Ensemble Learning through Rashomon Sets

Gianlucca Zuin[1,2], **Adriano Veloso**[1,2]

[1]C.S. Dept. – Universidade Federal de Minas Gerais
31.270-901 – Av. Antônio Carlos, 6627
Belo Horizonte – Brazil

[2]Kunumi
30130-131 – R. Rio Grande do Norte, 1435
Belo Horizonte – Brazil

{gzuin,adrianov}@dcc.ufmg.br

Abstract. *Creating models from previous observations and ensuring effectiveness on new data is the essence of machine learning. However, selecting models that generalize well to future data remains a challenging task. In this work, we investigate how models perform across datasets with distinct underlying data generation functions but constitute co-related tasks. The key motivation is to study the Rashomon Effect, which appears whenever the learning problem admits a set of models that all perform roughly equally well. Real-world problems often exhibit multiple local structures in data space, leading to a non-convex error surface and multiple high-performing models which literature suggests to be subject to the Rashomon Effect. Our approach is to stratify, during training, the solution space into model groups that are either coherent or contrasting given both performance and explanations. From these Rashomon groups, we build an ensemble ensuring that each constituent covers a distinct region of the solution space. We validate our approach by performing a series of experiments in both open and closed real-world datasets. Our method outperforms state-of-the-art techniques, improving AUROC by up to 0.20+ when the Rashomon ratio is large.*

Resumo. *Criar modelos a partir de observações e garantir sua eficácia em novos dados é a essencia do aprendizado de máquina. No entanto, selecionar modelos que generalizem bem para dados futuros continua sendo uma tarefa desafiadora. Neste trabalho, investigamos como os modelos se comportam em conjuntos de dados com funções de geração de dados distintas mas ainda correlacionadas. A motivação é estudar o Efeito Rashomon, que ocorre quando um problema admite a existência de vários modelos distintos com desempenho semelhante. Problemas do mundo real frequentemente exibem múltiplas estruturas locais nos dados, resultando em múltiplos modelos de alto desempenho sujeitos ao Efeito Rashomon. Propomos estratificar durante treino o espaço de soluções em grupos de modelos coerentes ou contrastantes. A partir desses grupos de Rashomon, contruimos um comitê onde cada constituinte cobre uma região distinta do espaço. Validamos nossa abordagem em conjuntos de dados abertos e reais. Nossa abordagem supera o estado-da-arte, melhorando a AUROC em até 0,20+ nos cenários onde a razão de Rashomon é alta.*

1. Introduction

Model selection is crucial in both industry and research, and the widely adopted approach is cross-validation. Although cross-validation generally provides robust risk estimation, it may fail for specific problems depending on the model selection goal. The empirical risk on a test set might not always correlate with real-world performance. Particularly, the empirical risk can be significantly influenced when different models perform similarly on the test set (Hinns et al. 2021).

The Rashomon Effect, also known as the multiplicity of good models (Breiman 2001), presents a phenomenon where many models perform equally well, yet they process data in substantially different ways, making it challenging to draw reliable conclusions or automate decisions based on a single model fit (Zuin et al. 2020; Zuin et al. 2023b). In this study, we investigate model performance in datasets with different underlying generator functions while constituting correlated tasks. A significant challenge arises when a cross-validated model, carefully selected during training, encounters data drawn from a different distribution during production. Cross-validation guarantees no longer apply to out-of-distribution data, resulting in unpredictable model behavior and rendering held-out performance an unreliable risk estimation. To address this issue, we extend our analysis beyond empirical risk.

Our main hypothesis posits that some models exhibit similar behavior only when data is drawn from the same distribution as seen during training. Instead of relying on a single model, which may struggle with complex datasets containing multiple local structures, we assemble contrasting models from different sub-populations of the solution space. We propose leveraging the Rashomon set and dividing it into subspaces, using the explanation of each model as a driver for the partitioning. Combining individuals from each subspace produces ensembles with varied perspectives, offering insights into the different facets of the problem. As each constituent offers a different explanation for the target phenomenon, the ensemble's output is directly linked to the trustworthiness of its prediction. Consensus among constituents indicates a match between the data distribution and the one seen during training, with all cross-validation guarantees holding. Disagreement suggests that the properties cannot be trusted.

We believe that diversity among individual models is crucial for gaining an understanding of any phenomenon. Further, we assume that problems are not tied to a single factor. The Rashomon Effect suggests the existence of multiple potential explanations for a given problem, all consistent with the data. To encourage diversity and identify patterns, we group models based on the similarity of their explanations. Ideally, this leads to dense groups where models share common explanatory factors. For each group, we select the most distinct models, also evaluating cohesion in a simulated dataset with perturbation. This results in an ensemble that is diverse in its constituents, incorporates high-performing models, summarizes the entire Rashomon set and solution space, and allows for an approximation of a risk metric under new data distributions based on constituent agreement. We coin this idea as the Rashomon Ensemble. Our approach involves the following steps:

1. Sample models from a pre-defined Rashomon subspace (set of models with equivalent empirical risk).

2. Compute the explanation vector of the sampled models and their pairwise similarity.
3. Perturbate a held-out test data through some data transformation.
4. Compute the pairwise distance in the transformed test set.
5. Split the Rashamon set into subgroups based on models' explanation vectors and distances.
6. Select a set of models with contrasting explanations and divergent predictions on the transformed data.
7. Build an ensemble and evaluate agreement to estimate reliability.

We validate our approach on a set of public datasets for reproducibility and demonstrate its robustness in simulated scenarios. Our results show that Rashomon ensembles consistently outperform state-of-the-art ensemble learning approaches if the Rashomon set is large enough. When exposed to data drift, our approach remained the performant one in most evaluated scenarios providing further evidence of its reliability. We proceed to employ the Rashomon ensembles in three real-world applications partnered with various industries and institutions, studying the impact of our approach.

2. Related Work

Many methods exist for capturing model uncertainty and improving prediction robustness, including ensemble modeling (Madras et al. 2020). Recent focus has centered on Neural Networks (NNs) and their intermediary features, particularly for Out-of-Distribution (OoD) detection (Chen et al. 2021). The main difference in our work lies in the analysis of additional unexplored axes, such as the decision-making process of a model via their explanatory factors (Lundberg and Lee 2017). A second key idea is to exploit the Rashomon Effect to look for models with similar performance during training. The Rashomon Effect defines a set of close-to-optimal models sharing similar performance (Fisher et al. 2019). We leverage the Rashomon set, defined relative to a reference model and allowing comparison based on performance and explanation proximity. Another fundamental aspect of our work and for comprehending the Rashomon set is the Rashomon ratio, as introduced by (Semenova and Rudin 2019). It represents the portion of models in the overall solution space that belong to a specific Rashomon set for a given problem. A high ratio suggests numerous diverse solutions, enabling the location of less complex and more robust models. A much more detailed discussion about related work can be found in (Zuin 2023).

3. Problem Formulation

We consider a supervised learning scenario and formulate a classification model as a function $f(X, Y; \theta)$ parameterized by θ that maps inputs $x_i \in X$ to labels $y_i \in Y$. During cross-validation, we train models on data D_{train} coming from a distribution T. To estimate the predictive risk of each function, we employ additional data D_{test} from the same distribution T and evaluate $f_n \in F$ on this independent and identically distributed data. The standard model selection step involves selecting the function that minimizes the empirical predictive risk, providing performance guarantees when future data follows the same distribution T. However, these guarantees do not hold when dealing with data coming from other distributions, such as in the case of data drift.

Our main objective is to build a diverse ensemble comprising different and contrasting explanations for the same problem. Additionally, we aim to estimate the reliability of our predictions under uncertainty arising from an unknown data distribution U, which may contain drift compared to the training data distribution T. To achieve this, we explore how models behave when the differences between different executions are only minor. We consider θ to encompass any choices made during training that lead to virtually similar models exhibiting contrasting performances. We then introduce drift to the test data and evaluate its effects on each model.

Instead of simply mixing multiple different structures into a single model and minimizing the objective function $f(x)$, we sample the model space by minimizing different functions $f(x')$, where $x' \subseteq x$ and $|x'| < |x|$ (Zuin et al. 2020). This sampling strategy resembles the Rashomon set concept, as it acknowledges the existence of multiple valid and diverse models that perform well in different regions of the data space. By exploring the Rashomon set and considering models with contrasting explanations (Zuin and Veloso 2019; Zuin et al. 2020), we can identify subgroups of correlated features and build ensembles with diverse models that contribute unique explanations for different facets of the data. This approach enhances the robustness of our solution by considering the multiplicity of performant models.

4. Rashomon Ensembles

We build our ensemble exploiting two concepts: diversity between individual models and stability between model explanation and empirical predictions (Shmueli 2010). Diversity is crucial for gaining a general understanding of a phenomenon, assuming that problems are not tied to a single explanatory factor, and explanatory factors may vary depending on factors that might not be directly intuitive. We can understand an explanatory factor as a vector obtained after using some explainability framework, such as SHAP (Lundberg and Lee 2017) and feature importance, to understand what drives model predictions. To promote diversity while finding patterns, we cluster models in $\mathcal{F}'$ based on the distance between their explanation vectors. Ideally, this creates numerous groups of models that are internally dense and separated from other models in terms of their explanatory factors. Stability, on the other hand, refers to models within a cluster being associated with the same explanatory factors and performing similar predictions.

To assess prediction-explanation stability, we consider this distance between the explanation vectors and project the found clusters into the prediction space. This allows us to locate different Rashomon subgroups inside the Rashomon set and select models from each subspace. If we evaluate one constituent model at a time, the remaining constituents of the ensemble serve as hint models to address new data distribution. Following our aforementioned hypothesis, if a candidate constituent agrees with the remainder of the ensemble, this is indicative of prediction stability. However, to study the Rashomon set for a given problem, we need to sample models from the complete model space (see Algorithm 1 for our ensemble learning approach).

Deriving an Ensemble: We assume a factorial combinatorial space encompassed by all feature combinations constrained to a single learning algorithm. To induce the Rashomon set, we aim to find a set of relevant features K that characterize an evaluated subspace. These features show complex correlations among a specific set of features and the target

```
Input: Set of available features $F$, train dataset $Z$, number of models to sample $n$,
       maximum model width $m$, and error margin $\epsilon$
Output: List of models constituting the ensemble

initialize pool $P$ with $n$ models containing random combinations of features from $F$
$H_{ref} \leftarrow$ choose a reference model to establish the Rashomon set
set $R$ as an empty list
for each $H_i \in P$ do
    if $\mathbb{E}[L(H_i, Z)] \leq \mathbb{E}[L(H_{ref}, Z)] + \epsilon$ then
        $R.insert((H_i, explanation(H_i))$
    end
end
cluster $R$ into $C$ given the explanation of each $H_i \in R$
find the $D$ clusteroids of $C$
set $E$ as an empty list
for each cluster $c \in C$ do
    $H_c \leftarrow$ the candidate model for expansion
    while $|H_c| \leq m$ do
        find the feature $f$ that minimizes $\mathbb{E}[L(\{H_c, f\} + \sum_{H_d}^{D-H_c}\{H_d\}, Z)]$
        assert $\{H_c, f\} \subset c$
        $H_c.insert(f)$
    end
    $E.insert(H_c)$
end
return $E$
```

Algorithm 1: Rashomon ensemble algorithm.

label, and the same correlations are not necessarily observed strongly in other regions of the data space, thus inducing a Rashomon subspace. The complete model space is characterized by models from size $s = 1$ to $|F|$, this being the set of all possible features. If we also consider the $\emptyset$ model to be a part of the complete model space, then there are ${}_FC_0 + {}_FC_1 + ... + {}_FC_F$ models. We limit our scope to problems where $|K| << |F|$, as otherwise it is unlikely that there exist multiple Rashomon subsets. If we sample an arbitrary model from the complete model space, the probability of this model not containing K is $({}_FC_K - 1)/{}_FC_K$.

Splitting the Rashomon Set: To split the Rashomon set into clusters, we represent how a model f' explains a phenomenon as a d-dimensional vector $S(f') = [e_1; e_2; ...; e_d]$ showing which features $[x_1, x_2, ...x_d]$ drive the model's prediction. We use K-Means clustering with a suitable number of clusters, identified by maximizing the silhouette value. This splits the Rashomon set into well-divided clusters based on their explanatory factors, leading to concise and distinguishable clusters.

Prediction Distance: We compare models within the Rashomon set to estimate the risk under an unknown distribution U. We compute the Jensen-Shannon distance (JSD) (Endres and Schindelin 2003) as our metric of choice for a measure of risk, indicating how similar the predictions of the two models are. Let P be the probability distributions returned from a model f_p, and we wish to compute a metric that estimates the risk of selecting it in production. Further, let Q be the probability distribution from a model f_q that ideally behaves similarly to f_p, and M be the mean of P and Q. The Jensen-Shanon distance can be computed as the mean Kullback–Leibler divergence (D_{KL}) between $D_{KL}(Q||M)$ and $D_{KL}(P||M)$.

Constituent search: Not all variables are relevant for prediction, and some features may even be detrimental. To find a set of relevant features to induce the Rashomon set, we represent the model space as a directed acyclic graph (DAG) in which each node represents a distinct feature subset, and vertex $A \to B$ is connected if B can be reached by simple feature addition from A, thus representing a transitive reduction of the more complex combinatorial complete model space (Zuin et al. 2021; Zuin et al. 2022a; Zuin et al. 2022b; Zuin et al. 2023a). This modeling approach presents two desirable properties: the first being that any vertex is reachable from the $[\emptyset]$ model, the second being that there exists a topological ordering, an ordering of all vertices into a sequence such that for every edge, the start vertex occurs earlier in the sequence than the ending vertex of the edge for any feature set path. These properties imply a partial ordering of the graph starting from the root node, which allows us to search it in an orderly manner. It has been shown that this modeling approach is effective for the task at hand (Zuin et al. 2021; Zuin et al. 2022a). This allows us to search the $F!$ combinatorial space.

5. Empirical Results

We assess the statistical significance of our measurements through a pairwise t-test with p-value ≤ 0.05 and 5-fold cross-validation. No hyperparameter tuning was performed in any of the algorithms employed, opting to keep their default values across all datasets. We evaluate the performance of both classical and state-of-the-art algorithms. In the presence of problems with many possible contrasting or competing explanations, employing the Rashomon sets as a method for obtaining ensemble constituents can be useful. Even in the absence of such structures, diversity is a desirable characteristic for any ensemble as it allows the end model to cover a wider region of the solution space. To support this statement and to verify whether Rashomon sets provide a suitable tool for model space partitioning, we propose splitting the Rashomon space into clusters, grouped by the explainability vectors of each model, and creating ensembles from the clusteroids. We compute the SHAP feature importance of each model and then run the K-Means algorithm to find a partition of the model space.

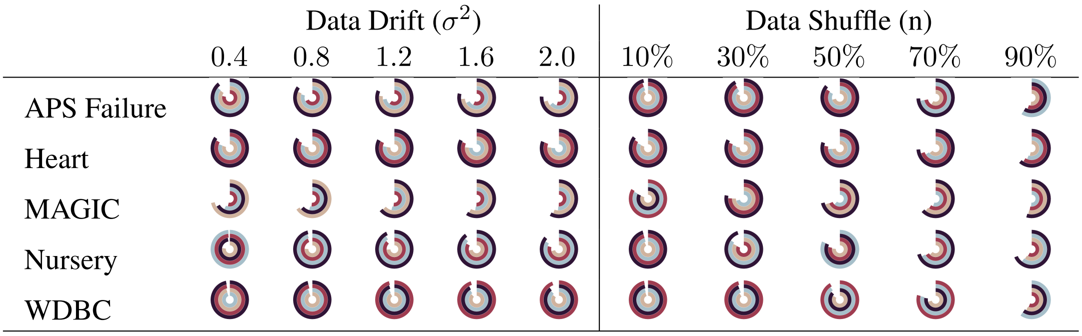
Open datasets: We include in our benchmark suite datasets from the UCI machine learning repository (Asuncion and Newman 2007) and the OpenML database (Bischl et al. 2017) on binary classification tasks. Table 1 summarizes a comparison between the proposed Rashomon ensemble and classic and state-of-the-art algorithms. For a fair comparison to the ensemble and boosting methods, we only employed decision trees as base constituents. In our experiments, we sampled $100,000$ decision trees to guarantee a minimum subset diversity and trained a meta-model to combine constituent outputs in a stacking ensemble. We verify that whenever the Rashomon ratio is relatively high ($\geq 5\%$), our proposed approach outperforms the alternatives.

To evaluate the robustness of Rashomon ensembles to distribution drift, we conducted experiments related to out-of-distribution data. We considered two scenarios: the addition of Gaussian noise and shuffling feature values to evaluate the reliance on core

Table 1. AUROC Benchmark suite results on binary classification tasks.

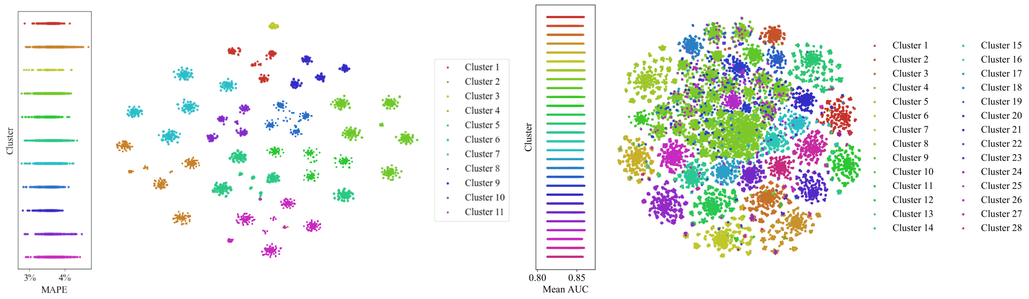
Benchmark			Baseline Algorithm						Rashomon	
Dataset	Instances	Features	DT	AdaBoost	Random Forest	XGB	LGBM	CatBoost	Ensemble	Ratio
APS	76000	172	.866	.824	.869	.835	.853	.888	**.911**	12.4%
Diabetes	101766	1691	.544	.614	.599	.615	.616	**.619**	**.618**	17.4%
Heart	303	171	.748	.787	.826	.796	.830	.834	**.839**	50.3%
MADELON	2000	502	.764	.598	.694	.828	.832	**.852**	*.746*	$< 0.5\%$
MAGIC	19020	102	.808	.830	**.857**	.837	.850	.850	.848	19.4%
Nursery	12630	784	.999	**.999**	**.999**	.991	**.999**	**.999**	**.999**	83.2%
Speeddating	8378	123	.650	**.673**	.630	.639	.642	.668	*.632*	$< 0.5\%$
WDBC	569	903	.949	**.973**	.967	.963	.967	**.974**	**.974**	21.5%
Wine	4898	13	.762	.722	.802	.755	.764	.782	**.805**	8.9%

Table 2. Performance of Random Forest ■, LightGBM ■, CatBoost ■ and Rashomon ensembles ■. Mean AUROC after 30 repetitions.



key features. In the first scenario, we added Gaussian noise with increasing σ^2 values to the datasets, mimicking shifts in the data distribution. We then evaluated the performance of Rashomon ensembles and other models under these perturbations. The results of this data drift scenario are summarized in Table 2, where each approach's performance is represented as a ring plot ordered by performance. The mean AUROC after 30 repetitions is provided as a measure of performance. In the second scenario, we shuffled the feature values within the datasets, disrupting the relationship between features and the target variable. We aimed to evaluate whether models could extrapolate from global information rather than relying on specific local patterns. The results can also be found in Table 2.

Unique collaborative datasets: To validate the effectiveness of our approach in real-world scenarios, we present the results of our evaluation across three distinct applications conducted in collaboration with various companies and institutions: stainless steel surface defects detection, COVID-19 hemogram detection from blood counts, and energy consumption forecasting. In all cases, new unique handcrafted datasets were created to explore each of the mentioned problems. Although these problems may seem vastly different, they share a common characteristic: the absence of a clear consensus among specialists on the best solution. Instead, they appear to exhibit multiple possible and effective solutions without a definitive optimal model or explanatory factors. This implies the presence of a large Rashomon set, fit for the application of our approach. Figure 1 showcases the Rashomon groups found in some of these studies. Further, as discussed in



((a)) Brazilian Energy.

((b)) COVID-19.

Figure 1. Example of Rashomon groups. We cannot observe a clear relationship between model performance and cluster assignment.

(Zuin 2023), our approach also enables detecting drift per instance and is easily adopted by domain experts, who can filter explanation groups to better align with the business.

The quality of duplex stainless steel is often threatened by the presence of surface defects. Slivers increase production costs as they remain undetected in intermediate processing stages, being observed only during the final inspection of the finished product. In partnership with *APERAM South America* we created a dataset containing the chemical compositions and metallurgical process variables of 122 duplex stainless steel production runs, from which 71 presented at least one defective plate. This corresponds to a dataset with nearly 500 stainless steel plates for studying the slivering problem, to which we applied our Rashomon ensemble technique and achieved a $.839$ AUROC in (Zuin et al. 2021). The task was formulated as a binary classification problem to predict which component combinations are likely to be associated with sliver formation. Once representative models were separated, we asked for insights from the metallurgical experts. The main lesson was that there were cases where some conclusions did not fit with realistic scenarios. After filtering those patterns, the most relevant ones were turned into production rules and employed in the 2019 and 2020 steelmaking process. A reduction of over 50% in the occurrence of heating slivers was reported, showing the potential of this strategy in real-world problems and validating the proposed framework.

In late 2019, the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) emerged in Wuhan, China, sparking a global outbreak in the subsequent weeks. Collaborating with *Grupo Fleury*, we amassed data from $900,220$ unique individuals, including $809,254$ Complete Blood Counts (CBC) and $1,088,385 RT - PCR$ exams. Among these, 21% $(234,466)$ tested positive for COVID-19, with fewer than 0.2% $(1,679)$ yielding inconclusive results. We also collected $120\,807$ CBCs performed between 2016 to 2019 of $16\,940$ individuals who tested positive for other respiratory viruses. Leveraging our Rashomon Ensemble technique, we predicted COVID-19 RT-PCR outcomes solely from CBC data, achieving an AUROC of 0.917 (Zuin et al. 2022a). Our method repurposed the readily available and cost-effective CBC test, enabling a fast and cheap preliminary diagnosis while accounting for potential confounding diseases. To the authors' knowledge, this study developed the most extensive COVID-19 dataset.

In our last case study, in collaboration with *Stanford University*, we developed a counterfactual model to identify the drivers of energy consumption in Brazil, culminating in multiple studies (Zuin et al. 2022b; Zuin et al. 2023a; Sun et al. 2023). Employing the Rashomon approach, we achieved a MAPE of 2.69 and an R^2 value of 0.848, also enabling quantitative assessments of extreme events' impacts, such as the COVID-19 pandemic, blackouts, and heatwaves. An unprecedented heatwave occurred in October 2020, breaking century-old temperature records. Our method detected anomalous climate data as early as May, showcasing its robustness and potential.

6. Conclusion and Final Remarks

In this study, we proposed a novel approach for ensemble learning based on explainability that enables estimating the prediction risk in production. We address the challenge of model selection by identifying a Rashomon subset of models that perform similarly but process data differently. By inducing perturbations on a held-out test set, we simulate out-of-distribution data and assess ensemble loss of predictive power as constituent models diverge. Our approach relies on ensemble diversity, leveraging that our constituent's behavior may diverge when faced with data from distributions that do not match the one seen in training. We demonstrate consistent gains in AUROC compared to other techniques in tasks where we verify the existence of multiple local structures in data. We validated our approach to real-world problems, achieving high performance in COVID-19 prediction and energy consumption forecasting. We also highlight the importance of expert inputs in refining the final model (Veloso et al. 2023), as demonstrated in a stainless steel case study which led to significant improvements in the production processes. Future work includes exploring ensembles with different algorithms and refining model selection methods for improved performance across more diverse datasets, as preliminarily explored in one of our most recent studies (Zuin et al. 2023b).

References

[Asuncion and Newman 2007] Asuncion, A. and Newman, D. (2007). Uci machine learning repository.

[Bischl et al. 2017] Bischl, B., Casalicchio, G., Feurer, M., Hutter, F., Lang, M., Mantovani, R. G., van Rijn, J. N., and Vanschoren, J. (2017). Openml benchmarking suites and the openml100. *stat*, 1050:11.

[Breiman 2001] Breiman, L. (2001). Statistical modeling: the two cultures. *Statistical Science. A Review Journal of the Institute of Mathematical Statistics*, 16(3):199–231.

[Chen et al. 2021] Chen, C., Yuan, J., Lu, Y., Liu, Y., Su, H., Yuan, S., and Liu, S. (2021). Oodanalyzer: Interactive analysis of out-of-distribution samples. *IEEE Trans. Vis. Comput. Graph.*, 27(7):3335–3349.

[Endres and Schindelin 2003] Endres, D. M. and Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7).

[Fisher et al. 2019] 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.*, 20:177:1–177:81.

[Hinns et al. 2021] Hinns, J., Fan, X., Liu, S., Raghava Reddy Kovvuri, V., Yalcin, M. O., and Roggenbach, M. (2021). An initial study of machine learning underspecification using feature attribution explainable ai algorithms: A covid-19 virus transmission case study. In *Pacific Rim International Conference on Artificial Intelligence*. Springer.

[Lundberg and Lee 2017] Lundberg, S. and Lee, S. (2017). A unified approach to interpreting model predictions. In *Annual Conf. on Neural Information Processing Systems*, pages 4768–4777.

[Madras et al. 2020] Madras, D., Atwood, J., and D'Amour, A. (2020). Detecting extrapolation with local ensembles. In *International Conference on Learning Representations*.

[Semenova and Rudin 2019] Semenova, L. and Rudin, C. (2019). A study in rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. *CoRR*, abs/1908.01755.

[Shmueli 2010] Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3):289–310.

[Sun et al. 2023] Sun, T., Zanocco, C. M., Zuin, G., Buechler, R., Flora, J. A., Galuppo, F., Soto, H., Veloso, A., and Rajagopal, R. (2023). Assessing climate change impacts on electricity consumption: Micro to macro perspectives from california's commercial sector to brazil's national grid. *AGU23*.

[Veloso et al. 2023] Veloso, A., Caramelli, P., Borges, K. B. G., Araújo, D., Zuin, G., Alves, T. H., and Ziviani, N. (2023). Processo centrado no humano para elaboração de modelos baseados em aprendizado de máquina e usos. Brazil patent BR 102021015411-0 A8.

[Zuin 2023] Zuin, G. (2023). *Ensemble Learning through Rashomon Sets*. PhD thesis, Universidade Federal de Minas Gerais.

[Zuin et al. 2022a] Zuin, G., Araujo, D., Ribeiro, V., Seiler, M. G., Prieto, W. H., Pintão, M. C., dos Santos Lazari, C., Granato, C. F. H., and Veloso, A. (2022a). Prediction of sars-cov-2-positivity from million-scale complete blood counts using machine learning. *Communications medicine*, 2(1):1–12.

[Zuin et al. 2022b] Zuin, G., Buechler, R., Sun, T., Zanocco, C., Castro, D., Veloso, A., and Rajagopal, R. (2022b). Revealing the impact of extreme events on electricity consumption in brazil: A data-driven counterfactual approach. In *2022 IEEE Power & Energy Society General Meeting (PESGM)*, pages 1–5.

[Zuin et al. 2023a] Zuin, G., Buechler, R., Sun, T., Zanocco, C., Galuppo, F., Veloso, A., and Rajagopal, R. (2023a). Extreme event counterfactual analysis of electricity consumption in brazil: Historical impacts and future outlook under climate change. *Energy*, page 128101.

[Zuin et al. 2020] Zuin, G., Chaimowicz, L., and Veloso, A. (2020). Deep learning techniques for explainable resource scales in collectible card games. *IEEE Transactions on Games*, pages 1–1.

[Zuin et al. 2021] Zuin, G., Marcelino, F., Borges, L., Couto, J., Jorge, V., Laurindo, M., Barcelos, G., Cunha, M., Alvarenga, V., Rodrigues, H., et al. (2021). Predicting heating sliver in duplex stainless steels manufacturing through rashomon sets. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

[Zuin et al. 2023b] Zuin, G., Parreiras, L., Melo, L., Barros, G., Lomeu, H., Melo, B., Marini, W., Lott, D., and De Souza, M. (2023b). An ensemble approach for inconsistency detection in medical bills: A case study. In *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 573–578. IEEE.

[Zuin and Veloso 2019] Zuin, G. and Veloso, A. (2019). Learning a resource scale for collectible card games. In *2019 IEEE Conference on Games (CoG)*, pages 1–8.

[Zuin et al. 2020] Zuin, G., Veloso, A., Portinari, J. C., and Ziviani, N. (2020). Automatic tag recommendation for painting artworks using diachronic descriptions. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.