the remaining examples used as a test set, as this permits a fast and unbiased evaluation based on optimal solutions of the underlying mathematical models. Note that we also report some results using larger training set sizes N in the Appendix A[1] to illustrate the scalability of our method and the consistency of our empirical findings. We generate five different random splits and report both the average values and standard deviations in our experiments. The two fairness metrics considered are statistical parity (Equation (2)) and equal opportunity (Equation (3)). For each random split of each dataset, we determine the optimal loss $\widehat{\mathcal{L}}_S(h_S)$ and the majority classifier (i.e., a constant classifier predicting the most frequent class within the training set) loss $\widehat{\mathcal{L}}_S(h_{\text{maj}})$. Then, the ϵ parameter is chosen so that the loss upper bound lies between these two extreme losses : $(1-p)\widehat{\mathcal{L}}_S(h_S) + p\widehat{\mathcal{L}}_S(h_{\text{maj}})$ with $p \in \{1\%, 5\%, 10\%, 20\%\}$. Notably, the 0%-Rashomon set includes the optimal models and the 100%-Rashomon set includes all models not worse than a majority classifier. The predictive performances of the reference models (achieving the optimal loss $\widehat{\mathcal{L}}_S(h_S)$) are reported in the Appendix B. They confirm that the models' training accuracies are in line with the literature and that they generalize well. Since the Rashomon set is constructed based on training accuracy, we focus our fairness analysis on the training set as well.

Hyperparameters. Our experiments using scoring systems use the same set of possible values for all coefficients: $\Omega_{j \in \{1,..,M\}} = \{0, \pm 1, \pm 2, \pm 5, \pm 10, \pm 20, \pm 30, \pm 50\}$. Sparsity values α (i.e., maximum numbers of non-zero coefficients in the scoring systems) range from 1 to $M+1$ (to account for the additional bias coefficient). Based on preliminary experiments, we fix the skeleton of the decision diagrams to a maximum of 12 internal nodes, distributed across 5 consecutive levels as follows: $(1, 2, 3, 3, 3)$. We consider sparsity values (i.e., the maximum number of active nodes in the decision diagrams) ranging from $\alpha = 4$ to $\alpha = 12$.

Exploration of the Rashomon set. We use the `Gurobi` solver (Gurobi Optimization, LLC 2023) through its Python binding to solve Problems (9) for scoring systems and (18) for decision diagrams. Each solver execution is done on 16 threads using a computing cluster with Intel Platinum 8260 Cascade Lake @2.4GHz CPUs. To speed up our experiments, we exploit the fact that increasing either the allowed sparsity value α or the Rashomon set parameter ϵ relaxes the problem, so we can rely on previously found solutions to hot start the solver. Specifically, each run (for a fixed dataset, random split, sparsity value α, and Rashomon set parameter ϵ) is limited to one hour of CPU time and 36 GB of RAM. For runs where no feasible solution was found or optimality was not proven, we reuse solutions obtained from more constrained versions of the problem (i.e., tighter values of α or ϵ) and restart the solver. Convergence was reached in all runs after at most five such iterations.

[1]The appendices for this paper can be found in the complete version available on ArXiv.

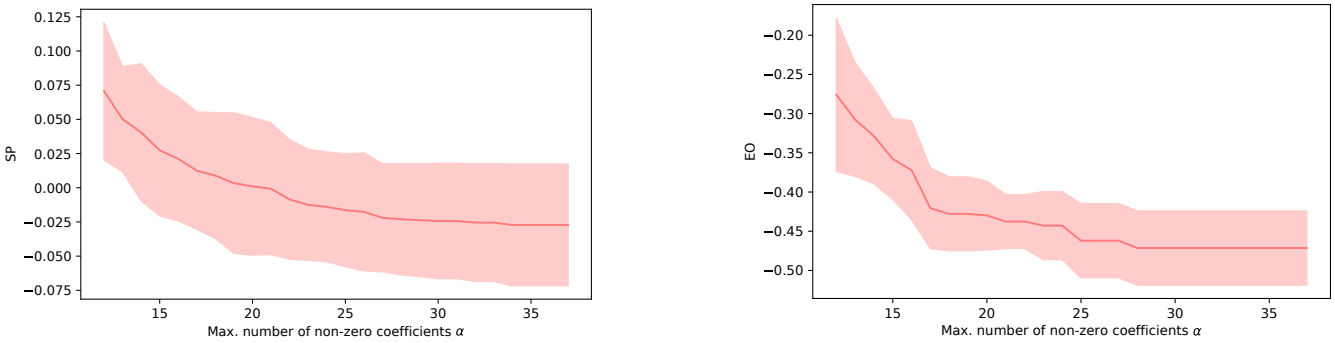

Figure 1: Minimum statistical parity (SP) and equal opportunity (EO) achievable on the training data for the UCI Adult Income dataset, within a 20%-Rashomon set for scoring systems, as a function of the chosen sparsity value α. Negative values favour group G_2 (females) over group G_1 (males), for predicting a high salary. We display both average value and standard deviation.

5.2 Results

We now highlight our key empirical findings and illustrate each of them with a subset of representative results. Complete results, including all datasets, fairness metrics, α and ϵ parameters, and larger training set sizes are provided in the Appendix B for both considered hypothesis classes.

Result 1. Sparsity restricts the range of achievable fairness values and may harm certain protected groups. As discussed earlier, tightening the enforced sparsity value α confines the search to a subset of the hypothesis space $\mathcal{H}_I \subset \mathcal{H}$, which can limit the tradeoffs between various objectives (Dziugaite, Ben-David, and Roy 2020), including fairness and predictive performance. While this result could be expected, our framework allows us to precisely and certifiably quantify this effect. Furthermore, the extent to which sparsity restricts the possible tradeoffs between fairness and predictive performance indicates the severity of the tension between the three desiderata.

For instance, Figure 1 shows the minimum achievable fairness values within a 20%-Rashomon set of scoring systems as a function of the enforced sparsity α for both the statistical parity (left) and equal opportunity (right) metrics on the UCI Adult Income dataset. Negative values for both metrics indicate a bias in favor of group G_2 (females) in predicting high salaries. By quantifying the minimum achievable value, we effectively measure the extent to which females can be advantaged over males given the specified sparsity and performance desiderata.

As expected, tightening the sparsity α reduces the range of achievable fairness. This suggests that enforcing sparsity excludes models with extreme fairness values, highlighting a conflict between these two criteria. Notably, the left plot shows that scoring systems with fewer than $\alpha = 20$ non-zero coefficients systematically disadvantage group G_2 (females), as indicated by the positive minimum fairness values. Hence, if high sparsity is legally required, the resulting outcome imbalance favoring group G_1 (males) could be justified under the principle of "business necessity".

Result 2: Different fairness metrics exhibit different tradeoffs with sparsity. A comparison of the two plots in Figure 1 reveals that the impact of sparsity on the minimum achievable fairness within a 20%-Rashomon set varies depending on the fairness metric considered. Specifically, sparsity consistently disadvantages females in terms of statistical parity (left plot). However, this is not the case for equal opportunity (right plot), where the minimum achievable value remains negative. This difference can be attributed to the fact that, as shown in Equation (3), equal opportunity is conditioned on the true labels and therefore aligns more closely with predictive accuracy, whereas statistical parity does not.

Result 3. High predictive performance requirements restrict the range of achievable fairness values. Table 3a shows the minimum and maximum achievable statistical parity for different Rashomon set parameters ϵ for scoring systems on the Default of Credit Card Clients dataset. We compare two sparsity levels: $\alpha = 15$ (corresponding to the scoring system with the best achievable loss) and $\alpha = 9$ (a sparser, arbitrary value). As previously noted, the range of achievable fairness values narrows with tighter sparsity (smaller α). At fixed sparsity, tightening the predictive performance constraint ϵ further restricts the achievable fairness range. Again, since enforcing tighter performance requirements amounts to shrinking the Rashomon set, this result could be expected. However, the extent to which it is the case indicates the severity of the tension between the two desiderata. Furthermore, it also allows discovering systematic biases, which, since the approach certifiably finds the minimum and achievable fairness values, can be used as legal arguments. For instance, tightening the predictive performance constraint can systematically disadvantage certain protected groups, as evidenced by the fact that the maximum achievable fairness becomes negative for $\epsilon \leq 10\%$ when $\alpha = 9$. This implies that females (group G_2) are (on average) predicted to default on payment more often than males (group G_1) in a systematic manner. In other words, when building a scoring system within 10% of the best predictive performance and limited to at most 9 non-zero coefficients (for interpretability), discrimination against females (in terms of statistical parity) is certifiably inevitable in the Default of Credit Card Clients dataset.

$\alpha = 15$	$\epsilon = 1\%$	$\epsilon = 5\%$	$\epsilon = 10\%$	$\epsilon = 20\%$
Min SP	-0.160	-0.185	-0.215	-0.236
	± 0.060	± 0.062	± 0.057	± 0.048
Max SP	0.059	0.077	0.120	0.137
	± 0.099	± 0.094	± 0.077	± 0.072
$\alpha = 9$	$\epsilon = 1\%$	$\epsilon = 5\%$	$\epsilon = 10\%$	$\epsilon = 20\%$
Min SP	N/A	-0.098	-0.124	-0.176
		± 0.068	± 0.063	± 0.067
Max SP	N/A	-0.017	-0.009	0.053
		± 0.110	± 0.108	± 0.057

(a) Scoring systems

$\alpha = 7$	$\epsilon = 1\%$	$\epsilon = 5\%$	$\epsilon = 10\%$	$\epsilon = 20\%$
Min SP	-0.279	-0.280	-0.310	-0.330
	± 0.139	± 0.138	± 0.112	± 0.103
Max SP	0.193	0.200	0.229	0.244
	± 0.090	± 0.087	± 0.084	± 0.079
$\alpha = 4$	$\epsilon = 1\%$	$\epsilon = 5\%$	$\epsilon = 10\%$	$\epsilon = 20\%$
Min SP	-0.272	-0.273	-0.292	-0.312
	± 0.143	± 0.142	± 0.128	± 0.114
Max SP	0.181	0.197	0.202	0.215
	± 0.096	± 0.088	± 0.081	± 0.073

(b) Decision diagrams

Table 3: Minimal and maximal statistical parity (SP) achievable on the training data for the Default of Credit Card Clients dataset, within different ϵ-Rashomon Sets, for two different sparsity values α, for our experiments on the two considered hypothesis classes. Negative values indicate higher default in payment prediction rates for group G_2 (females) compared to group G_1 (males). N/A indicates that there exists no scoring system satisfying both the sparsity and predictive performance desiderata. We report both the average value and standard deviation.

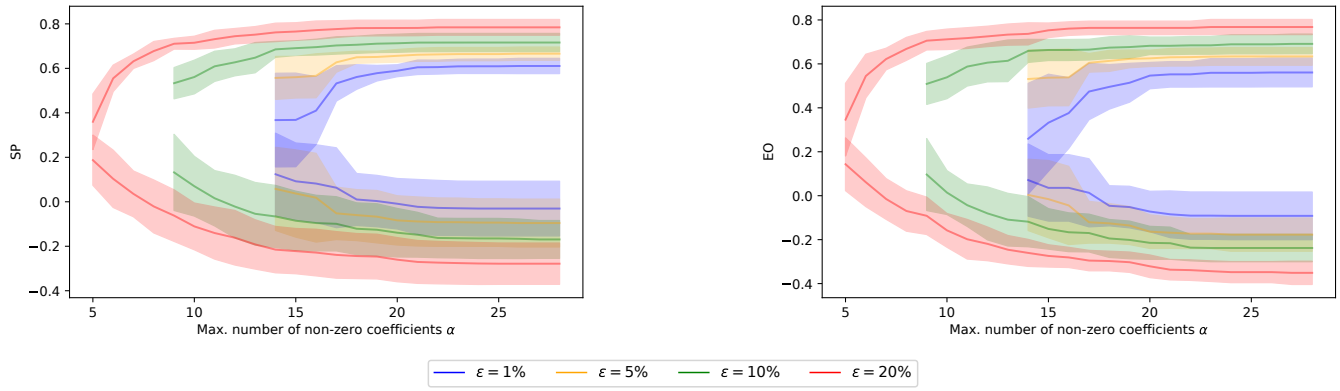

Figure 2: Minimum and maximum statistical parity (SP) and equal opportunity (EO) achievable on the training data for the COMPAS dataset, within different ϵ-Rashomon sets for scoring systems, as a function of the chosen sparsity value α. Positive values indicate higher recidivism prediction rates for group G_1 (African-Americans) compared to group G_2 (the remaining of the population). We report both the average value (line) and standard deviation (colored area).

Result 4. Accuracy, fairness, and sparsity have complex interplays. Figure 2 plots the minimum and maximum achievable fairness values as a function of the desired sparsity level α for different ϵ parameters. The experiments were conducted on scoring systems using the COMPAS dataset and two fairness metrics. This visualization reveals the complex interplays between the three desiderata: predictive performance, fairness, and sparsity. As previously noted, enforcing tighter sparsity (smaller α) narrows the range of achievable fairness values (represented by the gap between the minimum and maximum plotted curves of a given color). Considering tight predictive performance constraints also limits the achievable sparsity values, as indicated by the fact that the curves corresponding to small Rashomon set parameters are unable to reach the smallest sparsity values. For instance, scoring systems within the 1%-Rashomon set exhibit at least 14 non-zero coefficients, while the 20%-Rashomon set contains scoring systems with only 5 non-zero coefficients. Again, our approach offers a precise quantification of the tension between interpretability and predictive accuracy for a given hypothesis class. Here, "business necessity" could justify the inability to reach a target sparsity value.

Moreover, the fact that the curves are not centered around zero highlights inherent tradeoffs between predictive performance and fairness. Specifically, it means that, for a given sparsity level (α) and predictive performance constraint (ϵ), the extent to which one protected group can be favored over the other is greater than the reverse — revealing an asymmetry in how fairness can be achieved. Additionally, when a curve crosses the x-axis (i.e., the horizontal line at SP=0, or EO=0), one protected group becomes systematically disadvantaged across all models in the Rashomon set. For instance, scoring systems with $\alpha = 5$ non-zero coefficients within a 20%-Rashomon set have a minimum statistical parity value of 18.7%, meaning that all models in the Rashomon set consider higher recidivism risks for African-Americans. This bias is also observed with the equal opportunity fairness metric, which is conditioned on the true labels – suggesting that tight enough sparsity requirements systematically amplify existing biases for this experiment. Furthermore, as can

be seen in the complete results provided in the Appendix B, the tradeoffs between accuracy, fairness and sparsity are also influenced by the training data. In particular, the extent to which sparsity or accuracy restrain the range of achievable fairness values differs across datasets. While further investigations on this aspect could be conducted, key factors include the data intrinsic biases towards the considered protected groups, the respective correlations of the different attributes with the labels and the sensitive attributes, as well as the complexity of the underlying classification task.

Result 5. The complexity of the hypothesis class at hand strongly influences the observed tradeoffs. Table 3b reports the minimum and maximum achievable statistical parity for different Rashomon set parameters ϵ, based on our experiments with decision diagrams. We compare two sparsity levels: $\alpha = 7$ (corresponding to the decision diagram with the best achievable loss) and $\alpha = 4$ (a sparser value). The main trends align with the key findings from our experiments on scoring systems: for a fixed sparsity α, tightening the predictive performance requirement (smaller ϵ) restricts the possible fairness ranges. Similarly, for a fixed ϵ, enforcing tighter sparsity further reduces the achievable fairness ranges. The influence of the hypothesis class on the tradeoffs between accuracy, fairness, and sparsity is evident when comparing the results in Table 3b with those in Table 3a (which correspond to scoring systems learned on the same data splits). Decision diagrams offer a broader range of tradeoffs, with fairness ranges that are less constrained by performance and sparsity requirements. Notably, the minimum and maximum values of statistical parity systematically cross zero, implying that disparate impacts are hardly excusable by "business necessity". In other words, the resulting Rashomon sets systematically contain both models favoring group G_2 (females) and models favoring group G_1 (males). These wider ranges are possible because the hypothesis class of decision diagrams is significantly more complex than that of scoring systems. Indeed, the considered decision diagrams partition the input space using multivariate splits (Florio et al. 2023), with each internal node functioning as a linear classifier. In contrast, an entire scoring system corresponds to a linear classifier with integer coefficients: a single internal node of a multivariate decision diagram generalizes it, and we have: $\mathcal{H}_{\text{scoring systems}} \subset \mathcal{H}_{\text{diagrams}}$, even for decision diagrams involving a single internal node. However, this increased complexity comes at the expense of interpretability: understanding the resulting models is more difficult for humans due to the use of multivariate splits. This explains why scoring systems remain very popular in high-stakes applications such as medicine (Rudin et al. 2022). Indeed, the choice of the hypothesis space is another crucial dimension of the complex interplays between the considered ethical desiderata in machine learning. Thorough quantification of the tradeoffs between fairness, sparsity, and predictive accuracy —facilitated by our proposed framework— can empower stakeholders to make informed decisions when navigating these complex interdependencies.

Result 6. Increasing the number of training examples N can tighten the range of achievable fairness values, further highlighting existing discrimination. In the Appendix A, we report results of our experiments using larger training set sizes N. They demonstrate that the proposed approach scales well, and that the observed trends generalize to larger values of N. In fact, they are even exacerbated, as the trained scoring systems better fit the data distribution: the range of achievable fairness values for a given sparsity level (α) and predictive performance threshold (ϵ) narrows with increasing N. For instance, when $N = 500$, the minimum achievable EO within 20%-Rashomon sets for scoring systems on the COMPAS dataset becomes positive when the number of non-zero coefficients is small ($\alpha \leq 6$), as can be seen in Figure 2. This means that for tight enough sparsity requirements, all models in the Rashomon set consider higher recidivism risks for African-Americans, as evidenced through a systematically higher true positive rate. When $N = 4000$ (Figure 3d in the Appendix A), this bias persists across all sparsity requirements and becomes more pronounced as sparsity is tightened.

6 Discussion

This study has demonstrated that mathematical programming approaches can be used to explore the Rashomon set of any hypothesis class without enumeration by making generic modifications to a given baseline learning problem. Specifically, we introduced a framework to characterize fairness and sparsity within the Rashomon set and validated its versatility using two popular types of interpretable models: scoring systems and decision diagrams. The resulting tools enable the identification of sparser, less discriminatory alternative models, representing a significant step toward meeting legal and ethical requirements, despite the inherent challenges (Laufer, Raghavan, and Barocas 2025). The proposed methodology and software can be used by practitioners willing to enforce fairness desiderata to estimate how much accuracy or sparsity they should be ready to sacrifice for different hypothesis classes.

Our extensive experiments highlighted the complex interplays between predictive accuracy, fairness, and sparsity. Our framework not only certifiably quantifies these interplays but also identifies model parameters leading to extreme values, effectively guiding the search for fairer and sparser alternatives. Importantly, we observed that imposing strict predictive performance or sparsity criteria might inherently disadvantage a protected group, underscoring the need for a thorough characterization of these tradeoffs.

The research directions stemming from this work are diverse. First, we propose extending our generic framework to other hypothesis classes, such as rule-based models (Lawless et al. 2023) and tree ensembles, by leveraging recent advances in mathematical programming formulations for interpretable machine learning (Gambella, Ghaddar, and Naoum-Sawaya 2021; Rudin et al. 2022). Additionally, the declarative nature of the framework supports the integration of various additional desiderata, including alternative fairness or robustness metrics, as well as business-specific requirements. Overall, this makes it a promising tool for characterizing the tensions among key properties related to trustworthiness in machine learning.

Acknowledgments

This research was enabled by support provided by Calcul Québec and the Digital Research Alliance of Canada, as well as funding from the SCALE-AI Chair in Data-Driven Supply Chains. The authors would like to thank the anonymous reviewers for their valuable suggestions.

References

Aïvodji, U.; Arai, H.; Fortineau, O.; Gambs, S.; Hara, S.; and Tapp, A. 2019. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*, 161–170. PMLR.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica (2016). *ProPublica, May*, 23.

Aziz, H.; Cseh, Á.; Dickerson, J. P.; and McElfresh, D. C. 2021. Optimal Kidney Exchange with Immunosuppressants. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence, AAAI 2021*, 21–29. AAAI Press.

Black, E.; Koepke, L.; Kim, P.; Barocas, S.; and Hsu, M. 2024. The Legal Duty to Search for Less Discriminatory Algorithms. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024*, Non–archival paper.

Breiman, L. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3): 199–231.

Ciaperoni, M.; Xiao, H.; and Gionis, A. 2024. Efficient Exploration of the Rashomon Set of Rule-Set Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024*, 478–489. ACM.

Coker, B.; Rudin, C.; and King, G. 2021. A theory of statistical inference for ensuring the robustness of scientific results. *Management Science*, 67(10): 6174–6197.

Cooper, A. F.; Lee, K.; Choksi, M. Z.; Barocas, S.; Sa, C. D.; Grimmelmann, J.; Kleinberg, J. M.; Sen, S.; and Zhang, B. 2024. Arbitrariness and Social Prediction: The Confounding Role of Variance in Fair Classification. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence, AAAI 2024*, 22004–22012. AAAI Press.

Coston, A.; Rambachan, A.; and Chouldechova, A. 2021. Characterizing fairness over the set of good models under selective labels. In *International Conference on Machine Learning*, 2144–2155. PMLR.

Dai, G.; Ravishankar, P.; Yuan, R.; Neill, D. B.; and Black, E. 2025. Be Intentional About Fairness!: Fairness, Size, and Multiplicity in the Rashomon Set. *arXiv preprint arXiv:2501.15634*.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference, ITCS 2012*, 214–226.

Dziugaite, G. K.; Ben-David, S.; and Roy, D. M. 2020. Enforcing interpretability and its statistical impacts: Trade-offs between accuracy and interpretability. *arXiv preprint arXiv:2010.13764*.

EEOC., T. U. March 2, 1979. Uniform guidelines on employee selection procedures.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2015*, 259–268. ACM.

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.

Florio, A. M.; Martins, P.; Schiffer, M.; Serra, T.; and Vidal, T. 2023. Optimal Decision Diagrams for Classification. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023*, 7577–7585. AAAI Press.

Gambella, C.; Ghaddar, B.; and Naoum-Sawaya, J. 2021. Optimization problems for machine learning: A survey. *Eur. J. Oper. Res.*, 290(3): 807–828.

Ganesh, P.; Chang, H.; Strobel, M.; and Shokri, R. 2023. On The Impact of Machine Learning Randomness on Group Fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023*, 1789–1800. ACM.

Grover, S. S. 1995. The business necessity defense in disparate impact discrimination cases. *Ga. L. Rev.*, 30: 387.

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.

Gurobi Optimization, LLC. 2023. Gurobi Optimizer Reference Manual.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29: 3315–3323.

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

Kohavi, R. 1994. Bottom-Up Induction of Oblivious Read-Once Decision Graphs: Strengths and Limitations. In *Proceedings of the 12th National Conference on Artificial Intelligence, AAAI 1994*, 613–618. AAAI Press / The MIT Press.

Langenkamp, M.; Costa, A.; and Cheung, C. 2020. Hiring fairly in the age of algorithms. *arXiv preprint arXiv:2004.07132*.

Laufer, B.; Raghavan, M.; and Barocas, S. 2025. Fundamental Limits in the Search for Less Discriminatory Algorithms–and How to Avoid Them. In *ACM CS&Law 2025*.

Lawless, C.; Dash, S.; Gunluk, O.; and Wei, D. 2023. Interpretable and Fair Boolean Rule Sets via Column Generation. *Journal of Machine Learning Research*, 24(229): 1–50.

Marx, C.; Calmon, F.; and Ustun, B. 2020. Predictive multiplicity in classification. In *International Conference on Machine Learning*, 6765–6774. PMLR.

Mata, K.; Kanamori, K.; and Arimura, H. 2022. Computing the Collection of Good Models for Rule Lists. In *18th International Conference on Machine Learning and Data Mining (MLDM 2022)*.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Oliver, J. 1993. Decision graphs – an extension of decision trees. In *Proceedings of the 4th international workshop on artificial intelligence and statistics (AISTATS)*, 343—350.

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.

Rudin, C.; Chen, C.; Chen, Z.; Huang, H.; Semenova, L.; and Zhong, C. 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16: 1–85.

Simson, J.; Pfisterer, F.; and Kern, C. 2024. One Model Many Scores: Using Multiverse Analysis to Prevent Fairness Hacking and Evaluate the Influence of Model Design Decisions. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024*, 1305–1320. ACM.

Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; and Lakkaraju, H. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES 2020*, 180–186.

Ustun, B.; Tracà, S.; and Rudin, C. 2014. Supersparse Linear Integer Models for Interpretable Classification. *arXiv*.

Verma, S.; and Rubin, J. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness, FairWare 2018*, 1–7. Association for Computing Machinery.

Watson-Daniels, J.; Parkes, D. C.; and Ustun, B. 2023. Predictive Multiplicity in Probabilistic Classification. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023*, 10306–10314. AAAI Press.

Xin, R.; Zhong, C.; Chen, Z.; Takagi, T.; Seltzer, M.; and Rudin, C. 2022. Exploring the whole rashomon set of sparse decision trees. *Advances in neural information processing systems*, 35: 14071–14084.

Yeh, I.-C.; and hui Lien, C. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2, Part 1): 2473–2480.

Zhong, C.; Chen, Z.; Liu, J.; Seltzer, M.; and Rudin, C. 2024. Exploring and interacting with the set of good sparse generalized additive models. *Advances in neural information processing systems*, 36.