Rashomon Capacity: Measuring Predictive Multiplicity in Probabilistic Classification

Hsiang Hsu [1] **Flavio P. Calmon** [2]

Abstract

Predictive multiplicity occurs when classification models with nearly indistinguishable average performances assign conflicting predictions to individual samples. When used for decision-making in applications of consequence (e.g., lending, education, criminal justice), models developed without regard for predictive multiplicity may result in unjustified and arbitrary decisions for specific individuals. We introduce a new measure of predictive multiplicity in probabilistic classification called Rashomon Capacity. We show that Rashomon Capacity yields principled strategies for disclosing conflicting models to stakeholders. Our numerical experiments illustrate how Rashomon Capacity captures predictive multiplicity in various datasets and learning models, including neural networks. The tools introduced in this paper can help data scientists measure, report, and ultimately resolve predictive multiplicity prior to model deployment. Full paper available at https://arxiv.org/abs/2206.01295.

1. Introduction

Rashomon effect, introduced by Breiman (2001), describes the phenomenon where a multitude of distinct predictive models achieve similar training or test loss. The set of almost-equally performing models for a given learning problem is called the *Rashomon set* (Fisher et al., 2019; Semenova et al., 2019). In classification, models across the Rashomon set can have high *predictive multiplicity* (Marx et al., 2020): classifiers with similar average performance may assign wildly different predictions to a sample.

Predictive multiplicity captures potential individual-level harm introduced by an arbitrary choice of a single model in the Rashomon set. When such a model is used to support automated decision-making in sectors dominated by a few companies or Government—labeled *Algorithmic Leviathans* in Creel & Hellman (2021, Section 3)—predictive multiplicity results in unjustified and systemic exclusion of individuals from critical opportunities and introduce feedback loops that amplify systemic biases. For example, Governments are increasingly turning to algorithms for grading exams that grant access to higher-level education (e.g., UK (Smith, 2020)). Here, accounting for predictive multiplicity is critical: an arbitrary choice of a single model may lead to an unwarranted restriction of educational opportunities to an individual student. In applications such as criminal justice and healthcare, models that do not account for predictive multiplicity are at risk of causing individual-level harm by supporting decisions that may at first appear to be data-driven, but are in fact the result of arbitrary choices during training (e.g., parameter initialization). Predictive multiplicity must be reported to stakeholders in, for example, model cards (Mitchell et al., 2019).

We introduce a new predictive multiplicity metric called *Rashomon Capacity* for probabilistic classifiers[1]. Unlike prior metrics based on thresholded (i.e., 0 or 1) predictions, Rashomon Capacity captures more nuanced variations in scores among models in the Rashomon set for a target input sample, and possesses several properties that a predictive multiplicity metric must satisfy to simplify its interpretation by stakeholders. The computation of Rashomon Capacity also sheds light on a strategy for resolving predictive multiplicity. Different approaches have been proposed for dealing with multiplicity, including randomizing between competing classifiers (Creel & Hellman, 2021) and bagging (Breiman, 2001). However, the size of the Rashomon set can be large, making strategies that require randomizing predictions across the entire Rashomon set potentially arbitrary and impractical. In contrast, we apply standard results from convex analysis to show that Rashomon Capacity for an input sample can be entirely captured by at most c models in the Rashomon set, where c is the number of

[1]Department of Computer Science, Harvard University, Cambridge, USA [2]Department of Electrical Engineering, Harvard University, Cambridge, USA. Correspondence to: Hsiang Hsu <hsianghsu@g.harvard.edu>.

ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments, Baltimore, Maryland, USA, 2022. Copyright 2022 by the author(s).

[1]A probabilistic classifier is a model that maps an input sample onto a probability distribution, referred as a score, over a discrete set of classes. Examples of probabilistic classifiers include logistic regression and a neural network with a softmax output layer.

predicted classes. This results holds *regardless of the size of the Rashomon set*. Thus, when c is small, the predictions produced by the competing classifiers can be communicated to a stakeholder, empowering them to decide how to resolve conflicting decisions. *Omitted proofs, experimental details and codes will be attached upon acceptance.*

2. Related Work

The Rashomon effect impacts model selection (Rudin, 2019; Hancox-Li, 2020; D'Amour et al., 2020), explainability (Pawelczyk et al., 2020), and fairness (Coston et al., 2021). Rudin (2019) suggested that, given the choice of competing models, machine learning (ML) practitioners should select interpretable models *a priori*, rather than selecting a black-box model with conjectural explanations afterwards. Hancox-Li (2020) and D'Amour et al. (2020) further argued that epistemic patterns, e.g., causality, should be specified in the ML pipeline, and the selected models from the Rashomon set should be able to reflect these patterns. Competing models in the Rashomon set may not only render conflicting explanations for predictions (Pawelczyk et al., 2020) and measures of variables' importance (Fisher et al., 2019), but also have inconsistent performance across population sub-groups (Coston et al., 2021).

We next introduce notations and existing metrics for measuring predictive multiplicity.

Notations. We consider a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ for a classification task with c classes/labels, where each sample pair $(\mathbf{x}_i, \mathbf{y}_i) \overset{i.i.d.}{\sim} P_{X,Y}$ with support $\mathcal{X} \times \Delta_c$, where Δ_c is the c-dimensional probability simplex. Let $[\cdot]_j$ denotes the j^{th} entry of a vector, $\mathbf{e}_k$ be a length-c indicator vector, i.e., $[\mathbf{e}_k]_k = 1$, and $[\mathbf{e}_k]_j = 0 \ \forall j \neq k$. We denote by $\mathcal{H}$ a *hypothesis space*, i.e., a set of candidate probabilistic classifier is parameterized by $\theta \in \Theta \subseteq \mathbb{R}^d$ that approximate $P_{Y|X=\mathbf{x}_i}$, i.e., $\mathcal{H} \triangleq \{h_\theta : \mathcal{X} \to \Delta_c : \theta \in \Theta\}$. The loss function used to evaluate model performance is denoted by $\ell : \Delta_c \times \Delta_c \to \mathbb{R}^+$ (e.g., cross-entropy) and $\hat{L}(h) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(h(\mathbf{x}_i), \mathbf{y}_i)$ the empirical risk.

Rashomon set and pattern Rashomon ratio. Given a Rashomon parameter $\epsilon \geq 0$, Semenova et al. (2019) and Marx et al. (2020) respectively define a Rashomon set and an ϵ-level set as $\mathcal{R}(\mathcal{H}, \epsilon) \triangleq \{h \in \mathcal{H}; \hat{L}(h) \leq \epsilon\}$. Determining $\mathcal{R}(\mathcal{H}, \epsilon)$ is essentially a level set estimation problem (Mason et al., 2021), and is computationally infeasible when the hypothesis space $\mathcal{H}$ is large (Bachoc et al., 2021). For predictive multiplicity in classification problems, Semenova et al. (2019, Defn. 12) further proposed *pattern Rashomon ratio* for binary classification, which is the ratio of the count of all possible binary predicted classes given by the functions in the Rashomon set to that given by the functions in the hypothesis space. Note that the computational complexity of the pattern Rashomon ratio grows exponentially with the number of samples, and could be an "expensive" metric for predictive multiplicity when applied on a large dataset.

Ambiguity and discrepancy. Marx et al. (2020) proposed ambiguity and discrepancy to measure multiplicity in terms of the thresholded outputs (i.e., predicted classes) of a classifier in terms of classification accuracy. In probabilistic classification, thresholding may mask similar predictions produced by competing models and artificially increase multiplicity: output scores can be almost equal across different classes, yet the (thresholded) predicted classes can be very different. For example, two scores $[0.49, 0.51]$ and $[0.51, 0.49]$ for a binary classification problem can lead to entirely different predicted classes—1 and 0, respectively—and ultimately overestimate predictive multiplicity. This subtle, yet important difference motivates us to reconsider existing metrics and introduce a new predictive multiplicity metric for probabilistic classifiers that output scores.

3. Measuring Predictive Multiplicity of Probabilistic Classifiers

We outline desirable properties of predictive multiplicity metrics for probabilistic classifiers which provide guidelines for the creation of new multiplicity metrics in future research. Next, we formally define Rashomon Capacity in terms of the KL-divergence between the output scores of classifiers in the Rashomon set.

Properties. Consider a Rashomon set $\mathcal{R}(\mathcal{H}, \epsilon)$ for a classification problem with c classes. We collect all possible output scores for a sample $\mathbf{x}_i \in \mathcal{D}$ and define the *ϵ-multiplicity set* as $\mathcal{M}_\epsilon(\mathbf{x}_i) \triangleq \{h(\mathbf{x}_i) \mid h \in \mathcal{R}(\mathcal{H}, \epsilon)\} \subseteq \Delta_c$. Let $m(\cdot)$ be a measure of predictive multiplicity, and $m(\mathcal{M}_\epsilon(\mathbf{x}_i))$ be the predictive multiplicity of sample $\mathbf{x}_i$. Ideally, we expect $m(\mathcal{M}_\epsilon(\mathbf{x}_i))$ to be a bounded value in $[1, c]$, since at least one class is assigned to sample $\mathbf{x}_i$, and at most c different classes could be assigned to $\mathbf{x}_i$. If $m(\mathcal{M}_\epsilon(\mathbf{x}_i)) = 1$, only one score is produced for $\mathbf{x}_i$ and all predictions in $\mathcal{M}_\epsilon(\mathbf{x}_i)$ are exactly the same. Similarly, if $m(\mathcal{M}_\epsilon(\mathbf{x}_i)) = c$, then there must exist c models $\{h_1, \cdots, h_c\} \subseteq \mathcal{R}(\mathcal{H}, \epsilon)$ such that $h_j(\mathbf{x}_i) = \mathbf{e}_j$. Finally, $m(\mathcal{M}_\epsilon(\mathbf{x}_i))$ should be monotonic. We summarize the desirable properties of predictive multiplicity metrics in the following definition.

Definition 1. *Let $\mathcal{S}_c \triangleq \{\{\mathbf{y}\} \mid \mathbf{y} \in \Delta_c\}$ be the set of singleton sets in Δ_c, and $\sigma(\mathcal{S}_c)$ its corresponding σ-algebra. We say that a function $m : \sigma(\mathcal{S}_c) \to \mathbb{R}$ is a* predictive multiplicity metric *if for any $\mathcal{A}, \mathcal{B} \in \sigma(\mathcal{S}_c)$, (i) $1 \leq m(\mathcal{A}) \leq c$; (ii) $m(\mathcal{A}) = 1$ if and only if $|\mathcal{A}| \leq 1$; (iii) $m(\mathcal{A}) = c$ if and only if $\mathbf{e}_k \in \mathcal{A}$ for $k \in [c]$, i.e., $\mathcal{A}$ contains the corner points of Δ_c; (iv) $m(\mathcal{A}) \leq m(\mathcal{B})$ if $\mathcal{A} \subseteq \mathcal{B}$.*

We introduce next a predictive multiplicity metric called Rashomon Capacity that satisfies all properties above.

Rashomon Capacity. Our goal is to quantify predictive multiplicity by the score difference assigned to each point $\mathbf{x}_i$ in $\mathcal{D}$. It is natural to adopt divergence measures for distributions to capture the "variation" of scores in $\mathcal{M}_\epsilon(\mathbf{x}_i)$. From a geometric viewpoint, a larger spread in scores indicates a greater amount of predictive multiplicity for a sample.

Assume a probability (or "weight") distribution P_M across models in $\mathcal{R}(\mathcal{H}, \epsilon)$ (and therefore each score in $\mathcal{M}_\epsilon(\mathbf{x}_i)$), where M denotes the random variable of selecting/sampling the models in the Rashomon set. Intuitively, if P_M assigns mass 1 to a single model and 0 to all other models in the Rashomon set, then the output of only one model is considered. Conversely, if P_M is the uniform distribution, then the outputs of every model in the set are equally weighed. Given a divergence measure between distributions $d(\cdot\|\cdot)$, we quantify the spread of the scores in $\mathcal{M}_\epsilon(\mathbf{x}_i)$ by

$$\rho(\mathcal{M}_\epsilon(\mathbf{x}_i), P_M) \triangleq \inf_{\mathbf{q}\in\Delta_c} \mathbb{E}_{h\sim P_M} d(h(\mathbf{x}_i)\|\mathbf{q}). \quad (1)$$

Here, the minimizing $\mathbf{q}$ acts as a "center of gravity" or "centroid" for the outputs of the classifiers in the Rashomon set for a chosen distribution P_M across models. Analogously, the quantity $\rho(\mathcal{M}_\epsilon(\mathbf{x}_i), P_M)$ can be understood as a measure of "spread" or "inertia" across model outputs. We select the distribution P_M that results in the largest spread in scores:

$$C_d(\mathcal{M}_\epsilon(\mathbf{x}_i)) \triangleq \sup_{P_M} \rho(\mathcal{M}_\epsilon(\mathbf{x}_i), P_M). \quad (2)$$

A natural candidate for $d(\cdot\|\cdot)$ is KL-divergence, and we name the spread in scores measured as *Rashomon Capacity*.

Definition 2. *Given a sample $\mathbf{x}_i$ and the ϵ-multiplicity set $\mathcal{M}_\epsilon(\mathbf{x}_i)$, the Rashomon Capacity is defined as*

$$C(\mathcal{M}_\epsilon(\mathbf{x}_i)) \triangleq \sup_{P_M} \inf_{\mathbf{q}\in\Delta_c} \mathbb{E}_{h\sim P_M} D_{KL}(h(\mathbf{x}_i)\|\mathbf{q}). \quad (3)$$

Moreover, we define $m_C(\mathbf{x}_i) \triangleq 2^{C(\mathcal{M}_\epsilon(\mathbf{x}_i))}$.

The quantity $C(\mathcal{M}_\epsilon(\mathbf{x}_i))$ is ubiquitous in information theory; in fact, $C(\mathcal{M}_\epsilon(\mathbf{x}_i))$ is the *channel capacity* (Cover, 1999) of a channel $P_{Y|M}$ whose rows are the entries of $\mathcal{M}_\epsilon(\mathbf{x}_i)$. This connection motivates the name "Rashomon Capacity" and is useful for proving that $m_C(\mathbf{x}_i)$ is indeed a predictive multiplicity metric, stated in the next proposition.

Proposition 1. *The function $m_C(\cdot) = 2^{C(\mathcal{M}_\epsilon(\cdot))} : \mathcal{X} \to [1, c]$ satisfies all properties of a predictive multiplicity metric in Definition 1.*

Computation. The definition of Rashomon Capacity does not assume a finite cardinality of the Rashomon set. Remarkably, even when the Rashomon set has infinite cardinality, the value of Rashomon Capacity for a sample can be recovered by considering only a small number of models in the Rashomon set. In fact, for each sample $\mathbf{x}_i$, there exists a subset of at most c models that fully captures the variation in scores. This statement is formalized by the next proposition, which can be proven by applying Carathéodory's theorem (Carathéodory, 1911).

Proposition 2. *For each sample $\mathbf{x}_i \in \mathcal{D}$, there exists a subset $\mathcal{A} \subseteq \mathcal{M}_\epsilon(\mathbf{x}_i)$ with $|\mathcal{A}| \leq c$ that fully captures the spread in scores for $\mathbf{x}_i$ across the Rashomon set, i.e., $m_C(\mathbf{x}_i) = 2^{C(\mathcal{A})}$. In particular, there are at most c models in $\mathcal{R}(\mathcal{H}, \epsilon)$ whose output scores yield the same Rashomon Capacity for $\mathbf{x}_i$ as the entire Rashomon set.*

With the discrete $\mathcal{A}$, Rashomon Capacity can be computed by the Blahut–Arimoto (BA) algorithm (Blahut, 1972; Arimoto, 1972). In terms of Rashomon Capacity, the (at most) c conflicting scores capture the predictive multiplicity *across the entire Rashomon set*. The stakeholder can then choose to randomize between scores, accept the average score, or apply another appropriate strategy.

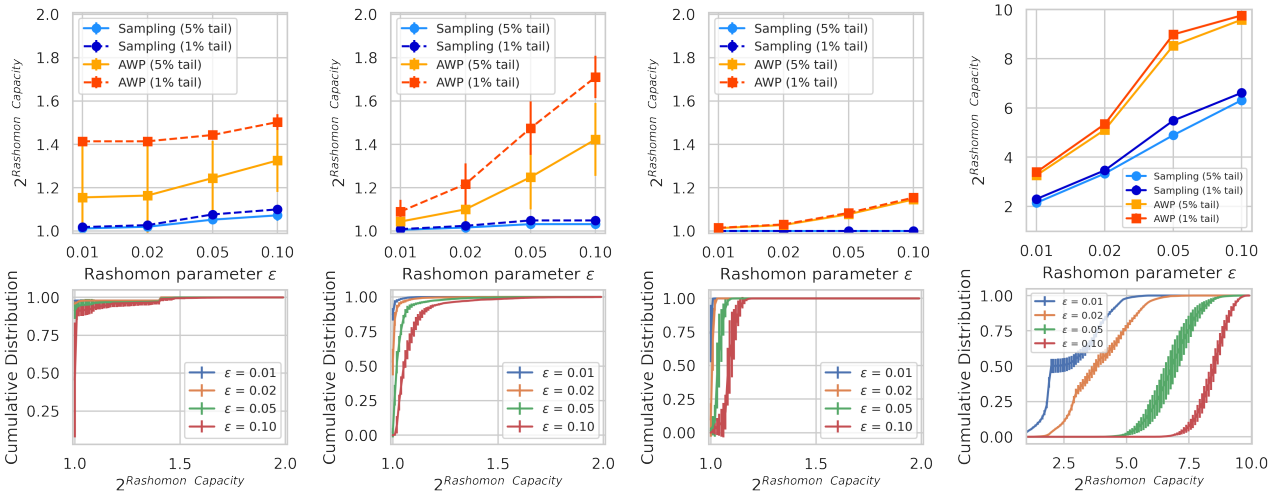
4. Empirical Study

We illustrate how to measure, report, and resolve predictive multiplicity of probabilistic classifiers using Rashomon Capacity on UCI Adult (Lichman, 2013), COMPAS (Angwin et al., 2016), HSLS (Ingels et al., 2011), and CIFAR-10 datasets (Krizhevsky et al., 2009). The HSLS is an education dataset, collected from high school students in the USA, whose features include student and parent information, and the binary label Y is students' 9^{th}-grade math test scores. We include the CIFAR-10 dataset to demonstrate how to report Rashomon Capacity in multi-class classification. We adopt feed-forward neural networks for the first three datasets, and a convolutional neural network VGG16 (Simonyan & Zisserman, 2014) for CIFAR-10. All numbers reported are evaluated on the test set.

Measuring and reporting Rashomon Capacity. We describe one simple method for navigating the Rashomon set next and, later in the section, we also consider sampling models in the Rashomon set via random initialization of parameters prior to training (see Section 2 for alternative strategies). Assume the models in $\mathcal{H}$ are parameterized. Given a sample $\mathbf{x}_i$, we obtain models with output predictions $\mathbf{p}_k$ by approximately solving the following optimization problem which maximizes the output score forall class $k = [c]$:

$$\mathbf{p}_k = h_{\hat{\theta}}(\mathbf{x}_i), \text{ where } \hat{\theta} = \underset{\theta\in\Theta,\ h_\theta\in\mathcal{R}(\mathcal{H},\epsilon)}{\arg\max} [h_\theta(\mathbf{x}_i)]_k. \quad (4)$$

To solve (4), for each k, we set the objective to be $\min_{\theta\in\Theta} -[h_\theta(\mathbf{x}_i)]_k$, compute the gradients, and update the



(a) UCI Adult (80.28%) (b) COMPAS (66.22%) (c) HSLS (70.39%) (d) CIFAR-10 (81.67%)

Figure 1. For each dataset (percentage is test accuracy), the top figure shows the mean and standard error of the largest 1% and 5% (1% tail and 5% tail in the legend) Rashomon Capacity among all the samples with difference Rashomon parameter ϵ. Two methods are used to obtain models from the Rashomon set, `AWP` (4) and random `sampling`. The bottom figure shows the cumulative distribution of the Rashomon Capacity of all the samples obtained by `AWP`. Each point is generated with 5 repeated splits of the dataset.

parameter θ until $h_\theta \notin \mathcal{R}(\mathcal{H}, \epsilon)$, i.e., $\hat{L}(h_\theta) > \epsilon$. Given a pre-trained model in the Rashomon set, (4) can be viewed as an adversarial weight perturbations (`AWP`) technique to explore the Rashomon set (Wu et al., 2020; Tsai et al., 2021). With the discrete set of scores collected by solving (4), the Rashomon Capacity can be computed by the BA algorithm.

We preform two methods, random `sampling` with different weight initialization seeds and `AWP` (4), to obtain 100 models from the Rashomon set, and report the Rashomon Capacity in Fig. 1. In particular, we show the mean of the largest 1% and 5% Rashomon Capacity, and the cumulative distribution of the Rashomon Capacity across the samples. As the Rashomon parameter increases, both `sampling` and `AWP` lead to higher Rashomon Capacity since the Rashomon set gets larger. The `AWP` (4) achieves higher Rashomon Capacity than random `sampling` as `AWP` intentionally explores the Rashomon set that maximizes the scores variations. It is important to keep in perspective that *each sample* in the high-Rashomon Capacity tail displayed in Fig. 1. corresponds to an *individual* who receive conflicting predictions. In applications such as criminal justice and education, conflicting predictions for even one individual should be reported in, e.g., model cards (Mitchell et al., 2019).

Resolving predictive multiplicity. We propose a greedy model selection procedure to reduce the number of competing classifiers for resolving predictive multiplicity. Given R competing classifiers, the goal is to select r models ($r < R$) that result in distributions of the Rashomon Capacity similar to that of the original R models. Starting from a dataset $\mathcal{D}$ and a Rashomon set $\mathcal{R}(\mathcal{H}, \epsilon)$, this can be implemented by (i) initializing a set $\mathcal{A}$ of models by randomly selecting a model in $\mathcal{R}(\mathcal{H}, \epsilon)$, (ii) growing $\mathcal{A}$ by adding one model from $\mathcal{R}(\mathcal{H}, \epsilon)$ that maximizes the average Rashomon Capacity across $\mathcal{D}$, and (iii) stopping until there are r models in $\mathcal{A}$. This greedy model selection is inspired by Property 4 (monotonicity) in Definition 1, since including the models to the set $\mathcal{A}$ does not reduce capacity.

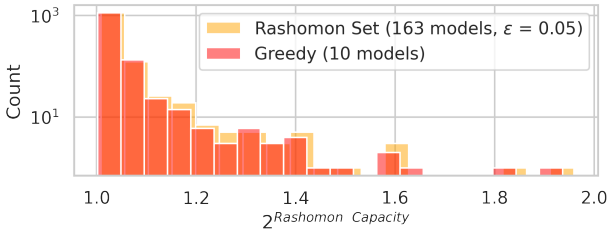


Figure 2. The distributions of the Rashomon Capacity for COMPAS datasets with mean test accuracy 67.35%, obtained by sampling models from the Rashomon set and applying greedy model selection procedure (Greedy in the legend) on the sampled models.

In Fig. 2, we sampled 163 and 52 models from the Rashomon sets for COMPAS and HSLS datasets respectively. Here, the hypothesis space are feed-forward neural networks. Observe that only a small subset of the sampled models, selected by the greedy model selection procedure, is required to recover the distribution of the Rashomon Capacity. On COMPAS dataset, the 10 models obtained by the greedy model selection procedure capture the Rashomon Capacity computed with the original 163 models., i.e., these 10 models "explain" most of the score variations.

Acknowledgement

This material is based upon work supported by the National Science Foundation under grants CAREER 1845852, IIS 1926925, and FAI 2040880, and by Meta Ph.D. fellowship.

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. *ProPublica*, 2016.

Arimoto, S. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.

Bachoc, F., Cesari, T., and Gerchinovitz, S. The sample complexity of level set approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 424–432. PMLR, 2021.

Blahut, R. Computation of channel capacity and rate-distortion functions. *IEEE transactions on Information Theory*, 18(4):460–473, 1972.

Breiman, L. Statistical modeling: The two cultures. *Statistical science*, 16(3):199–231, 2001.

Carathéodory, C. Über den variabilitätsbereich der fourier'schen konstanten von positiven harmonischen funktionen. *Rendiconti Del Circolo Matematico di Palermo (1884-1940)*, 32(1):193–217, 1911.

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

Cover, T. M. *Elements of information theory*. John Wiley & Sons, 1999.

Creel, K. and Hellman, D. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision making systems. *Virginia Public Law and Legal Theory Research Paper*, 2021.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

Fisher, A., Rudin, C., and Dominici, F. 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, 2019.

Hancox-Li, L. Robustness in machine learning explanations: does it matter? In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 640–647, 2020.

Ingels, S. J., Pratt, D. J., Herget, D. R., Burns, L. J., Dever, J. A., Ottem, R., Rogers, J. E., Jin, Y., and Leinwand, S. High school longitudinal study of 2009 (hsls: 09): Base-year data file documentation. nces 2011-328. *National Center for Education Statistics*, 2011.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images (technical report). *University of Toronto*, 2009.

Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

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

Mason, B., Camilleri, R., Mukherjee, S., Jamieson, K., Nowak, R., and Jain, L. Nearly optimal algorithms for level set estimation. *arXiv preprint arXiv:2111.01768*, 2021.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.

Pawelczyk, M., Broelemann, K., and Kasneci, G. On counterfactual explanations under predictive multiplicity. In *Conference on Uncertainty in Artificial Intelligence*, pp. 809–818. PMLR, 2020.

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

Semenova, L., Rudin, C., and Parr, R. A study in rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. *arXiv preprint arXiv:1908.01755*, 2019.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Smith, H. Algorithmic bias: should students pay the price? *AI & society*, 35(4):1077–1078, 2020.

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

Wu, D., Xia, S.-T., and Wang, Y. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020.