"A 6 or a 9?": Ensemble Learning Through the Multiplicity of Performant Models and Explanations

GIANLUCCA ZUIN, Universidade Federal de Minas Gerais, Brazil and Instituto Kunumi, Brazil
ADRIANO VELOSO, Universidade Federal de Minas Gerais, Brazil and Instituto Kunumi, Brazil

Creating models from past observations and ensuring their effectiveness on new data is the essence of machine learning. However, selecting models that generalize well remains a challenging task. Related to this topic, the Rashomon Effect refers to cases where multiple models perform similarly well for a given learning problem. This often occurs in real-world scenarios, like the manufacturing process or medical diagnosis, where diverse patterns in data lead to multiple high-performing solutions. We propose the Rashomon Ensemble, a method that strategically selects models from these diverse high-performing solutions to improve generalization. By grouping models based on both their performance and explanations, we construct ensembles that maximize diversity while maintaining predictive accuracy. This selection ensures that each model covers a distinct region of the solution space, making the ensemble more robust to distribution shifts and variations in unseen data. We validate our approach on both open and proprietary collaborative real-world datasets, demonstrating up to 0.20+ AUROC improvements in scenarios where the Rashomon ratio is large. Additionally, we demonstrate tangible benefits for businesses in various real-world applications, highlighting the robustness, practicality, and effectiveness of our approach.

CCS Concepts: • **Computing methodologies** → *Classification and regression trees*; **Ensemble methods**; • **Human-centered computing**;

Additional Key Words and Phrases: Rashomon Effect, Ensemble Learning, Explainability

1 INTRODUCTION

Model selection is crucial in industry and research, and the widely adopted approach is cross-validation. Although cross-validation generally provides robust risk estimation [4], it can fail for specific problems depending on the goal of model selection, and empirical risk in a test set might not always correlate with real-world performance [24]. Empirical risk can be significantly affected when different models perform equally well on the test set [55]. This limitation of relying solely on empirical risk motivates us to explore alternative evaluation approaches that capture subtle differences in model behavior.

The Rashomon Effect, also known as the multiplicity of good models [16], presents a phenomenon where many models perform equally well. However, these models may process data in substantially different ways, making it challenging to draw reliable conclusions or automate decisions based on a single model fit [129]. This inherent diversity in model behavior emphasizes why a single performance metric can be misleading. A significant challenge arises when a cross-validated model, carefully selected during training, encounters data drawn from a different distribution during production. In these cases, even small internal differences among models may lead to divergent outcomes. Cross-validation guarantees no longer apply to out-of-distribution data, resulting in unpredictable model performance and rendering held-out performance an unreliable risk estimate.

Authors' Contact Information: Gianlucca Zuin, gianlucca@kunumi.com, Universidade Federal de Minas Gerais, Brazil and Instituto Kunumi, Brazil; Adriano Veloso, adriano@kunumi.com, Universidade Federal de Minas Gerais, Brazil and Instituto Kunumi, Brazil.

ACM 1556-472X/2025/9-ART
https://doi.org/10.1145/3767735

To address this issue, we extend our analysis beyond empirical risk and explore additional axes to identify more robust models. In doing so, we not only measure performance but also examine how models adapt their behavior across varying data distributions.

Our main hypothesis posits that models exhibit similar behavior only when data are drawn from the same distribution as seen during training. If this holds, we can sample different models and verify their outputs during production (i.e., once deployed and used on new data). Disagreement among models would imply that the data are drawn from an unknown distribution, leading to untrustworthy predictions. Conversely, if the models agree, we gain extra confidence in the correctness of their predictions, as diverse models converge on the same conclusion. These models can be used to build an ensemble, and we may establish the Jensen-Shannon distance between the outputs of the ensemble constituents as a production risk metric.

However, we believe that diversity among individual models is still crucial for gaining an understanding of the data. The Rashomon Effect suggests that multiple explanations can exist for a given phenomenon, each consistent with the observed data. To systematically capture this diversity, we can group models based on the similarity of their explanations. Ideally, this leads to dense groups of models that share common factors, from which the most distinct representatives are selected. The traditional approaches attempting to induce a single model from all sub-populations may perform poorly in complex problems where data inherently contain several local structures and sub-populations [8, 70, 72, 128]. Our strategy instead culminates in an ensemble that is both diverse and robust, with all its constituents being performant models focused on a subset of features, a local structure in the data.

Further, as each constituent offers a different explanation for the target phenomenon, the ensemble's output is directly linked to the trustworthiness of predictions. The consensus among the constituents indicates a match between the data distribution and the one seen during training, with all the cross-validation guarantees. Disagreement suggests that cross-validation properties may not be trusted. In our experiments, we consider a simple voting scheme ensemble and the returned voting ratio, that is, the probability predicted by our ensemble, as a measure of agreement. For binary classification problems, such as those evaluated in our experiments, ratios close to 50% indicate the highest uncertainty. We coined this concept the Rashomon Ensemble. In summary, our approach involves the following steps:

(1) Sampling models from a pre-defined Rashomon subspace (i.e., a set of models with equivalent performance), achieved by drawing different random feature subsets.
(2) Computing the explanation for each sampled model and quantifying the pairwise similarities among them.
(3) Perturbing held-out test data through appropriate transformations to evaluate model stability.
(4) Measuring the pairwise distances on the perturbed dataset to capture divergence in predictions.
(5) Segmenting the Rashomon Set into subgroups based on the models' explanation vectors and distance metrics.
(6) Selecting a set of models with contrasting explanations and divergent responses on the perturbed data.
(7) Constructing an ensemble and evaluating the degree of agreement among models as a proxy for production risk.

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. Further, we also employ the Rashomon ensembles in four real-world applications partnered with various industries and institutions, studying the impact of our approach. We demonstrate how our approach leads to a tangible impact on business, with reported gains of over *R*$1.5 million and a patent being filed. Our results also consistently show

that Rashomon ensembles outperform state-of-the-art ensemble learning approaches when the Rashomon Set is sufficiently large.

The remainder of this work is structured as follows: Section 2 introduces background to support the understanding of the devised techniques, explaining concepts such as the Rashomon Effect, Rashomon Ratio, and our understanding of model explanation. Section 3 discusses related research, including data drift, domain adaptation, ensemble learning, and prior applications of the Rashomon Effect in machine learning. Section 4 formalizes our proposed Rashomon Ensemble, outlining its theoretical framework and implementation. Sections 5 and 6 present experimental evaluations. Section 5 addresses the analysis of our approach's robustness and limitations under publicly available datasets, while Section 6 focuses on multiple real-world case studies using unique collaborative data. Finally, Section 7 brings about the conclusion and proposes directions for future research.

2 BACKGROUND

The idea of capturing model uncertainty by exploring the relationship between test points and the learned model is not new. Typical approaches include building an ensemble of models and measuring inter-model variance [77], or learning a scoring rule that captures ambiguity in targets [66, 67]. However, most recent research on this topic has mainly focused on Neural Networks and how they learn intermediate features. More specifically, state-of-the-art approaches to Out-of-Distribution (OoD) detection enrich the intermediate feature space beyond what would ordinarily be learned via supervised learning alone, such as encouraging a model to learn high-level task-agnostic semantic features [114], or employing an additionally labeled outlier dataset during training [53]. When one cannot access the intermediate feature space, most of the mentioned approaches fail. As noted by Chen et al. [20], this type of approach has two drawbacks: first, models trained to identify OoD may fail to cover the entire data distribution; second, explaining the source of OoD may be non-trivial.

The key difference in our work lies in the analysis of additional unexplored axes, such as the decision-making process of a model via its explanatory factors [75, 125]. A second key idea is to exploit the Rashomon Effect by looking for models with similar performance during training. Both of these propositions enable explanation of the risk metric by assigning importance to the factors leading to each model decision and comparing them. Further, our approach is algorithm-agnostic and reproducible with any model that handles tabular data. We therefore summarize three pivotal points underlying our approach: acknowledging that production data may fall outside the training distribution, recognizing the multiplicity of high-performing models, and analyzing the explanatory factors behind model decisions

Underspecification: Underspecification in deep learning arises when models achieve similar in-sample performance but present divergent behaviors on out-of-sample data. This is problematic when some models perform significantly worse in production, creating challenges for proper model selection [24]. Although much of the underspecification literature focuses on deep neural networks, the phenomenon largely arises from the elevated number of optimized parameters [18, 91]. Mei and Montanari [83] state that this issue is common to any machine learning pipeline. D'Amour et al. observed that repeating a training process can generate many models with identical test performance but significantly different behaviors, even when only minor perturbations are introduced, such as using a different random seed. This, in turn, differentiates each model by small arbitrary learning decisions. Although these differences are usually considered minor, the consequence is varying degrees of performance in the real world. As such, underspecification is closely tied to the Rashomon Effect.

Rashomon Effect: The Rashomon Effect, as analyzed by Fisher et al. [36], refers to the set of models with accuracy close to that of the optimal model. From this set, they formally defined the concept of the *Rashomon Set*, which represents the subspace of the universe of models summarizing the range of effective prediction strategies an optimal analyst might choose. The Rashomon Effect is further explored by Semenova et al. [100], who provide

pertinent definitions concerning the generalization of the Rashomon Set, its shape, and its volume. In particular, they explore under which situations it is possible to obtain a sample of the model space such that the Rashomon properties of this subspace are similar to those of the full universe.

The Rashomon Set thus comprises a collection of close-to-optimal models that share similar explanations and performance due to the Rashomon Effect. For this, we need a comparison to some key reference model, denoted as h_{ref}. Fisher suggests that h_{ref} can be derived from expert knowledge or from some quantitative decision rule implemented in practice. This prespecified reference model serves as a baseline for performance. Thus, if we establish ϵ as the maximum accepted error relative to h_{ref} for considering a model part of the Rashomon Set, we can denote it as:

$$R(H, \epsilon) := \{h \in H : E[L(h, Z)] \leq E[L(h_{ref}, Z)] + \epsilon\} \tag{1}$$

where E denotes expectations concerning the population distribution, L is some nonnegative loss function, H is the hypothesis space, and $Z = [YX]$ is the data. The ϵ metric takes into account models that might be arrived at due to differences in data measurement, processing, filtering, model parameterization, covariate selection, or other analysis choices.

Further, let $X_1 \in X$ be the set of all features that model h_{ref} relies on to reach a prediction. This reliance metric has a direct relationship with model explanation. We can expect models that rely too heavily on X_1 to be prone to high variance, leading to poor performance. Likewise, models that rely too little on X_1 are prone to high bias, also leading to poor performance. Model reliance (MR) of variable X_1 can be computed as the increase in expected loss when the contribution of this variable is removed by random permutation. The range of all possible MR values within this class gives rise to the notion of Model Class Reliance (MCR), which helps define a minimum and maximum MR value to classify a model f within the set defined by h_{ref}, ϵ, and X_1. These models comprise the set of high-performing models that also share a similar reliance on X_1 as the reference model h_{ref}, as proposed by Fisher et al.

Rashomon Ratio: The concept of the Rashomon ratio, as introduced by Semenova and Rudin [100], quantifies the fraction of models within the hypothesis space that perform nearly as well as a reference model. Given a hypothesis space H and a subset $R \subseteq H$ of good models (the Rashomon Set), the ratio is defined as:

$$R_{ratio} = \frac{|R|}{|H|} \tag{2}$$

Here, the notation $|\cdot|$ is used to denote the size of a set, and its interpretation depends on the hypothesis space shape. For discrete spaces, that is, when H is discrete and finite, $|R|$ and $|H|$ represent the cardinalities of the Rashomon Set and the full hypothesis space, respectively:

$$|R| = \sum_{h \in H} \mathbf{1}\{h \in R\}, \quad |H| = \sum_{h \in H} 1$$

in which $\mathbf{1}\{\cdot\}$ is the indicator function.

For continuous hypothesis spaces, that is, in scenarios where H is continuous or infinite, $|R|$ and $|H|$ can be interpreted as volumes under a chosen measure $\mathcal{V}(\cdot)$, such that:

$$|R| \equiv \mathcal{V}(R) \quad \text{and} \quad |H| \equiv \mathcal{V}(H)$$

in which the ratio $\frac{\mathcal{V}(R)}{\mathcal{V}(H)}$ is well-defined under the assumption of a uniform prior over H.

However, even for discrete hypothesis spaces, the exact computation of R_{ratio} would involve evaluating every model $f \in H$, which may be computationally infeasible. Thus, we approximate the Rashomon ratio by random sampling: models are drawn from H and the set, the fraction of sampled models, that lie in $\hat{R}$ serve as an empirical

estimate $\hat{R}_{ratio}$ of this discrete hypothesis subspace. If the sample is large enough, this estimate holds guarantees of similarity to the true Rashomon Set, as stated by Fisher et al. [36].

Data drift: Let T be the train distribution of the source data, and U be some unknown distribution from another dataset. Candela et al. [19] defines data drift as a change in the joint distribution of features. That is:

$$P(x_t, y_t) \neq P(x_u, y_u) \tag{3}$$

Probably approximately correct learning relies on the assumption that data are independently and identically distributed to estimate the empirical risk of a learning function. If we observe data drift, we cannot guarantee that the empirical risk is close to the true risk.

We can decompose $P(x, y) = P(x) \times P(y|x)$. Thus, if data drift occurs, it may stem from two sources: a change in $P(x)$ (covariate drift) or a change in $P(y|x)$ (concept drift). As stated by Moreno-Torres et al. [85], covariate drift is tied to the distribution of the variables, while concept drift implies that the relationship between the target and predictors changes between datasets. Finally, both $P(x)$ and $P(y|x)$ may differ significantly from the original distributions, which is defined as dual drift. Overall, data drift can be stated as a phenomenon in which the statistical properties of a target domain change over time in an arbitrary way [74].

In this work, we address data drift by building ensembles composed of models trained on independent feature subsets. We propose that if a small portion of features suffers from drift, accurate predictions can still be obtained from the unaffected features. Further, if the distributions T and U are completely different, we must be able to signal the low reliability of the prediction. Finally, to ensure diversity in constituent models, we rely on the notion of explainability.

Model Explanation: Instead of the model reliance metric proposed by Fisher et al., another possible approach is presented by Lundberg and Lee [75]. Shapley Additive Explanations (or simply SHAP) use Shapley values to interpret a prediction model. We represent how model f' explains the data as a d-dimensional vector $S(f') = s_1, s_2, \ldots, s_d$, showing which features contribute most to the prediction. The Shapley value is a concept in cooperative game theory introduced by Shapley [101]. In each game, a unique distribution of the rewards generated by the cooperation of all players is provided. Many other feature attribution methods exist [17, 96, 98], but as highlighted by Hinns et al. [55], the sound mathematical foundation and ease of implementation make SHAP ideal for identifying underspecification. Further, SHAP is the only method with the three desirable properties:

- Local accuracy: the explanations truthfully explain the model.
- Missingness: missing features have no attributed impact on the decisions.
- Consistency: if a model changes so that some feature's contribution increases or stays the same regardless of the other features, that feature's attribution should not decrease.

In summary, the Rashomon Set reveals the existence of multiple valid model solutions with similar performance. Their explanations allow us to examine each model's decision rationale and understand the importance of different features, highlighting what makes each model in the Rashomon Set distinct. Further, it enables us to discern the unique aspects of each model that contribute to varied performance and responses under data drift. By combining these approaches, we gain a broader understanding of the problem and build ensembles of diverse models that provide complementary explanations for different facets of the data. This ensemble enhances the robustness of our solution, as each model's behavior under varying conditions is better understood and accounted for.

3 RELATED WORK

Data drift is usually associated with the notion of online learning, in which a model is applied to production and is constantly updated as new instances arrive. Under online learning, a model must handle new concepts as they arrive, properly tuning itself to new data distributions. The main challenge consists of the fact that, as data

drifts toward these new concepts, it negatively impacts the accuracy of the models that are learned based on past training instances [45]. Therefore, early identification and adaptation to data drift are key aspects of such systems. Lu et al. [73] provides a basic framework underlying general drift detection:

- Stage 1 (Data Retrieval): retrieval of chunks from data streams to infer data distribution.
- Stage 2 (Data Modeling): extraction of key features that present the most impact on the system in the presence of drift.
- Stage 3 (Test Statistics Calculation): the measurement of a dissimilarity or distance metric.
- Stage 4 (Hypothesis Test): evaluation of the statistical significance of the measured metric.

The main differences between methods lie in stages 3 and 4. Concerning stage 3, two of the main categories of drift identification are error-based and data-based algorithms. Most error-based drift detection employs a base classifier and tracks the change in the online error rate. The main hypothesis behind these methods relies on the fact that the base model will misclassify new instances when data drifts, thus increasing the error rate. This is the core idea behind the DDD of Minku and Yao [84]. There are many other error-based methods, but, as stated by Lu et al. [73], DDD is perhaps the most referenced method. Under their framework, other methods can be summarized by changes to some stage of the drift detection, such as employing another hypothesis testing [41] or changing some detail of the evaluated metric [7].

Data-based drift detection algorithms rely on directly quantifying the dissimilarity between the distribution of historical and new data. The standard strategy is to define a fixed window for the past and a sliding window for new data during the online learning process [62]. If we ignore Stage 1 of the drift detection framework, the problem turns into a multivariate two-sample test evaluating if samples come from the same distribution. However, there remains a problem concerning actual and virtual drift.

The decomposition of Equation 3 presents the sources of data drift: covariate drift and concept drift. Covariate drift is often called virtual drift due to drift in $P(x)$ not affecting the decision boundary of models [95]. Retraining a model under covariate drift might not be necessary, as the learned conditional $P(y|x)$ remains unchanged. This is not the case for dual drift, however, when both $P(x)$ and $P(y|x)$ exhibit a shift under new data. It is important to highlight that the aforementioned approaches to drift detection are well-suited for online learning scenarios, which is not the case for our proposed problem. We can only compute error-based metrics if we know the correct label of new incoming instances. Sliding window data-based methods depend on the notion of temporal relationships. Further, knowledge of the labels of novel instances is necessary to differentiate between dual and virtual drift, which might not be possible in scenarios outside of online learning. We propose building an ensemble of models and using the intra-constituent agreement as a proxy for error rate, as described in Section 4.

Existing research has explored methods to address the challenge of data distribution shifts. Domain adaptation, for instance, aims at mitigating performance degradation when a model trained on a source domain is applied to a different but related target domain. Farahani et al. categorize these approaches into shallow and deep methods, emphasizing strategies such as feature alignment, instance re-weighting, and adversarial training for settings where only source labels are available [35]. These methods often focus on aligning feature distributions or adapting the model parameters to the target data [71, 92, 108].

Another related approach that tackles this problem is domain generalization, in which one seeks to train models that can generalize well to unseen target domains without access to this target data during training [86]. The current approaches can be organized into categories such as data manipulation, representation learning, and learning strategies [**?**]. For example, Mixup-based augmentation [118] enhances diversity by modifying training data through linear interpolation, while domain-adversarial training [42] explicitly aligns distributions across domains by adversarial optimization.

A third approach consists of test-time adaptation methods. Unlike domain adaptation or generalization, which operate during training, this approach adapts pre-trained models directly to unlabeled test data in real time [68]

and, as highlighted by Liang et al., they broadly fall into three categories: (i) test-time domain adaptation, which leverages pseudo-labeling or clustering to align entire target domains with source knowledge [69, 94]; (ii) test-time batch adaptation, which adjusts normalization statistics or fine-tunes parameters on small batches [99, 119]; and (iii) online test-time adaptation, which incrementally updates models on streaming data while attempting to mitigate catastrophic forgetting [88, 113]. Other approaches also include entropy minimization [112], contrastive consistency [21], or memory-augmented prototypes [31]. However, these methods assume that models must be adapted to the target distribution, requiring access to source model parameters and computational resources for updates. These may be prohibitively limiting in scenarios where model stability, interpretability, or deployment efficiency are crucial.

While these existing methods, be they domain or test-time, focus on adapting models to shifted data, our approach leverages the inherent diversity of high-performing models in the Rashomon Set. Instead of modifying parameters or normalizing statistics, we detect data drift through disagreement among ensemble constituents. Models in the Rashomon Set, though equally accurate on training data, rely on distinct features and hold different decision boundaries. Significant prediction divergence on new data hints at distribution shifts. This process requires no model updates, source data access, or computational overhead after deployment, providing a lightweight, proactive indicator for safety-critical or resource-limited settings, as is the case for many real-world applications.

One of the core motivations in this work arises from the insight that a dataset might be heterogeneous, thus inducing a large Rashomon Set. There might exist regions of the data that show complex correlations among a specific set of features and the target label, and the same correlations are not necessarily so strongly observed in other regions. If this is true, it would be more suitable if local behavior were represented by a local model, which can be incorporated into an ensemble [126]. Sampling multiple local minima allows approximation of the global objective while expanding the representation space [30]. This idea of employing local models aligns with 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, we can identify subgroups of correlated features and build ensembles with diverse models that contribute unique explanations for different facets of the data [122]. Such an approach enhances robustness by leveraging the multiplicity of high-performing models with diverse decision-making processes.

Dembczyński et al. [28] focuses on understanding how one can learn a performant rule-based ensemble via boosting. Starting from the standard initial rule, they iteratively add new rules to obtain an ensemble that can cover most of the data. To validate their approach, they also define the concept of coverage through a $\phi(x)$, this being an arbitrary axis-parallel region in the attribute space. The diversity of constituents is measured solely by the coverage ϕ of each rule. As noted by D'Amour et al. [24], two rules may have the same coverage but exhibit divergent behavior in practice. Thus, using some other metric associated with the inner mechanism of the model and not simply the observed response may be relevant, such as a vector representation of the explainability of a model.

Grosskreutz [46] propose splitting dataset rows into subgroups given a set of restrictions over its columns, and apply this approach to an unsupervised problem. If the groups are large enough, the associated restrictions express some significant pattern in the data. Grosskreutz focuses on tasks where there is no target variable. However, one can employ an equivalent technique regardless of this fact, similar to Malik and Kender [79] and Knobbe and Valkonet [65]. All these works operate primarily within the data space, looking for relevant patterns, clusters, or subgroups that induce diverse models. Our approach, in contrast, operates within the model space, finding different groups of explanations. The Rashomon groups can be interpreted as a particular set of restrictions on the data, which in turn induce the subgroups presented. We improve upon previous work in the sense that the SHAP groupings aided by the Rashomon concept not only prune a large portion of the search

space but also provide a direct measure of model behavior similarity while tackling the problem of data drift detection in domains outside of online learning.

Regarding the Rashomon Effect, many works have exploited its implications to gain insights about the solution space. Marx et al. [82] explores the concept of predictive multiplicity, the ability of a prediction problem to admit competing models with conflicting predictions, which can be seen as a restriction on the Rashomon Set. Kissel and Mentch [64] searches for an entire collection of plausible models via a forward selection approach and resampling of the training dataset to account for uncertainty. Dong and Rudin [32] introduces the notion of a variable importance cloud, mapping every variable to its importance for the Rashomon Set, and experimenting on criminal justice, marketing data, and image classification tasks. [87] performs a similar approach using Shapley values as a measure of importance. There is also relevant literature regarding Rashomon Sets and a specific learning algorithm of choice. For instance, Ahanor et al. [3] and Danna et al. [25] both look for the set of near-optimal solutions for integer linear programs, while Xin et al. [117] restricts their analysis of the Rashomon Set to Decision Trees. To the best of the authors' knowledge, building an ensemble from the Rashomon set is a novel idea.

4 METHOD

We consider a supervised learning scenario and formulate a classification model as a function $h(X, Y; \theta)$ 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 $h_n \in H$ 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 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 executions are only minor. We consider θ to encompass any choices made during training that lead to similar models exhibiting contrasting performances. We then introduce drift to the test data and evaluate its effects on each model.

Instead of simply mixing different structures into a single model and minimizing the objective function $h(x)$, we sample the model space by minimizing different functions $h(x')$, where $x' \subseteq x$ and $|x'| < |x|$, as in [128]. This sampling strategy approximates the Rashomon Set, acknowledging 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, 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 well-performing models with varying decision-making processes.

We build our ensemble exploiting two concepts: diversity between individual models and stability between model explanation and empirical predictions [102]. Diversity is crucial for gaining a general understanding of a phenomenon, assuming that problems are not tied to a single cause, which may vary in ways that are not directly intuitive. To promote diversity while finding patterns, we cluster the set of sampled models H' based on the distance between their explanation vectors (i.e., SHAP values). In our experiments, we employ Euclidean distance and k-means clustering, though alternative metrics and clustering methods could be applied. Ideally, this creates 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 cluster the model space based on the distance between the explanation vector associated with each model and project them into the prediction space. This allows us to locate different Rashomon subgroups inside the Rashomon Set and select models from each subspace. In practice, when we evaluate one constituent at a time, the remaining members of the ensemble act as reference models to verify consistency under new data distributions. If a candidate model's predictions agree with the remainder of the ensemble, it is indicative of prediction stability. Further, we also require a stability check to ensure that, when searching for optimal constituents, adding or removing features does not significantly alter the model's explanation vector relative to its cluster centroid and neighborhood. Finally, to study the Rashomon Set for a given problem, we need to sample models from the complete model space. Algorithm 1 describes the main steps of our ensemble learning approach. Here each model h is represented by the unique feature set that it employs. As such, we overload the notation h to denote both the model and its feature set, using them interchangeably in the algorithm description.

Algorithm 1: Rashomon Ensemble Algorithm

Input: Feature set F, evaluation dataset Z, number of initial models n, maximum model width m, error margin ϵ
Output: Ensemble of models M
Initialize pool P with n models, each built using a random subset of features from F.
Choose a reference model h_{ref} (e.g., a baseline method).
Initialize an empty Rashomon Set R.
for *each model* $h_i \in P$ **do**
  Evaluate h_i on Z.
  if $E[L(h_i, Z)] \leq E[L(h_{ref}, Z)] + \epsilon$ **then**
    Compute explanation vector $\mathrm{S}(h_i)$ (e.g., using SHAP).
    Add h_i along with $\mathrm{S}(h_i)$ to R.
Cluster the models in R based on the distance between their explanation vectors, forming clusters C.
For each cluster $c \in C$, select a representative model (e.g., the clusteroid) h_c.
Initialize the ensemble M with these representative models.
for *each cluster* $c \in C$ **do**
  Let h_c be the representative model for cluster c.
  while $|h_c| < m$ **do**
    Identify the feature $f \in F \setminus h_c$ that minimizes

$$E\left[L\left((M \setminus \{h_c\}) \cup \{h_c \cup \{f\}\}, Z\right)\right]$$

    while ensuring that $h_c \cup \{f\}$ remains consistent with the explanatory profile of cluster c (e.g., does not fall into another cluster).
    Update $h_c \leftarrow h_c \cup \{f\}$.
return M

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 (with size $|K|$) that characterize an evaluated subspace. These features show complex correlations with the target label, which may not appear as strongly in other regions of the data space, thus inducing a Rashomon subspace.

To mitigate the curse of dimensionality, we restrict the model space to subsets of size s where $|K| \leq s \leq S_{\max}$, with $S_{\max} \ll |F|$. This avoids the computational intractability of enumerating all $2^{|F|}$ subsets. The number of models containing the subset K is derived by fixing $|K|$ features and choosing the remaining $s - |K|$ from $|F| - |K|$:

$$\sum_{s=|K|}^{S_{\max}} \binom{|F| - |K|}{s - |K|} \tag{4}$$

The probability of sampling a model containing K is:

$$P_K = \frac{\sum_{s=|K|}^{S_{\max}} \binom{|F|-|K|}{s-|K|}}{\sum_{s=|K|}^{S_{\max}} \binom{|F|}{s}} \tag{5}$$

We limit our scope to problems where $|K| \ll |F|$, as otherwise, the Rashomon ratio diminishes due to the combinatorial scarcity of subspaces containing K. To ensure computational tractability, we restrict models to sizes s where $|K| \leq s \leq S_{\max}$, avoiding the curse of dimensionality inherent in high-dimensional feature spaces. If we sample an arbitrary model from this constrained space, the probability of it *not* containing K is $1 - P_K$. From Equation 5, to guarantee that the subspace K is present in at least one model with probability α, we need to sample at least η models:

$$\eta = \frac{\ln(1 - \alpha)}{\ln(1 - P_K)} \tag{6}$$

For example, for $|F| = 100$, $|K| = 4$, $S_{\max} = 10$, and $\alpha = 0.95$:

$$\eta \approx \frac{\ln(1 - 0.95)}{\ln(1 - 0.00005)} \approx 60{,}000$$

Time Complexity: In the experiments, we use Decision Trees as base models for the ensemble constituents. The time complexity of training and explaining a Decision Tree is $O(\log(F)ID + D^2)$ [75], where I is the number of instances, F is the number of features, and D is the maximum tree depth. Since we sample T trees, other steps present negligible complexity in comparison to the sampling stage, resulting in $O(TI)$ complexity. However, as we sample models, the Rashomon ensemble substantially reduces the number of models that need to be evaluated, making the approach feasible in practice. For instance, any model with a loss close to random guessing is unlikely to present itself as a useful constituent. Thus, we do not need to explain the entire model space, and need only concern ourselves with the Rashomon Set.

Splitting the Rashomon Set: To split the Rashomon Set into clusters, we represent how a model h' explains a phenomenon as a d-dimensional vector $S(h') = [e_1; e_2; ...; e_d]$ 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 compact and well-separated clusters. As discussed previously, there usually exists a small subset of key features that are only present in models from one cluster and absent in the remaining ones. The presence of this subset leads to these models being close in the feature preference space since cohesion values are relatively high and lead to concise and well-divided 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) [34] 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 h_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 h_p. As shown by MacKay [76], we can compute the error of P from Q by the cross-entropy between P and Q as:

$$H(P, Q) = H(P) + D_{KL}(P||Q) \tag{7}$$

We could also evaluate the Kullback-Leibler (KL) divergence between these models under the unknown distribution U, and thus estimate risk. The main drawback of employing the Kullback-Leibler divergence is that it is non-symmetric. That is, $D_{KL}(P, Q)$ might be different from $D_{KL}(Q, P)$. To avoid confusion, we instead opt to employ the Jensen-Shannon distance as our metric of choice for a measure of risk. If P and Q agree (low JSD), we have a strong indicator that U should be similar to T, and predictions can be trusted. Contrasting P and Q (high JSD) suggests that U differs from T, and the returned predictions cannot be trusted.

Constituent search: In summary, we verify that looking at the explanatory factors in isolation is not enough to observe meaningful patterns. In our preliminary experiments, we find instances of models with similar SHAP but contrasting predictions as well as contrasting SHAP but similar predictions. The choice of a h_q model to estimate the risk of the target constituents becomes a challenging task. We propose performing a controlled transformation in T to create a simulated production dataset. This should enable us to estimate model behavior in an out-of-distribution scenario. Namely, the transformation employed over data drawn from T consists of adding Gaussian noise to the input features such that $y_i = h(x_i + \epsilon_i)$ and $\epsilon_i \sim N(0, \sigma^2)$. We can then select models that have contrasting explanations and predictions. Among possible transformations, we choose Gaussian noise for its simplicity and the ease of computing exact feature distortion. In many real-world scenarios, Gaussian noise might not be the closest representative of divergence. However, we verify that this simple transformation is enough to induce large changes in model behavior and enable our ensemble learning approach.

Further, 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 $A \rightarrow B$ 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. This modeling approach has two desirable properties: (i) any vertex is reachable from the $[\emptyset]$ model, and (ii) a topological ordering exists 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 [122, 126].

We can, for example, apply the A* algorithm [51], employing as a heuristic the performance of the model represented by the feature set of a given vertex and the Jensen-Shannon distance to the predictions of the remaining Rashomon subgroup clusteroids. We hypothesize that there exists a set of optimal feature expansions that lead to the best-performing models for each specific base task. This allows us to search the $F!$ combinatorial space of feature subsets to select the best-performing specialized models and build the Rashomon ensemble.

5 OPEN DATASETS

We present our experiments related to the Rashomon Set for a given problem and the process of obtaining ensemble constituents using the Rashomon Sets. The goal is to explore the usefulness of Rashomon Sets as a method for model space partitioning and to understand their effectiveness in addressing the problem akin to underspecification in ensembles. To study the Rashomon Set for a given problem, we sampled models from the complete model space. We considered the $N!$ combinatorial space, encompassing all feature combinations constrained to a single learning algorithm. We aimed to evaluate whether Rashomon Sets could serve as a valuable tool for partitioning the model space and generating diverse ensembles.

To achieve this, we proposed a process of Rashomon Set partitioning based on clustering models by their explainability vectors. The K-Means algorithm was used to induce clusters, and we determined the optimal number of clusters (K) using silhouette scores. By creating ensembles composed solely of models located close to the centers of each Rashomon subgroup, we aimed to generate diverse ensembles capable of covering a wider region of the solution space.

5.1 Benchmark Suite and Datasets

To verify the effectiveness of the Rashomon ensemble learning technique, we considered a benchmark suite including a series of open-source datasets from the UCI machine learning repository [5] and the OpenML database [12]. The benchmark suite consists of the following datasets:

APS Failure: The dataset used for the 2016 IDA Industrial Challenge [23]. It consists of data collected from heavy Scania trucks in everyday usage, and the problem is formulated as a binary classification task to predict component failures for a specific component of the APS system after a small amount of noise was introduced to the data.

Diabetes Readmission: This dataset was submitted on behalf of the Center for Clinical and Translational Research, Virginia Commonwealth University [106]. It represents 10 years of clinical care at 130 US hospitals and integrated delivery networks. The problem is a binary classification task to predict whether a given patient will be readmitted to a hospital.

Heart Disease: This dataset from the Cleveland database focuses on the diagnosis of coronary artery disease [2]. The goal is to predict the presence of heart disease in the patient, with a severity indicator valued from 0 (no presence) to 4. We have focused on the binary counterpart of this problem, in which we simply attempt to distinguish presence (value 1, 2, 3, 4) from absence (value 0).

MADELON: This artificial dataset contains data points grouped in 32 clusters placed on the vertices of a five-dimensional hypercube and randomly labeled +1 or -1, and it was one of five datasets used in the NIPS 2003 feature selection challenge [49]. The problem is a binary classification task to separate examples into two classes.

MAGIC: This dataset is composed of a series of Monte Carlo simulations regarding the registration of high-energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope (Major Atmospheric Gamma Imaging Cherenkov Telescope project, MAGIC) [13]. The problem is a binary classification task to discriminate the patterns caused by primary gammas (signal) from the images of hadronic showers initiated by cosmic rays in the upper atmosphere (background).

Nursery: This dataset was derived from a hierarchical decision model originally developed to rank applications for nursery schools, thus constituting the Nursery Database [89]. The goal is to predict the final decision, ranging from not recommended to priority. We have focused on the binary counterpart of this problem, in which an applicant was given either a priority recommendation or not.

Speed Dating: This dataset was gathered from participants in experimental speed dating events from 2002 to 2004 [37]. The problem was formulated as a binary classification task to predict whether both participants would like to date each other again, given each participant's questionnaire responses and characteristics.

WDBC: This dataset is composed of features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass associated with breast cancer [107]. The problem was formulated as a binary classification task to predict the presence of malignant tumor cells.

Table 1. Mean AUROC results on binary classification tasks after 10 repetitions. Datasets with no pre-defined test set were subject to an 80-20 cross-validation split.

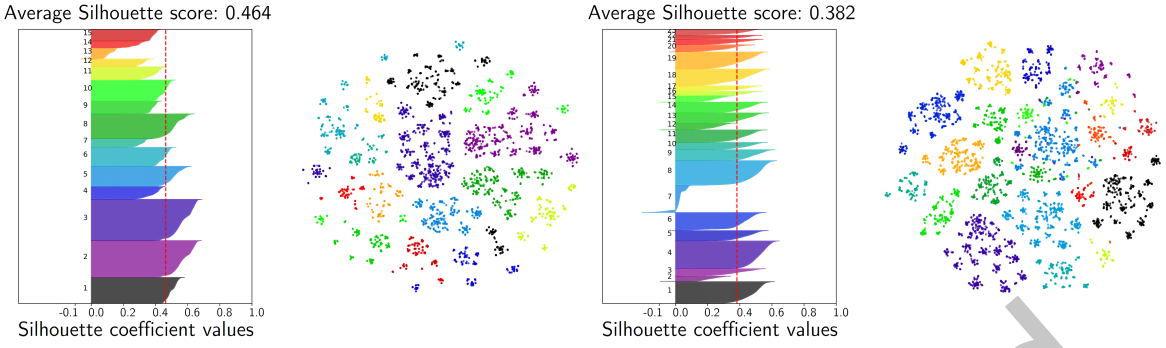
Benchmark			Baseline Algorithm								Rashomon	
Dataset	Rows	Cols	DT	Ada	RF	XGB	LGBM	CatBoost	GRANDE	NSGA-II	Ensemble	Ratio
APS	76000	172	.866	.824	.869	.835	.853	.888	.855	.859	**.911**	12.4%
Diabetes	101766	1691	.544	.614	.599	.615	.616	**.619**	-	**.619**	**.618**	17.4%
Heart	303	171	.748	.787	.826	.796	.830	.834	.825	.806	**.839**	50.3%
Nursery	12630	784	**.999**	**.999**	**.999**	.991	**.999**	**.999**	**.999**	**.999**	**.999**	83.2%
WDBC	569	903	.949	**.973**	.967	.963	.967	**.974**	**.975**	**.973**	**.974**	21.5%
Wine	4898	13	.762	.722	.802	.755	.764	.782	.802	**.804**	**.805**	8.9%
MAGIC	19020	102	.808	.830	.857	.837	.850	.850	**.897**	.809	.848	19.4%
MADELON	2000	502	.764	.598	.694	.828	.832	**.852**	.594	.674	*.746*	< 0.5%
Speeddating	8378	123	.650	.673	.630	.639	.642	.668	**.801**	.751	*.632*	< 0.5%

Wine Quality: This dataset is composed of chemical analysis of wines grown in the same region in Italy but derived from three different cultivars [1]. The analysis determined the quantities of 13 constituents found in each of the three types of wines, and the end goal is to predict the wine quality score, ranging from 0.0 to 8.0. We have focused on the binary counterpart of this problem, in which we wish to predict whether a given wine is of high quality (>5) or not.

5.2 Rashomon Ensemble Learning

Table 1 summarizes our comparison between our approach and classic and state-of-the-art algorithms. Specifically, we employ AdaBoost [40], Random Forests [15], XGBoost [22], LightGBM [61], Catboost [93], and GRANDE [81] as baseline algorithms. We also consider an evolutionary approach using the NSGA-II algorithm [27, 48] and evaluating the same number of models as our Rashomon Ensemble. The proposed evolutionary framework optimizes both the performance and feature diversity of the population and induces an ensemble from the best individuals of the last generation. For a fair comparison to the ensemble and boosting methods, we only employed decision trees as base constituents for our Rashomon ensembles. 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. No hyper-parameter tuning was employed, either on the baselines or the Rashomon ensembles, to ensure a fair comparison. The number of trees in all ensemble algorithms was limited to 50.

In the Nursery dataset, we verify that nearly all models lie inside the Rashomon space. This implies that the problem is relatively easy, and nearly any model is performant. In this scenario, the choice of using Rashomon ensembles or any other learning algorithm becomes less meaningful, and we observe that all baselines can achieve an AUROC of 0.99. In other scenarios, where the Rashomon ratio is large but not excessive (between 8% and 50%), we observe statistically significant gains when using our approach. The poor performance on the Speed Dating and MADELON datasets can be explained by the scarcity of contrasting explanations, represented by the small size of the Rashomon Set. The same cannot be said of the MAGIC dataset. One hypothesis for this behavior is related to MAGIC being a purely synthetic dataset. There might be some underlying pattern guiding the feature creation that is not present in the remainder datasets, which were crafted from different real-world problems. Figures 1a and 1b illustrate some Rashomon subspaces and their respective Rashomon partitions after sampling. The figures show a TSNE reduction of the Rashomon space and the optimal silhouette scores for each subgroup.



(a) Optimal k (15) for MAGIC models.

(b) Optimal k (23) for APS Failure models.

Fig. 1. TSNE reduction of the Rashomon space and optimal silhouette for each subgroup.

5.3 Evaluating Ensemble Composition and Robustness

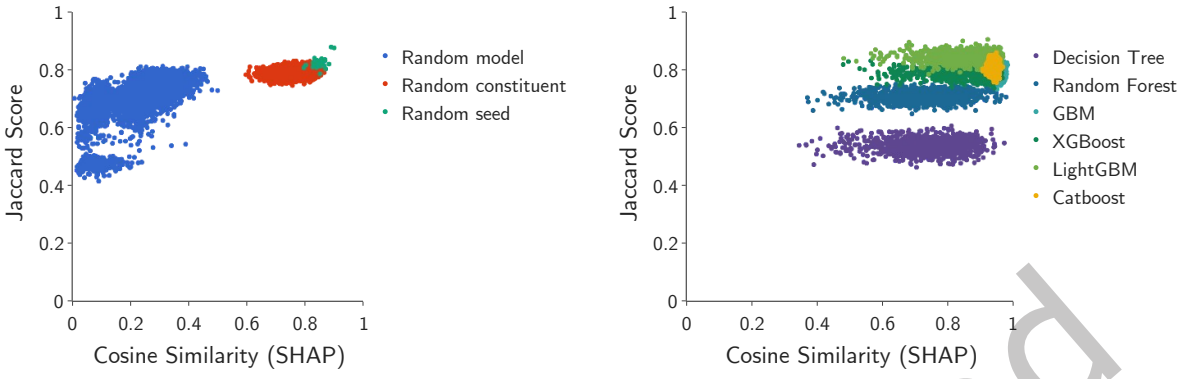
To assess the ensemble composition and robustness, we considered three scenarios to isolate the impact of key components in our Rashomon pipeline (Figure 2). Scenario I (green points) evaluates the stability of the full proposed method, in which we ran the complete pipeline 30 times with different random seeds to test sensitivity to stochasticity. Scenario II (red points) serves as a cluster-aware ablation, in which we retained the Rashomon subgroups but selected constituents randomly from each, over a total of 10,000 runs. This scenario represents omitting the intra-cluster optimization step. Finally, in Scenario III (blue points), we conduct a subgroup-free ablation. We ignored clustering entirely and selected models randomly from the entire Rashomon Set over 10,000 runs.

In Figure 2a, Scenario I (green) demonstrates stable similarity to the reference model across seeds, while Scenarios II (red) and III (blue) illustrate how omitting optimization or clustering increases variability. Figure 2b presents a comparison against the remaining baseline methods employed in our previous experiments.

To compare these different scenarios, we used a visualization scheme that jointly considered the Jaccard Index and SHAP values' similarity between the models found in each scenario and a reference model (performance in Table 1). The Jaccard Index provided insights into the degree of agreement between two sets of predictions, accounting for differences in prediction patterns that might be overlooked by standard performance metrics. On the other hand, the SHAP values similarity helped gauge the robustness of the explanations and whether the explanatory factors remained consistent despite the stochastic nature of the algorithms.

Upon analyzing the results presented in Figure 2, we observed that both Scenario I and Scenario II demonstrated models with high Jaccard Index and SHAP values similar to the reference model, indicating better ensemble compositions and increased robustness in terms of explanations. In contrast, Scenario III, which directly selected models from the whole Rashomon Set without considering subgroups, yielded models with statistically inferior performance compared to the reference model.

The outcomes of Scenarios I and II provide evidence that the Rashomon pipeline offers a viable solution to enhance ensemble robustness and credibility. By maintaining Rashomon subgroups and selecting ensemble constituents from each cluster, the Rashomon ensemble generation process appears to mitigate challenges related to ensemble underspecification, as discussed by D'Amour et al. [24], which mentions that only observing the performance of models poses an ineffective way to judge underspecification and thus, the potential multiplicity and divergence of seemly equal models under production settings. We also observe that performing the intra-cluster optimization step, represented by the green cloud of points on Figure 2a, severely reduces the variability



(a) Rashomon ensembles' sensibility to random constituent or seed selection.

(b) Remaining baseline methods' sensibility to random seed selection.

Fig. 2. Similarity to a reference model found from running the Rashomon and baseline pipelines, filtering models with statistically worse performance than the proposed threshold. MAGIC dataset.

on SHAP. That is, different runs of the algorithm seem to result in the same set of variables and key constituents being chosen. This same behavior cannot be verified for most of the remaining baseline methods in Figure 2b, except for Catboost.

5.4 Robustness to Distribution Drift

To evaluate the robustness of Rashomon ensembles to distribution drift, we conducted experiments to address concerns about out-of-distribution data. We considered two scenarios: the addition of Gaussian noise with increasing values of σ^2 to simulate data drift and shuffling feature values to evaluate the reliance on core 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 (Area Under the Receiver Operating Characteristic curve) 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. This scenario aimed to evaluate whether models could extrapolate from global information rather than relying on specific local patterns. The results of this data shuffle scenario are also presented in Table 2.

Upon analyzing the results, we observed that Rashomon ensembles consistently outperformed other models in both data drift and data shuffle scenarios, demonstrating their robustness to distribution changes. The Rashomon ensemble's ability to maintain superior performance under these perturbations showcases its capacity to adapt and generalize well to variations in data distributions.

The comparative analysis provided in Table 2 highlights the strengths of Rashomon ensembles in handling distribution drift. The findings suggest that the ensemble's ability to leverage diverse subgroups of models contributes to its robustness and adaptability, making it a promising approach for real-world applications where data distributions may evolve over time.

5.5 Intra-Model Associations

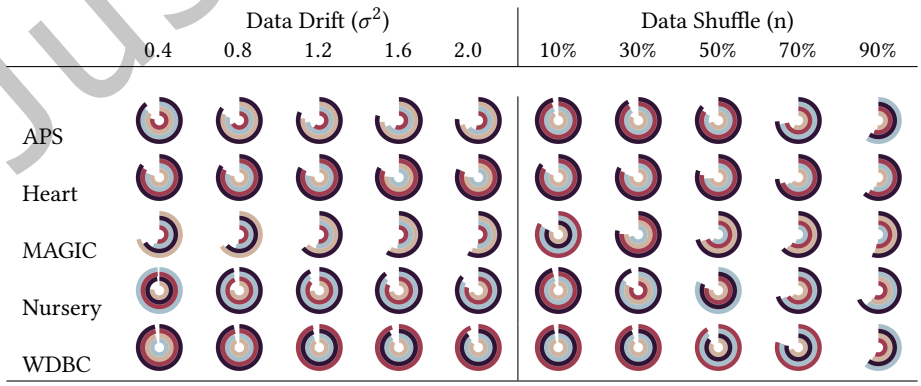
Out of the three datasets in which our approach was not able to beat the state-of-the-art, two can be explained by the Rashomon ratio. In most scenarios, a fair share of the sampled models presented performance statistically close to or superior to the all-in-one model, resulting in explanation diversity. In the cases of the Speed Dating and MADELON problems, less than 0.5% of the sampled models (less than 500) were comparable to the all-in-one model. This entailed scarcity in possible different solution paths and the possibility that nearly all the present features provided a complementary view of the solution. This harmed our sub-space division, leading to non-representative groups and poor predictive power.

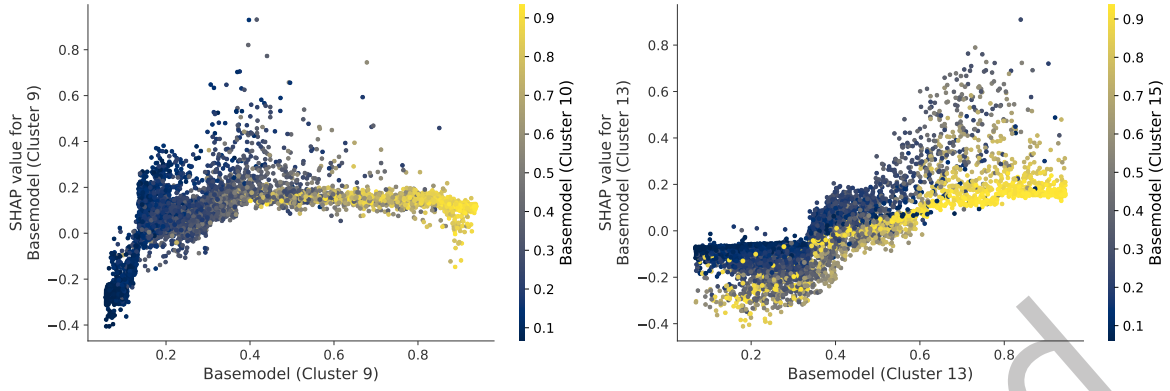
When exploring ensemble composition and robustness, we verified that once we filtered the underperforming models of III, this group population represented a sample of several ensembles that could be reached by direct optimizations over their constituents. On the opposite extreme, the group I represented the ensembles found following our proposed pipeline. It was important to remark that our approach depended on sampling the extremely complex model space, making it highly unlikely that the clusteroids found in each repetition were the same. However, the high Jaccard coefficient associated with the high cosine similarity between the SHAP vectors provided strong evidence that the centroids found in each repetition were contiguous, resulting in similar clusteroids that led to similar ensembles. Finally, group II represented a sample of possible optimization paths within the respective Rashomon Sets.

In all experiments, group III not only presented the lowest values of Jaccard and SHAP similarity but also consisted of the sparsest point cloud. Groups I and II were more cohesive and concentrated over high values of similarity with the reference model. When considering that all models had a statistically equal or higher performance than the reference model, it was reasonable to conclude that the pipeline involving Rashomon Sets reduced the impact of underspecification while retaining concise predictions. When further exploring drift by introducing Gaussian noise and shuffling feature values, the robustness of Rashomon ensembles became evident. In most explored scenarios, our approach remained the best-performing model, even when considering the MAGIC dataset, in which Rashomon ensembles had slightly worse performance than other models.

Further investigation of the relationships learned in the ensemble revealed a variety of interesting patterns, as illustrated by Figure 3 with partial dependence plots derived from the constituents' Shapley values. For instance, in Figure 3a, we observed that the ensemble learned to rely on the output of the 9th base model to give mostly positive predictions, with nearly all points above the 0.2 probability threshold presenting a positive SHAP value.

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





(a) Dependence between models 9 and 10.

(b) Dependence between models 13 and 15.

Fig. 3. Dependence between relevant base models in the MAGIC Rashomon ensemble.

We also verified that models 9 and 10 provided a complementary view of the problem, as higher prediction values of 9 were likewise associated with high prediction values of 10. Figure 3b showed that models 13 and 15 mostly contrast. Whenever there was disagreement, the relative importance of Model 13 increased. Similarly, model 13 presented Shapley values close to zero when both models agreed, resulting in a concentration of yellow points on the lower side of the distribution.

6 UNIQUE COLLABORATIVE DATASETS

To validate the effectiveness of our approach in real-world scenarios, we present the results of our evaluation across distinct applications conducted in collaboration with various companies and institutions: stainless steel surface defect detection, COVID-19 hemogram detection from blood counts, energy consumption forecasting, and medical bills auditing. 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.

6.1 Surface Defects in Stainless Steel Manufacturing

The quality of duplex stainless steel can be compromised by surface defects, such as slivers, which increase production costs due to their detection occurring only during the final inspection stage. In collaboration with *APERAM South America*, we analyzed the chemical composition and hot rolling process variables of duplex stainless steel plates. Chemical compositions were measured using spectrometers, capturing the relative abundances of 20 elements, while 1,160 temporal hot-rolling variables were collected. We extended the feature space by considering elemental ratios, increasing the total to 220 chemical attributes. For the hot rolling data, we discretized the temporal series into 30-second intervals and calculated statistical moments, resulting in 11,488 hot rolling features after filtering non-actionable variables. This data was used to predict the likelihood of sliver formation as a binary classification problem.

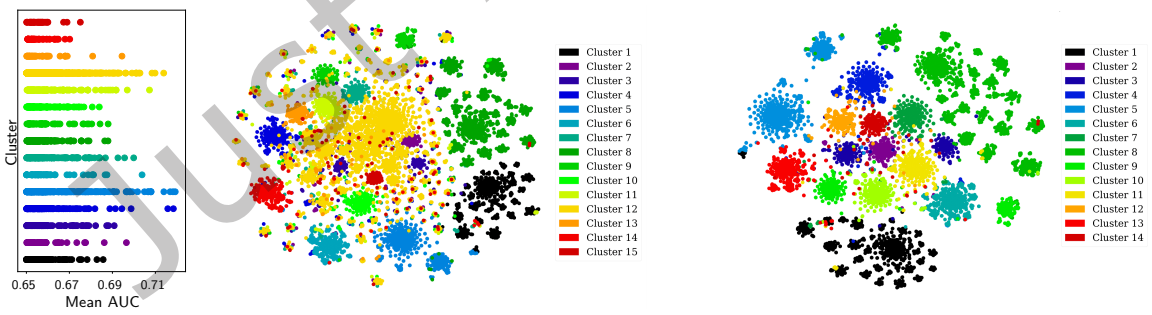
As described by Barbosa et al. [10], identifying factors contributing to sliver formation is challenging, as these defects can arise from a combination of process variables or chemical compositions at various steelmaking

stages. We hypothesize that different data structures might correspond to different models and that models with similar feature importance distributions likely reflect similar mechanisms. Our analysis identified correlations between feature sets and predictions, indicating that some features are more sensitive to specific defect formation mechanisms [126].

To explore the model space, we constructed a Rashomon Set by sampling 75,000 models for each feature set size until no significant performance improvement was observed, resulting in 1,049,999 models. Since no strong baseline exists in the literature, we compared it against an all-in-one model, which included all features and achieved an AUC of 0.62. Models surpassing this threshold formed the Rashomon Set, comprising 63,374 models (6.04%). For visualization purposes, Figure 4 presents a t-SNE projection of 2,500 models from the Rashomon Set, clustered by their explanations.

We selected the optimal model for each cluster using two approaches. In the first one, we simply elect the clusteroids as the ensemble constituents. In the second approach, we reframe the problem of model selection as a graph search problem. We consider each possible model as a node in a graph, and two models are connected if they can be reached from simple feature addition. For each cluster, we employ the A* search algorithm in this graph using as a heuristic the AUROC of each model and penalizing paths that lead to models that lie outside the cluster boundaries. Figure 5 compares the performance of our ensemble (both A* and clusteroid models) against state-of-the-art tree-based ensemble techniques and other classical algorithms with and without feature selection.

Once representative models were found, we asked for insights from metallurgical experts. The main lesson was that there were cases where some conclusions did not fit with realistic scenarios. For example, some models hinted towards increasing carbon concentration to such high levels that the steel plates could not be classified as Duplex anymore. These inconsistencies highlight the advantage of our approach, which allows domain experts to discard unrealistic models at production time without involving data scientists and the need to retrain and evaluate new models, since, during training time, we sample from the complete model. We verified that this sort of approach drastically increases the power of domain experts and helps build trust in the models, as they feel in control of these domains and business-specific decisions concerning model development. This insight led to a patent further explained in Section 6.3. After filtering unrealistic patterns, the most relevant ones were turned into production rules and employed in the 2019 and 2020 steelmaking processes. A reduction from 49% to 3% in



(a) AUCs. (b) All found clusters. (c) Filtering the sparse cluster.

Fig. 4. T-SNE visualization of the sampled Rashomon space for models trained on the steel plate defects problem. Each point represents a model. Models are placed according to the defect explanations assigned to each steel plate so that models that possess similar SHAP values are placed next to each other in space. The color indicates the cluster for which the model was assigned.

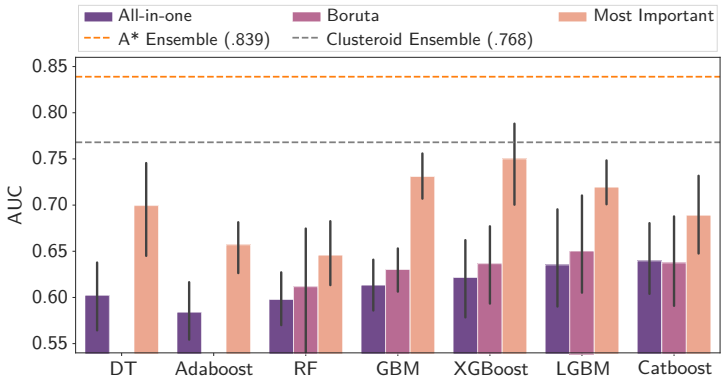


Fig. 5. Comparison of different algorithms to our approach in the steel manufacturing defects problem. Even when employing the clusteroid ensemble, in which most constituents are underperforming, our approach exceeds other state-of-the-art results.

the occurrence of heating slivers was reported, showing the potential of this strategy in real-world problems and validating the proposed framework.

6.2 Diagnosing COVID-19 from Complete Blood Counts

In late 2019, Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) emerged in Wuhan, China, leading to a global outbreak of Coronavirus Disease 2019 (COVID-19) within weeks [60, 116]. By the time of writing, over 630 million COVID-19 cases and 6.67 million deaths have been reported worldwide. Diagnosing COVID-19 is complicated by initial symptoms such as fever, dry cough, and tiredness, which overlap with many respiratory diseases [29]. Complete blood counts (CBC) are commonly used for diagnostic purposes [111]. As a low-cost test, CBC measures various analytes and can provide insights into potential diseases, including infectious ones. However, correlating specific CBC results with particular diagnoses can be challenging, as similar changes may occur across different diseases.

In analyzing complete blood counts of individuals with COVID-19 infection in isolation, we find some changes to be quite characteristic of the disease [38, 39, 57], hinting at the potential for automated detection and screening of the disease using machine learning. However, many possible analyte combinations might lead to the same conclusion regarding a target disease, thus elucidating the Rashomon Effect and posing a suitable problem to deploy our ensembling approach. Numerous models have been proposed for automated COVID-19 diagnosis using CBC and omics data. We argue that the detection performance of these models may be biased due to non-unique patterns not exclusive to SARS-CoV-2. Our study utilizes a dataset from 2016 to 2021, in collaboration with *Grupo Fleury*, encompassing blood tests and RT-PCR results across Brazil for COVID-19 and other pathologies, including Influenza-A and H1N1 [122].

Data collection for 2020 and 2021 includes 900, 220 unique individuals, 809, 254 CBCs, and 1, 088, 385 RT-PCR tests, with 21% (234, 466) positive results and less than 0.2% (1, 679) inconclusive results. This work does not consider demographic, prognostic, or clinical data, such as ethnicity or hospitalization. We frame the task as a binary classification problem and analyze two timeframes: the early pandemic stage (the first wave of COVID-19 in Brazil) and a second stage post-November 2020, coinciding with the emergence of the *P1* variant that led to a health crisis in Amazonas [52].

Our algorithm's first step involved sampling 100, 000 models from the complete model space, examining both raw analytes and analyte ratios as features. Previous studies have explored the use of CBC for COVID-19

detection through various machine-learning methods and, as such, provide reference points for our Rashomon space induction. Notable examples include a naive Bayes classifier with an AUROC of 0.84 [6], a gradient boosting machine achieving 0.81 [104], and a neural network and random forest model reaching an AUROC of 0.94 [9]. Other studies report AUROC values ranging from 0.88 to 0.94, with one analysis involving 114, 957 individuals in a COVID-negative cohort [105].

Based on this literature, we established a performance threshold of AUROC 0.81 to define a minimally performant model for our Rashomon Set, resulting in a sampled model space $\mathcal{H}'$ containing 47,708 models (47.71% performing better than the literature threshold). This substantial Rashomon Set indicates that blood-related features are significant for preliminary disease diagnosis. Figure 6 illustrates the induced Rashomon space, highlighting divisions after clustering models based on their explanatory vectors.

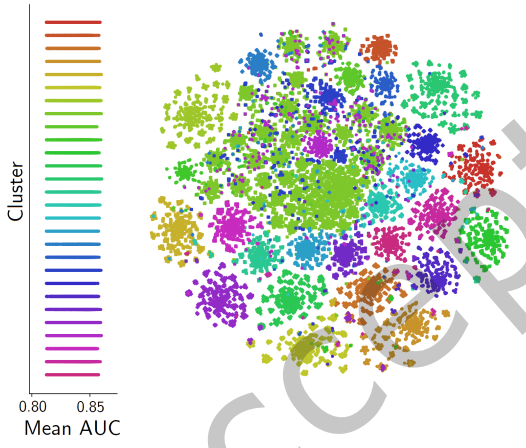


Fig. 6. TSNE visualization of the COVID-19 Rashomon space of models trained to predict COVID-19 diagnosis. No clear relationship exists between cluster assignment and predictive power. Cluster 11 appears to be more spread over the space, overlapping with other clusters, while the remaining ones are mostly concise, reminiscent of the steel-plate defects of Rashomon space.

In line with previous findings in the literature, not all CBC analytes are relevant for differentiating the target diseases, and some may detract from model performance. To refine feature selection for each representative Rashomon model, we represented the model space as a directed acyclic graph (DAG), where each node corresponds to a feature subset. The A* algorithm [51] was applied, utilizing the AUROC of models represented by each vertex as a heuristic. Once models were selected, their suitability as Rashomon constituents was assessed. Our hypothesis posits that models exhibiting disagreement under data drift and with diverse explanations can form a more robust ensemble. In Figure 7, we introduced Gaussian noise to normalized features and examined the probability distributions returned by each constituent. We observed a direct correlation between noise and confidence intervals, indicating increased divergence among models under drift and thus supporting the appropriateness of our constituent selection.

By mid-November 2020, Brazil entered the second wave of COVID-19, which eventually led to the collapse of the health system in Manaus, the capital of Amazonas, a state in Brazil [33]. One of the explanations raised by the local government was the emergence of a new COVID-19 variant, known as $20J/501Y.V3$ - or simply P.1 [52]. To evaluate the performance of our COVID-19 model as the SARS-CoV-2 virus mutates, we trained it on two distinct

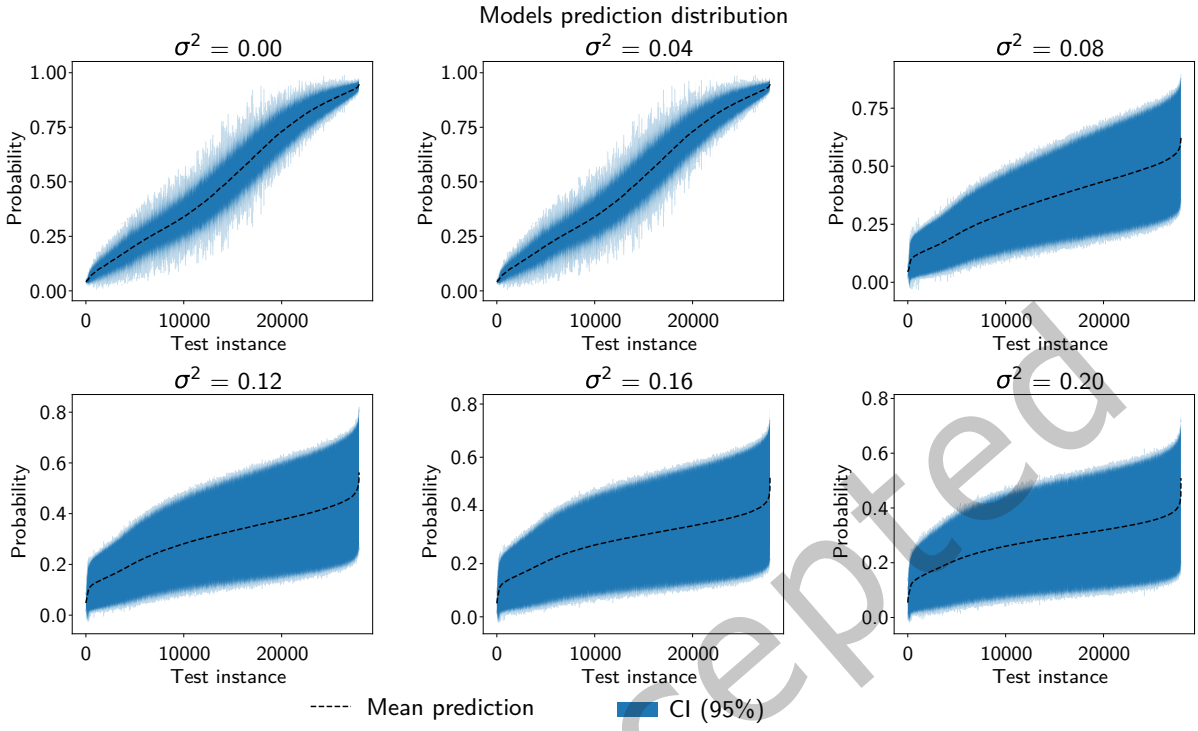


Fig. 7. Effect of introducing noise to input features of models trained for COVID-19 diagnosis from blood analytes.

points in time. The first one, which we will refer to as the 'First-wave model', was trained using the training set associated with the first wave. The second, which we will refer to as the 'Second-wave model', was trained using the training set associated with the second wave in Brazil. Figure 8 presents AUROC values obtained from these models during the pandemic until March 2021, utilizing a 7-day sliding window and depicting the prevalence of COVID-19 cases over time. We focus on three periods of interest: when the reproduction number exceeded 1.00, during the holiday season, and during Carnival, which included gatherings despite event cancellations.

We evaluate the COVID-19 Rashomon ensemble on both periods using the model trained with data up to October, illustrated in Figure 9. Performance on both periods appears to be comparable, thus implying that the constituents were able to properly generalize to the second wave. We also verify that the Rashomon ensemble remains a suitable approach, outperforming all constituents in either scenario. Further, the empirical risk found during training can be used to estimate the empirical risk on production, as no significant divergences were observed. Overall, leveraging our Rashomon ensemble technique, we predicted COVID-19 RT-PCR outcomes using CBC data, achieving an AUROC of 0.917.

6.3 Including specialists in the model creation process

A patent was filed focusing on incorporating experts and decision-makers into the AI model development pipeline [110]. The core of the patent is to foster a sense of shared responsibility and co-creation, where researchers make decisions regarding the technical aspects while domain experts provide guidance on domain-specific topics

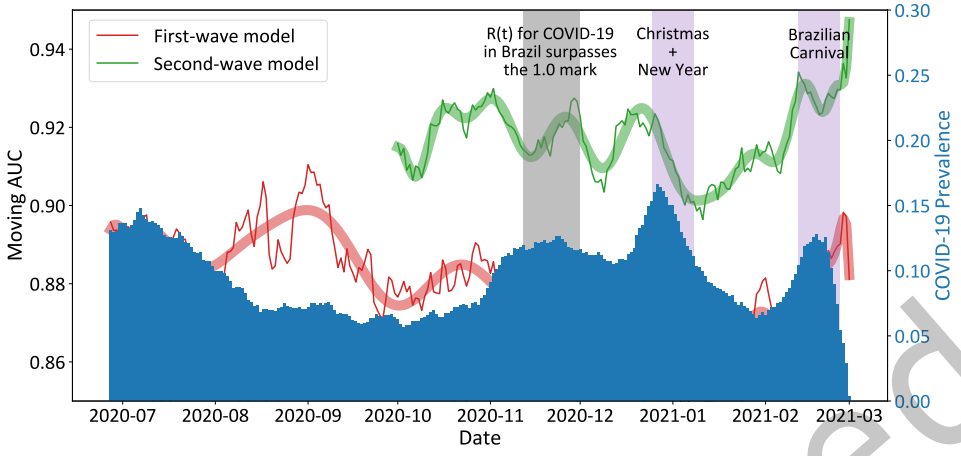


Fig. 8. AUROC fluctuation over time considering a 7-day sliding window on 357 956 CBCs. The red line represents the model trained only on the first wave of COVID-19 in Brazil data (up to 2020-06), while the green line represents a model trained with data immediately before the start of the second wave of COVID-19 in Brazil (up to 2020-10). Thinner lines depict the measured AUROC values, while thicker lines illustrate their respective trends. The second-wave model can retain performance during the second wave while the performance of the first-wave model deteriorates. Key events are marked in gray and purple.

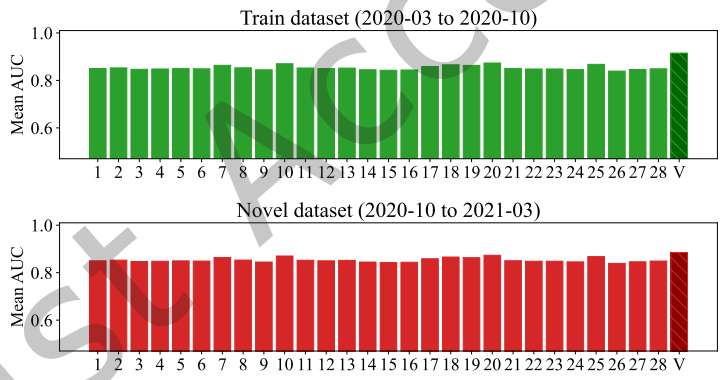


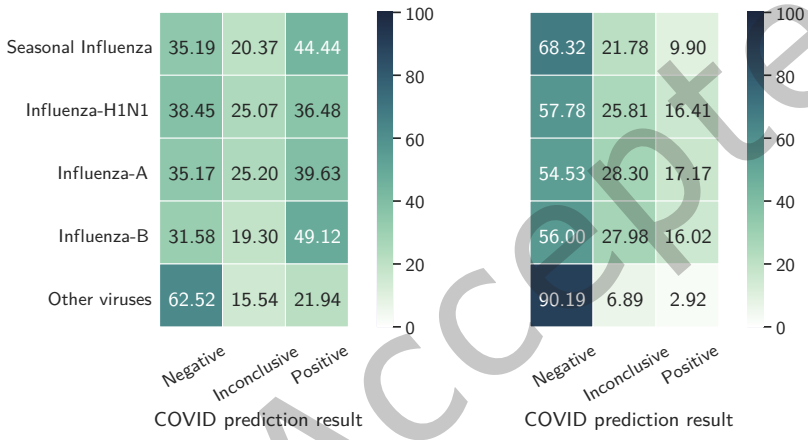
Fig. 9. Comparison of model performances across periods. Each constituent model is represented by the Cluster from which it hailed. Both in the train and novel datasets, we can observe that all constituent models behave similarly. We should expect data distributions from late 2020 and early 2021 to be properly represented in data before October 2020.

such as feature selection, exclusion, and the interpretation of discovered patterns. This balance ensures that the model development process integrates both technical rigor and domain knowledge.

Among the methods included in the patent is the usage of the developed Rashomon ensembles framework, which identifies multiple contrasting patterns in the data as per the Rashomon Effect. Frequently, some of these patterns are closely aligned with experts' experience and existing domain literature, enhancing the specialists'

trust in the model. This alignment thus leads to more constructive discussions, especially regarding patterns that deviate from established knowledge. Since experts find some patterns consistent with their mental models, we verified that they are more likely to engage with the results and question potential gaps in the literature rather than dismissing them outright.

In the specific case involving COVID-19 diagnosis from blood-count data, this methodology prompted experts to question whether a model trained solely on COVID-19 cases could distinguish it from other respiratory diseases. This line of inquiry proved crucial in improving the model's final performance, demonstrating the value of expert involvement in refining and validating the patterns uncovered by the AI models. Figure 10 shows how different models perform specifically on individuals who were infected by some viruses in 2019. The ideal result would be all predictions being negative for COVID-19. However, models trained solely on COVID-19 data failed to do so (Figure 10a). Including viruses other than SARS-CoV-2 during training increases the performance of 2019 data (Figure 10b).



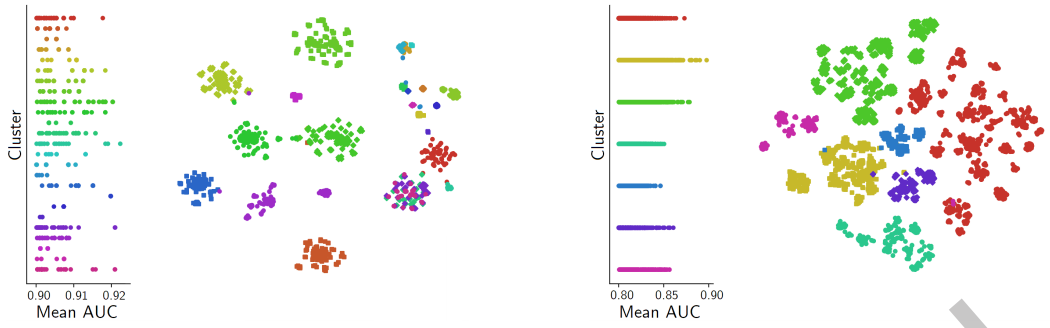
(a) Before expert input. (b) After expert input.

Fig. 10. Results of different models evaluated on 11 116 CBCs from 2019 individuals with confirmed RT-PCR results for diverse viruses, including Influenza-A, Influenza-B, Influenza-H1N1, and Seasonal Influenza. Left is a model trained only on SARS-CoV-2 data. CBC (–) includes only COVID-19 (–). The model to the right was trained using data of diverse viruses, including SARS-CoV-2. CBC (–) also includes viruses other than SARS-CoV-2.

6.4 Data Divergence in Hospital Settings

We investigated the behavior of Rashomon ensembles under train and production divergence using two datasets: one related to COVID-19 and another to Alzheimer's disease. Both datasets contain information from different contexts, indicative of shifts in data distribution. Due to the lack of a strong baseline in the literature, we adopted the "all-in-one" approach to establish a reference model (f_{ref}) and quantify uncertainty (ϵ) for the ensemble.

COVID-19: This dataset is an initiative of the São Paulo Research Foundation (FAPESP) and includes pseudonymized data from two Brazilian hospitals: *Beneficência-Portuguesa Hospital* (HBP) with 91, 648 exams and *Sírio-Libanês Hospital* (HSL) with 37, 643 exams. The data encompasses clinical and laboratory exams as well as hospitalization information. The binary classification task aims to predict the death prognosis of COVID-19 patients 20 days prior. The training dataset comprises exams from 453 individuals hospitalized at HBP, while our considered production dataset consists of exams from 4, 018 individuals hospitalized at HSL, highlighting potential distribution drift.



(a) COVID-19 (b) Alzheimers'

Fig. 11. TSNE visualization of the Rashomon space of each problem.

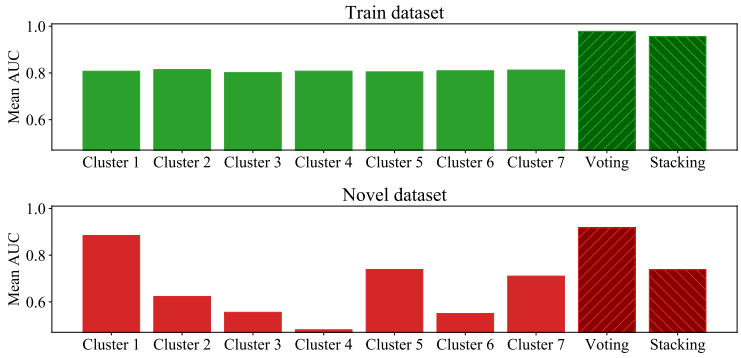
Alzheimer's Disease: This dataset includes patients with suspected Alzheimer's Disease symptoms, featuring attributes such as gender, age, education level, and lab results. The binary classification task predicts whether a patient is diagnosed with Alzheimer's. Data is sourced from the Geriatrics and Neurology departments, each presenting unique socio-economic characteristics. The training dataset consists of 154 exams from the geriatrics department, and our considered production dataset consists of 166 exams from the neurology department, ensuring divergence due to the inclusion of non-geriatric patients.

The all-in-one approach involves creating a comprehensive model that utilizes the entire dataset for predictions. We trained a decision tree model with all available data and features, referred to as the "all-in-one" model, which captured overall patterns. The model achieved average AUROC values of 0.90 and 0.81 for the COVID-19 and Alzheimer's datasets, respectively, establishing benchmarks for minimal performance in the Rashomon Set. After sampling 100,000 models for each dataset, this resulted in subspaces containing 2, 554 and 6, 251 models, leading to Rashomon ratios of 2.5% and 6.2%. Figure 11 illustrates the Rashomon subspaces identified through clustering. We employed t-distributed Stochastic Neighbor Embedding (t-SNE) for visualization, where each point represents a model, colored by cluster assignment. K-means clustering was utilized to define subspaces, with the number of clusters determined by maximizing the silhouette value.

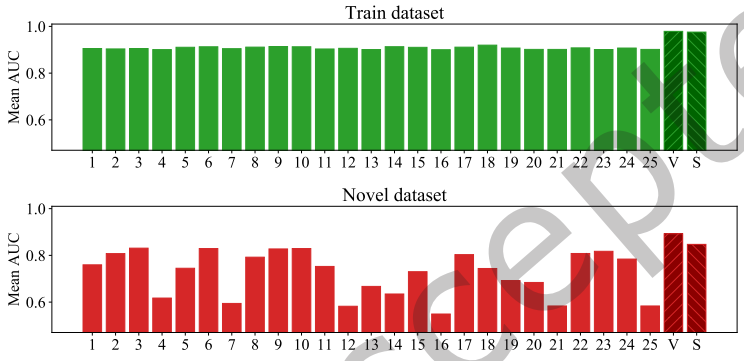
To evaluate model performance under train and production divergence, we combined models in a voting scheme. Since the task is a binary classification, the probability returned for the ensemble constitutes the absolute agreement rate of the constituents (e.g., for an ensemble of 10 constituents, a probability of 80% means that 8 out of the 10 constituents agreed on predicting the positive class). Thus, probabilities near 0% or 100% indicate strong agreement, while values close to 50% suggest uncertainty. Voting provides an interpretable measure of prediction reliability, beneficial in cases of production divergence and unknown data distributions. However, we also considered a stacking scheme where a meta-model learns to optimally combine the constituent outputs given the patterns present in the training data.

Figure 12 displays performance comparisons for each base model and ensemble on the COVID-19 and Alzheimer's datasets. While all constituent models performed similarly on training data, voting consistently outperformed stacking on new, unseen data, confirming our hypothesis regarding erratic behavior across models. Both ensemble techniques exceeded the performance of individual models and state-of-the-art methods.

To understand the relationship between model agreement and prediction confidence, we stratified test data points based on ensemble agreement and evaluated performance as shown in Figure 13. A direct correlation was observed between ensemble performance and intra-constituent agreement. When the models agreed, ensemble accuracy approached 1, indicating high confidence. Conversely, as agreement neared 50%, ensemble accuracy



(a) COVID-19 dataset pairs.



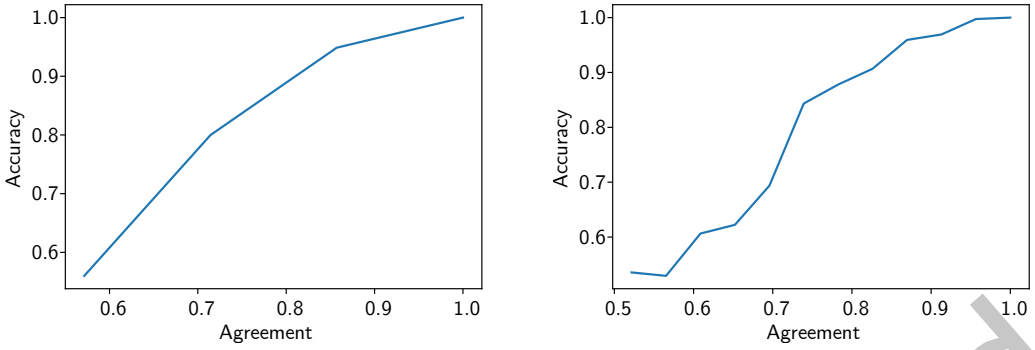
(b) Alzheimer's dataset pairs.

Fig. 12. Comparison of model performances across datasets. Each constituent model is represented by the Cluster from which it hailed. In the train datasets, we can observe that all constituent models behave similarly. On the novel datasets, under the unknown U distribution, performance becomes unpredictable. However, we can verify that the voting approach always outperforms the best constituent model, thus presenting itself as a suitable technique to mitigate this behavior.

resembled random guessing, supporting our hypothesis that model agreement is critical for prediction reliability. This relationship has significant implications for deploying Rashomon ensembles in dynamic data environments. Consensus among constituent models suggests instances similar to those encountered during training, enhancing trust in predictions. Conversely, divergence indicates that observations may fall outside training data, leading to unreliable predictions.

6.5 Determinants of energy consumption

Energy systems are increasingly linked with economic, social, and climate factors. Understanding these interconnections is crucial for electricity planning, particularly regarding how they impact electricity consumption and supply. Climate and weather significantly influence energy demand, with temperature being a key variable [26, 56]. Various temperature-related metrics effectively approximate energy consumption [59]. While degree days are commonly used for load forecasting, recent studies indicate that other weather variables, such as humidity, also affect electricity demand, particularly during hot days [78, 115]. This suggests a range of potential weather predictors, indicating the Rashomon Effect and aligning with our analytical approach. Given the strong connection



(a) COVID-19 production dataset.

(b) Alzheimer's production dataset.

Fig. 13. Relationship between Rashomon Ensemble accuracy and intra-constituent agreement. As we hypothesized, there exists a direct relationship between ensemble performance and agreement. When constituents agree, accuracy is close to 1, implying that the observed instance is similar to what was seen in training, and we can trust the prediction with high confidence.

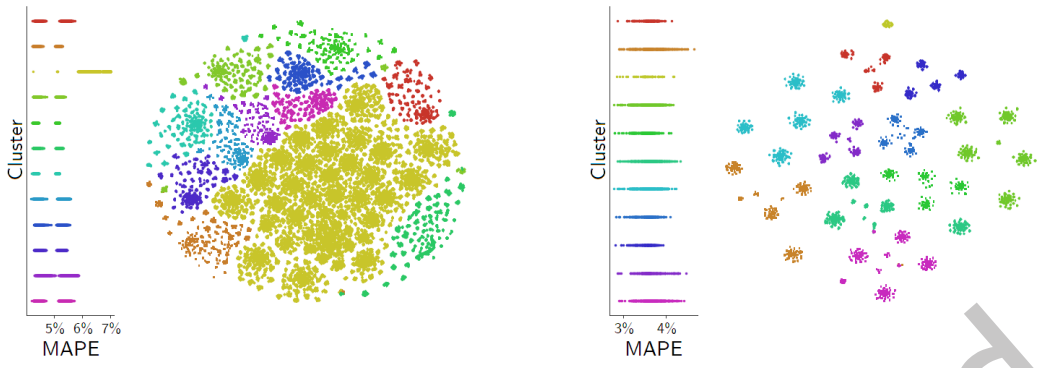
between weather and Brazil's electrical system, we focus on weather determinants of consumption, as presented in our previous work [123, 124].

We utilized two primary datasets in our study: the Brazilian National Energy System Operator (ONS) historical reports [90] and the ERA5 reanalysis [54]. Energy data from the ONS website spans from 1999 to the present, including daily measures of load, maximum consumption, mean and total daily megawatts (MWd), and hourly megawatts (MWh). The ERA5 dataset offers global hourly estimates from 1950-2021 for atmospheric variables at a spatial resolution of 0.25 degrees (approximately a 30 x 30 km grid). Our final weather feature subset includes daily temperature minimums, means, and maximums, humidity, wind speed, precipitation, heating degree days (HDD), cooling degree days (CDD), heat index, wind chill index, apparent temperature, and the derived HDD and CDD from respective indices.

The primary objective is to predict consumption in the absence of abnormal events, enabling a direct comparison between predicted and actual consumption. We formulate this as a regression problem. Given a set $w \in W$ of weather descriptors and a set $t \in T$ of time descriptors, we apply a function $f(w; t; \sigma)$ parameterized by σ to map a period to consumption. To exclude disruptive factors, we identify optimal subsets $W' \subset W$ and $T' \subset T$.

We propose three main groups of factors influencing electricity consumption based on existing literature: load growth, historical events, and weather [44]. We observed a logistic growth trend in yearly energy consumption. Normalizing daily consumption by the load growth function, derived from yearly load interpolation while filtering atypical events, enables the construction of a counterfactual model focused on weather and temporal factors. This allows for the development of various models $f'(w; t; \sigma')$ with different feature sets, forming an ensemble that captures potential explanation biases in line with the Rashomon Effect.

The first step in building our ensemble involves sampling models to estimate the Rashomon space. We chose to use the Mean Average Percentile Error (MAPE) for inducing Rashomon Sets. Figure 14 illustrates the found Rashomon space after sampling 100 000 models and also depicts the impact of different choices for the MAPE ϵ threshold. For MAPE of 7%, we found a Rashomon ratio of .82 (82 114 models presented $MAPE \leq 0.07$) while decreasing this threshold to 5% reduced the ratio to .04 (4 064 models presented $MAPE \leq 0.05$). In Figure 14a, we noted that underperforming models clustered together, hinting that including these models in the ensemble



(a) $\epsilon = 7\%$. (b) $\epsilon = 5\%$.

Fig. 14. Induced Rashomon spaces with ϵ thresholds of 7% and 5% MAPE. Overestimating ϵ results in a larger Rashomon space and undesirable correlations between cluster assignments, explanatory factors, and performance.

might not be productive. By properly tuning the ϵ value, we reduced the Rashomon space and extracted more meaningful clusters, as depicted in Figure 14b.

Following our algorithm, we searched for optimal representatives within each explanation cluster. Once again, we represented the model space as a directed acyclic graph (DAG) and searched for optimal constituents. In Figure 15, we added Gaussian noise to the normalized features and observed the normalized consumption estimates from each model. In the absence of noise, all models behaved similarly within a narrow confidence interval. However, introducing noise led to increased divergence among models, confirming the direct relationship between noise levels and the width of the confidence interval, indicating reduced ensemble reliability. Although the mean ensemble prediction remained stable, minor noise introduced significant prediction variability.

As shown in all the other experiments, agreement is a key metric to measure the reliability of the Rashomon ensembles. However, defining agreement in regression contexts is challenging since visual inspections of confidence intervals are impractical. While in classification, agreement occurs when two models predict the same class; in regression, we consider predictions 'similar' if they are within a context-dependent range. We selected the coefficient of variance (C_V) between regressors as our agreement measure, defined as:

$$C_V = \frac{\sigma}{\mu} \tag{8}$$

Here, σ represents standard deviation, while dividing by μ provides a dimensionless metric indicating the variability extent relative to population means. A higher C_V reflects greater dispersion and disagreement between constituents. In a voting scheme, the ensemble prediction is μ, while C_V indicates the degree of divergence among individual constituents.

We validated our ensemble by comparing performance with constituent prediction dispersion C_V, as shown in Figure 16. A direct relationship between these metrics was observed, with most instances below $C_V = 0.05$, indicating less than 5% dispersion among ensemble constituents. In such scenarios, we anticipate an MAPE below 4%, which is favorable. As dispersion increased, error escalated, confirming that disagreement among constituents compromises prediction reliability.

We evaluate the robustness of our approach through a case study in Brazil from 2001 to 2002. Severe drought and low reservoir levels in early 2001 raised concerns about a potential grid collapse, prompting the Federal Government to implement policies aimed at reducing energy consumption by 20% [11]. These measures included

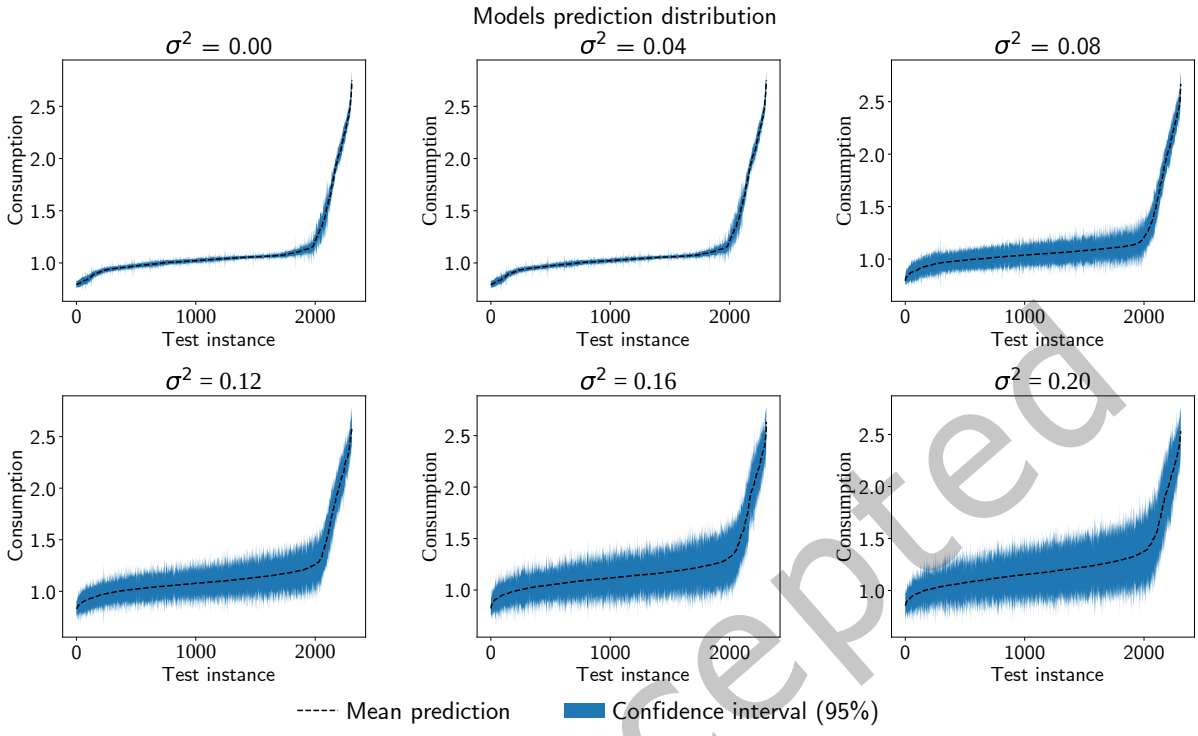


Fig. 15. Effect of noise on ensemble constituents' input features and predictions for models trained to predict Brazilian energetic consumption.

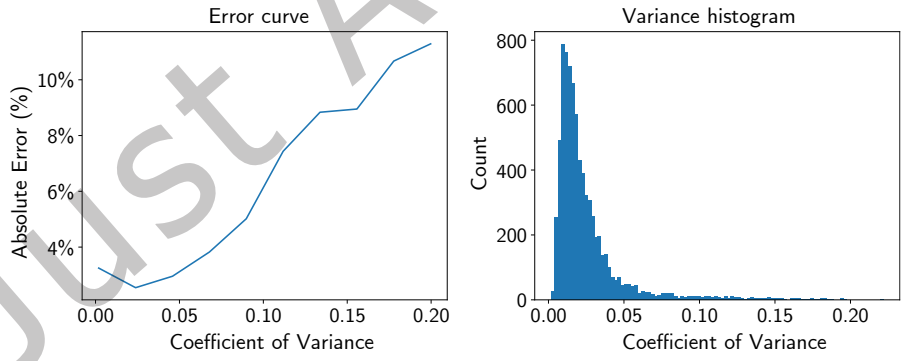


Fig. 16. Relationship between constituent agreement and ensemble performance when predicting energetic consumption. The coefficient of variance between constituent predictions was used as an agreement metric.

awareness campaigns, peak-hour price increases, and incentives to limit non-essential energy use. This period provides a clear opportunity to measure expected impacts, allowing for a direct evaluation of our method. Figure 17 illustrates the effects of these policies across regions. Our counterfactual model revealed mean and median

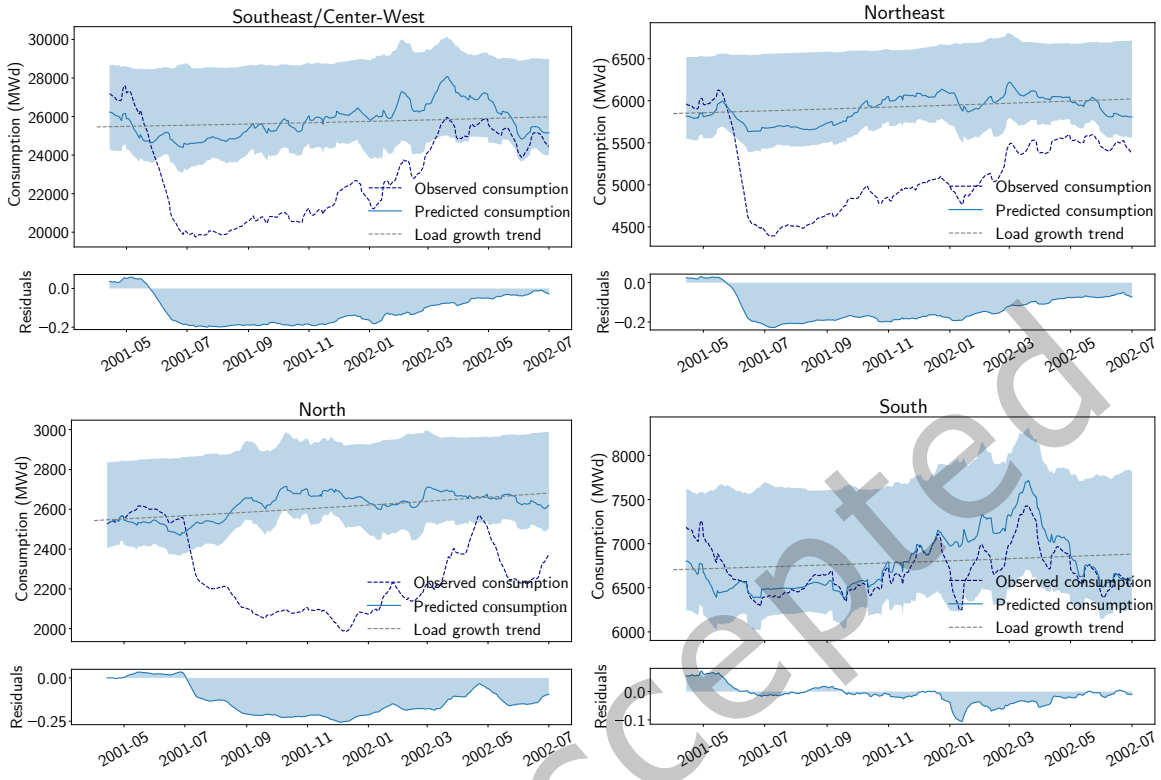


Fig. 17. Consumption during Brazil's 2001 *Apagão* (Blackout). The South region did not adhere as closely to the restrictions as the other regions.

relative residuals for the Center-West region from June 2001 to January 2002 of −18.1%(±1.6%) and −18.6%, closely aligning with the anticipated −20% reduction during the restriction period from July 1, 2001, to February 19, 2002. Similar trends were observed in the North and Northeast regions, with relative residuals of −19.1% and −18.9%, respectively.

We also examine the year 2020, particularly during the early COVID-19 pandemic, which imposed significant mobility restrictions and reduced global GDP. Figure 18 compares electricity consumption with the Oxford Stringency Index, which measures the strictness of COVID-19 policies [50]. During this period, there was a notable relationship between the drop in consumption and the Stringency Index in Brazil. From April to June, electricity consumption showed a mean relative residual of −8.87% ± 1.2%, in accordance with economic records [47].

The lack of demographic and behavioral variables enables us to assess the pandemic's impact through a counterfactual approach, with the expectation of uniform constituent errors in 2020. This period demonstrates the usefulness of the Rashomon ensembles and our proposed approach in contrast to other counterfactual techniques. Figure 19 shows ensemble and constituent performances when trained on data from 2014 to 2018 and applied to 2019 and 2020. As expected, errors for 2019 are similar to those seen in training, meaning similar weather variable distributions. However, in 2020, the Rashomon ensemble displayed erratic behavior, with a coefficient of variance rising from 0.04 in 2019 to 0.14 in May 2020, suggesting atypical weather patterns. Since this pertains to the

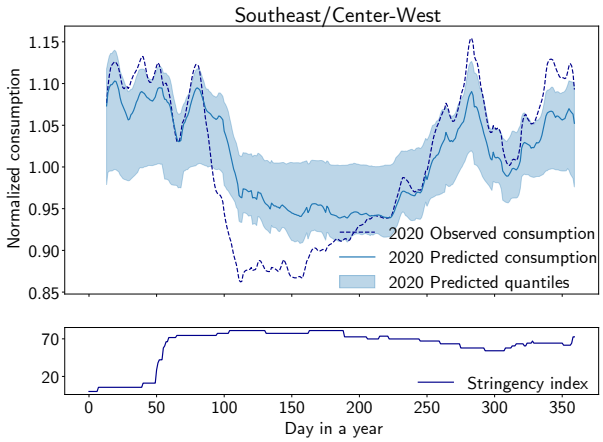


Fig. 18. Impact of COVID-19 on Brazil in 2020. Stringency data from Ritchie et al. [97].

high point of restrictions in Brazil, we can conclude that the pandemic effect shadowed this change in weather. In October, Brazil experienced one of its most intense heatwaves in history, breaking century-old temperature records [80]. Such extremes diverged from the 2014-2018 distribution. If May's prediction errors were solely pandemic-related, we would expect consistent performance across constituents. However, the erratic behavior implies that the weather throughout the year was unusual and evident months before the heatwave.

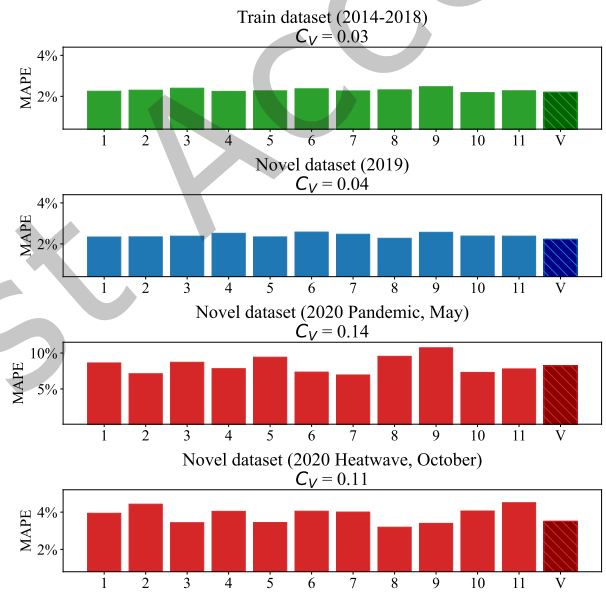


Fig. 19. Comparison of model performances across periods, with each constituent model represented by its cluster. In 2019, models showed low MAPE; in 2020, individual performances became erratic, indicating data from different distributions due to the October heatwave.

6.6 Auditing medical bills in large healthcare companies

Auditing hospital and outpatient bills is critical for identifying errors and overcharges, but the high volume of daily bills makes comprehensive auditing challenging [63, 109, 120]. We collaborated with Unimed-BH, the leading private healthcare provider in Minas Gerais, Brazil, to address this issue. With over 1.5 million patients, only a small fraction of bills are manually reviewed, leading to missed discrepancies and inefficient use of expert auditors. Automating this process allows auditors to focus on the most complex cases.

We developed a tool to rank bills based on a "discrepancy score", estimating the likelihood of inconsistencies [127]. Bills from Unimed-BH's accredited network are evaluated using a Rashomon ensemble, where each model contributes a distinct perspective on inconsistency. We define a dataset $X = \{x_1, x_2, \ldots, x_N\},\ x_i \in \mathbb{R}^D$ and a representation space $Z \in \mathbb{R}^K (K \ll N)$. The score function $\tau(\cdot) : X \to \mathbb{R}$ ranks bills by inconsistency likelihood, using an ensemble of models tuned to different features. Each model i maps data to Z (e.g., via autoencoders) or generates a score $\tau_i(\cdot)$ (e.g., via isolation forests), which informs the ranking of X.

Unlike in our previous experiments, we explored how to merge multiple algorithms in a single ensemble and weigh their importance. We employ a multi-armed bandit (MAB) approach, balancing exploration (auditing lower-ranked bills) and exploitation (focusing on high-ranked bills). This method allows us to incorporate varied algorithms without directly comparing their applicability [14]. The ensemble models inconsistency detection as a ranking task, where high-ranked bills are more likely to contain misaligned financial values. The MAB algorithm assigns weights to each model based on its performance with labeled data, combining models optimally to reflect diverse explanations. The final ensemble uses a weighted voting method, dynamically adapting to new data. Constituent models include:

- Generative methods: PCA, Denoising AutoEncoders.
- Regressor methods: Elastic Net, Lasso, Support Vector Regressor, XGBoost, LightGBM.
- Isolation methods: Isolation Forest, K-Nearest Neighbors.

Given that our data is only partially labeled (only monetary values of audited bills are known), we use this subset to tune model weights. As new patterns of inconsistency emerge monthly, the MAB Rashomon ensemble ensures that models with stronger predictive power receive higher weights, while less effective models are penalized, as shown in Figure 20. This setup also accommodates the integration of new models and maintains robustness against performance drift and degradation [43, 58].

Our initial experiments tested the hypothesis that combining models would outperform any single model in detecting inconsistencies and evaluated the monetary recovery potential of this approach. In February and March 2023, Unimed audited 31,355 hospital bills, with recorded adequacy values (in Brazilian Reais, BRL) indicating the recovery amount from inconsistencies. To establish a benchmark, we ranked bills by adequacy values to set a theoretical upper bound for recovery. We also used a baseline ensemble method where all constituent models contributed equally, allowing comparison with the MAB approach. Figure 21 illustrates the results of comparing our MAB Rashomon ensemble with other approaches in terms of adequacy. The baselines consisted of the theoretical maximum recovery, the voting scheme approach used previously, an oracle system selecting the best constituent before deployment, and a literature-based anomaly detection approach.

The analysis showed that reaching the 90% adequacy threshold - the minimum level for practical applicability - required 2,178 bills by the theoretically optimal ranking. In comparison, the Multi-Armed Bandit (MAB) ranking needed 10,219 bills, while the baseline voting scheme required 14,190 bills to achieve the same threshold. Despite these differences, all methods significantly outperformed traditional anomaly detection approaches such as isolation forests, as well as a naive baseline in which bills are audited in order of arrival, configuring a system without a priority ranking. Using the MAB Rashomon ensemble could reduce the number of bills audited by nearly 62% from the naive approach, from over 26,000 to approximately 10,000, while meeting the 90% adequacy threshold.

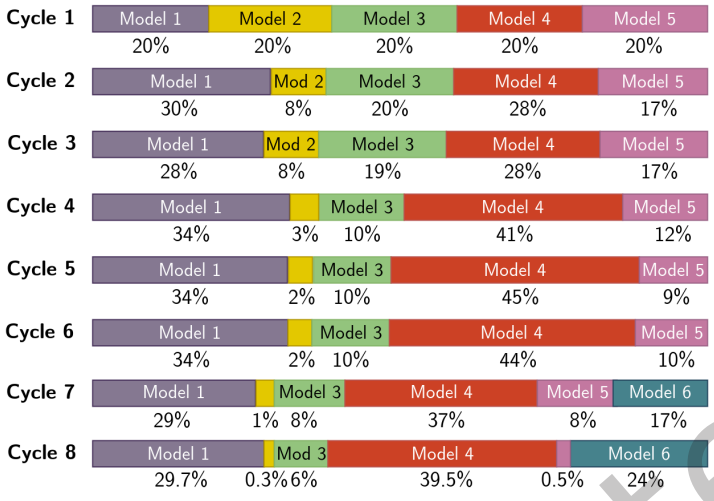


Fig. 20. Multi-armed bandit approach to assign ensemble weights. Models with inferior performance are penalized, and new models can be added.

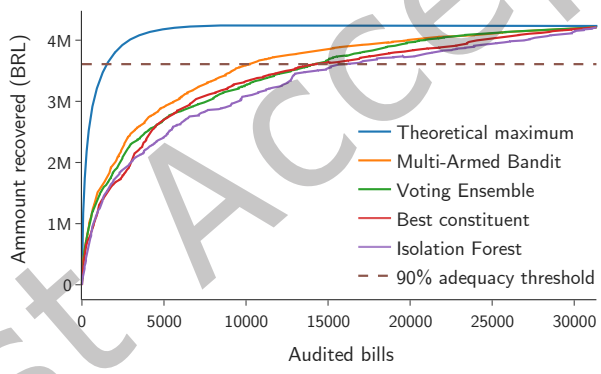


Fig. 21. Performance of the proposed approach in Brazilian Reais (BRL). The ensemble requires 28% fewer bills to reach the 90% adequacy threshold compared to the best individual model.

We proceeded to deploy our approach in Unimed's pipeline under specific constraints. The model was trained on bills from the previous three months, and daily rankings of medical bills were generated for auditing. To account for processing delays, we use a D-3 sliding-window rule, which generates the final daily ranking using the last three days' worth of bills. We accompanied the results of this deployment for 5 months, from April to August 2022. During this period, the solution alone recovered R$1,571,146 (USD 349,610) that would otherwise have been lost. From April to August, Unimed's current rule-based system and the Rashomon solution analyzed 8,570 bills. Of these, 3,327 were flagged by our algorithm, identifying 665 more inconsistent bills than the rule-based system. Unimed's business area decided that all bills flagged by the algorithm should be audited monthly, although only an average attainment rate of 89.9% was achieved due to workforce constraints. However, this resulted in an

increase of 63.4% in the number of bills flagged for auditing, with 39% of these bills being flagged exclusively by our algorithm. The monetary recovery increased by 38%, with the highest adequacy values aligning with the months where the auditors prioritized the bills recommended by our solution. These results further demonstrate the robustness and effectiveness of the Rashomon ensembles in real-world problems.

7 CONCLUSION

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 constituents' behavior may diverge when faced with data from distributions that do not match the one seen in training.

An extensive evaluation across various tasks demonstrated the effectiveness of our Rashomon ensemble, especially in scenarios with multiple local structures. In such cases, our approach consistently outperformed other state-of-the-art tree-based ensembling techniques, showcasing the capabilities of our approach. Even in scenarios in which local structures were less prevalent, our ensembles proved robust, maintaining high-performance levels. This adaptability becomes particularly valuable when predicting in domains where the inherent data generation functions may differ from what was observed during training.

Nevertheless, we acknowledge instances where our approach faced challenges. Specifically, when the Rashomon ratio - the proportion of models retained in the Rashomon Set - was relatively small. The scarce diversity among the constituent models limited performance gains. This finding emphasizes the importance of carefully considering the Rashomon ratio and the diversity of explanatory factors. While our approach consistently achieved superior results in most of our experiments, care should be exercised when deploying it in domains with a low Rashomon ratio. Alternative ensemble methods or model selection strategies may be more suitable in such cases. Therefore, we stress the importance of understanding each problem domain and evaluating the ensemble's diversity and performance before deployment. Our experiments have also revealed a direct relationship between model agreement and prediction accuracy.

Finally, we emphasize the importance of expert input in refining the final model and sets of variables, which resulted in a patent [110]. Due to our focus on explicability, we observed that when inducing our Rashomon Ensembles, some expert-known patterns frequently emerged among the constituents. This not only helped gain the experts' trust but also led to more insightful discussions regarding the remaining learned patterns. This was a main factor in achieving a tangible impact on business and a core aspect of the patent. Its applicability is demonstrated in both the stainless steel case study, which resulted in significant improvements in production processes, and the Unimed one, with gains of over *R*$1.5 million across 5 months.

CODE AND DATA AVAILABILITY

The code used for all machine learning analyses, made available for non-commercial use, has been deposited at https://doi.org/10.6084/m9.figshare.30081913 [121]. The datasets employed in this study are accessible as follows:

- **Open datasets:** Available directly from the UCI Machine Learning Repository [5] and the OpenML database [12].
- ***FAPESP* COVID-19 datasets:** Accessible upon request via covid19datasharing@fapesp.br.
- **Alzheimer datasets:** See details in [103].
- ***APERAM South America* stainless steels datasets:** See details in [126].
- ***Grupo Fleury* COVID-19 datasets:** See details in [122].
- **Brazilian energy datasets:** See details in [124].

- ***Unimed-BH* medical bills datasets:** See details in [127].

All datasets were used strictly in accordance with their respective terms of use, and any restrictions on data redistribution are noted in the corresponding references.

ACKNOWLEDGEMENTS

This work was funded by the authors' individual grants from Kunumi. *APERAM South America*, *Grupo Fleury*, and *Unimed-BH* kindly granted access to the data needed for the development of the case studies in this work.

REFERENCES

[1] S Aeberhard, D Coomans, and O De Vel. 1992. *Comparison of classifiers in high dimensional settings. Dept Math Statist, James Cook Univ, North Queensland.* Technical Report. Australia. Tech Rep.

[2] D Aha and Dennis Kibler. 1988. Instance-based prediction of heart-disease presence with the Cleveland database. *University of California* 3, 1 (1988), 3–2.

[3] Izuwa Ahanor, Hugh Medal, and Andrew C Trapp. 2022. DiversiTree: Computing Diverse Sets of Near-Optimal Solutions to Mixed-Integer Optimization Problems. *arXiv preprint arXiv:2204.03822* (2022).

[4] Sylvain Arlot and Alain Celisse. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys* 4, none (2010), 40 – 79.

[5] Arthur Asuncion and David Newman. 2007. UCI machine learning repository.

[6] Eduardo Avila, Alessandro Kahmann, Clarice Alho, and Marcio Dorn. 2020. Hemogram data as a tool for decision-making in COVID-19 management: applications to resource scarcity scenarios. *PeerJ* 8 (2020), e9482.

[7] Manuel Baena-Garcıa, José del Campo-Ávila, Raúl Fidalgo, Albert Bifet, R Gavalda, and Rafael Morales-Bueno. 2006. Early drift detection method. In *Fourth international workshop on knowledge discovery from data streams*, Vol. 6. 77–86.

[8] Chris Ballard and Wenjia Wang. 2016. Dynamic ensemble selection methods for heterogeneous data mining. In *2016 12th World Congress on Intelligent Control and Automation (WCICA)*. IEEE, 1021–1026.

[9] Abhirup Banerjee, Surajit Ray, Bart Vorselaars, Joanne Kitson, Michail Mamalakis, Simonne Weeks, Mark Baker, and Louise S Mackenzie. 2020. Use of machine learning and artificial intelligence to predict SARS-CoV-2 infection from full blood counts in a population. *International immunopharmacology* 86 (2020), 106705.

[10] Leoneros Acosta Barbosa, Geraldo André Fagundes, Leila Teichmann, and Afonso Reguly. 2007. Evaluation of sliver surface defects in cold-drawn steel bars. *Tecnologia em Metalurgia, Materiais e Mineração* 3, 4 (2007), 59.

[11] Cesar Endrigo Alves Bardelin. 2004. *Os efeitos do racionamento de energia elétrica ocorrido no Brasil em 2001 e 2002 com ênfase no consumo de energia elétrica.* Ph. D. Dissertation. Universidade de São Paulo.

[12] Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Frank Hutter, Michel Lang, Rafael G Mantovani, Jan N van Rijn, and Joaquin Vanschoren. 2017. OpenML benchmarking suites and the OpenML100. *stat* 1050 (2017), 11.

[13] R.K. Bock, A. Chilingarian, M. Gaug, F. Hakl, T. Hengstebeck, M. Ji rina, J. Klaschka, E. Kotr c, P. Savický, S. Towers, A. Vaiciulis, and W. Wittek. 2004. Methods for multidimensional event classification: a case study using images from a Cherenkov gamma-ray telescope. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 516, 2 (2004), 511–528.

[14] Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. 2020. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 1–8.

[15] Leo Breiman. 2001. Random Forests. *Mach. Learn.* 45, 1 (Oct. 2001), 5–32.

[16] Leo Breiman. 2001. Statistical modeling: the two cultures. *Statistical Science. A Review Journal of the Institute of Mathematical Statistics* 16, 3 (2001), 199–231. With comments and a rejoinder by the author.

[17] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees.* Wadsworth and Brooks, Monterey, CA.

[18] Phuong Bui Thi Mai. 2021. Underspecification in Deep Learning. (2021).

[19] J Quiñonero Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2009. Dataset shift in machine learning. *The MIT Press* 1 (2009), 5.

[20] Changjian Chen, Jun Yuan, Yafeng Lu, Yang Liu, Hang Su, Songtao Yuan, and Shixia Liu. 2021. OoDAnalyzer: Interactive Analysis of Out-of-Distribution Samples. *IEEE Trans. Vis. Comput. Graph.* 27, 7 (2021), 3335–3349.

[21] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. 2022. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 295–305.

[22] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[23] Camila Ferreira Costa and Mario A. Nascimento. 2016. IDA 2016 Industrial Challenge: Using Machine Learning for Predicting Failures. In *Advances in Intelligent Data Analysis XV - 15th International Symposium, IDA 2016, Stockholm, Sweden, October 13-15, 2016, Proceedings (Lecture Notes in Computer Science, Vol. 9897)*, Henrik Boström, Arno J. Knobbe, Carlos Soares, and Panagiotis Papapetrou (Eds.). 381–386.

[24] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. 2020. Underspecification Presents Challenges for Credibility in Modern Machine Learning. arXiv:2011.03395 [cs.LG]

[25] Emilie Danna, Mary Fenelon, Zonghao Gu, and Roland Wunderling. 2007. Generating multiple solutions for mixed integer programming problems. In *International Conference on Integer Programming and Combinatorial Optimization*. Springer, 280–294.

[26] M Davies. 1959. The relationship between weather and electricity demand. *Proceedings of the IEE-Part C: Monographs* 106, 9 (1959), 27–37.

[27] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation* 6, 2 (2002), 182–197.

[28] Krzysztof Dembczyński, Wojciech Kotłowski, and Roman Słowiński. 2008. A general framework for learning an ensemble of decision rules. In *From Local Patterns to Global Models ECML/PKDD 2008 Workshop*.

[29] VMCH Dias, Marcelo Carneiro, Cláudia Fernanda de Lacerda Vidal, Mirian de Freitas Dal Ben Corradi, Denise Brandão, Clóvis Arns da Cunha, Alberto Chebabo, Priscila Rosalba Domingos de Oliveira, Lessandra Michelin, Jaime Luis Lopes Rocha, et al. 2020. Orientações sobre diagnóstico, tratamento e isolamento de pacientes com COVID-19. *Journal Infection Control* 9, 2 (2020), 56–75.

[30] Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*. Springer, 1–15.

[31] Yuhe Ding, Lijun Sheng, Jian Liang, Aihua Zheng, and Ran He. 2023. Proxymix: Proxy-based mixup training with label refinery for source-free domain adaptation. *Neural Networks* 167 (2023), 92–103.

[32] Jiayun Dong and Cynthia Rudin. 2020. Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence* 2, 12 (2020), 810–824.

[33] Francisco G Emmerich. 2021. Comparisons between the Neighboring States of Amazonas and Pará in Brazil in the Second Wave of COVID-19 Outbreak and a Possible Role of Early Ambulatory Treatment. *International journal of environmental research and public health* 18, 7 (2021), 3371.

[34] Dominik Maria Endres and Johannes E Schindelin. 2003. A new metric for probability distributions. *IEEE Transactions on Information theory* 49, 7 (2003), 1858–1860.

[35] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. 2021. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020* (2021), 877–894.

[36] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 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 (2019), 177:1–177:81.

[37] Raymond Fisman, Sheena S. Iyengar, Emir Kamenica, and Itamar Simonson. 2006. Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment*. *The Quarterly Journal of Economics* 121, 2 (05 2006), 673–697.

[38] David Foldes, Richard Hinton, Siamak Arami, and Barbara J Bain. 2020. Plasmacytoid lymphocytes in SARS-CoV-2 infection (Covid-19). *American Journal of Hematology* (2020).

[39] Vincenzo Formica, Marilena Minieri, Sergio Bernardini, Marco Ciotti, Cartesio D'Agostini, Mario Roselli, Massimo Andreoni, Cristina Morelli, Giusy Parisi, Massimo Federici, et al. 2020. Complete blood count might help to identify subjects with high probability of testing positive to SARS-CoV-2. *Clinical Medicine* 20, 4 (2020), e114.

[40] Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* 55, 1 (1997), 119–139.

[41] Isvani Frias-Blanco, José del Campo-Ávila, Gonzalo Ramos-Jimenez, Rafael Morales-Bueno, Agustin Ortiz-Diaz, and Yaile Caballero-Mota. 2014. Online and non-parametric drift detection methods based on Hoeffding's bounds. *IEEE Transactions on Knowledge and Data Engineering* 27, 3 (2014), 810–823.

[42] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research* 17, 59 (2016), 1–35.

[43] Mike Gashler, Christophe Giraud-Carrier, and Tony Martinez. 2008. Decision tree ensemble: Small heterogeneous is better than large homogeneous. In *2008 Seventh international conference on machine learning and applications*. IEEE, 900–905.

[44] Christos Giannakopoulos and Basil E Psiloglou. 2006. Trends in energy load demand for Athens, Greece: weather and non-weather related factors. *Climate research* 31, 1 (2006), 97–108.

[45] Paulo M. Gonçalves, Silas G.T. de Carvalho Santos, Roberto S.M. Barros, and Davi C.L. Vieira. 2014. A comparative study on concept drift detectors. *Expert Systems with Applications* 41, 18 (2014), 8144–8156.
[46] Henrik Grosskreutz. 2008. Cascaded subgroups discovery with an application to regression. In *Proc. ECML/PKDD*, Vol. 5211. Citeseer, 33.
[47] MARIA CAROLINA R GULLO. 2020. A economia na pandemia Covid-19: algumas considerações. *Rosa dos Ventos* 12, Esp. 3 (2020), 1–8.
[48] Aditya Gupta, Ishwari Singh Rajput, Gunjan, Vibha Jain, and Soni Chaurasia. 2022. NSGA-II-XGB: Meta-heuristic feature selection with XGBoost framework for diabetes prediction. *Concurrency and Computation: Practice and Experience* 34, 21 (2022), e7123.
[49] Isabelle Guyon, Steve Gunn, Asa Ben-Hur, and Gideon Dror. 2004. Result analysis of the NIPS 2003 feature selection challenge. *Advances in neural information processing systems* 17 (2004), 545–552.
[50] Thomas Hale, Anna Petherick, Toby Phillips, and Samuel Webster. 2020. Variation in government responses to COVID-19. *Blavatnik School working paper* (2020).
[51] Peter Hart, Nils Nilsson, and Bertram Raphael. 1968. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics* 4, 2 (1968), 100–107. https://doi.org/10.1109/tssc.1968.300136
[52] Daihai He, Guihong Fan, Xueying Wang, Yingke Li, and Zhihang Peng. 2021. The new SARS-CoV-2 variant and reinfection in the resurgence of COVID-19 outbreaks in Manaus, Brazil. *medRxiv* (2021).
[53] Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. 2019. Deep Anomaly Detection with Outlier Exposure. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.
[54] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. 2020. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society* 146, 730 (2020), 1999–2049.
[55] James Hinns, Xiuyi Fan, Siyuan Liu, Veera Raghava Reddy Kovvuri, Mehmet Orcun Yalcin, and Markus Roggenbach. 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, 323–335.
[56] Ching-Lai Hor, Simon J Watson, and Shanti Majithia. 2005. Analyzing the impact of weather variables on monthly electricity demand. *IEEE transactions on power systems* 20, 4 (2005), 2078–2085.
[57] Ben Hu, Hua Guo, Peng Zhou, and Zheng-Li Shi. 2020. Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology* (2020), 1–14.
[58] Faliang Huang, Guoqing Xie, and Ruliang Xiao. 2009. Research on ensemble learning. In *2009 International Conference on Artificial Intelligence and Computational Intelligence*, Vol. 3. IEEE, 249–252.
[59] YJ Huang, R Ritschard, J Bull, and L Chang. 1986. *Climatic indicators for estimating residential heating and cooling loads*. Technical Report. Lawrence Berkeley Lab., CA (USA).
[60] David S Hui, Esam I Azhar, Tariq A Madani, Francine Ntoumi, Richard Kock, Osman Dar, Giuseppe Ippolito, Timothy D Mchugh, Ziad A Memish, Christian Drosten, et al. 2020. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health – The latest 2019 novel coronavirus outbreak in Wuhan, China. *International Journal of Infectious Diseases* 91 (2020), 264–266.
[61] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 3146–3154.
[62] Daniel Kifer, Shai Ben-David, and Johannes Gehrke. 2004. Detecting change in data streams. In *VLDB*, Vol. 4. Toronto, Canada, 180–191.
[63] Melih Kirlidog and Cuneyt Asuk. 2012. A fraud detection approach with data mining in health insurance. *Procedia-Social and Behavioral Sciences* 62 (2012), 989–994.
[64] Nicholas Kissel and Lucas Mentch. 2021. Forward stability and model path selection. *arXiv preprint arXiv:2103.03462* (2021).
[65] Arno Knobbe and Joris Valkonet. 2009. Building classifiers from pattern teams. In *Proceedings of the ECML PKDD'09 workshop LeGo*. Citeseer, 77–93.
[66] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. 2017. Identifying Unknown Unknowns in the Open World: Representations and Policies for Guided Exploration. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (San Francisco, California, USA) *(AAAI'17)*. AAAI Press, 2124–2132.
[67] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6405–6416.
[68] Jian Liang, Ran He, and Tieniu Tan. 2024. A Comprehensive Survey on Test-Time Adaptation Under Distribution Shifts. *Int. J. Comput. Vision* 133, 1 (July 2024), 31–64.
[69] Jian Liang, Dapeng Hu, and Jiashi Feng. 2020. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*. PMLR, 6028–6039.
[70] Weixin Liang, Yining Mao, Yongchan Kwon, Xinyu Yang, and James Zou. 2023. Accuracy on the curve: On the nonlinear correlation of ml performance between data subpopulations. In *International Conference on Machine Learning*. PMLR, 20706–20724.

[71] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*. PMLR, 97–105.
[72] Mehmet E Lorasdagi, Ahmet B Koc, Ali T Koc, and Suleyman S Kozat. 2025. Fitting Multiple Machine Learning Models With Performance Based Clustering. *IEEE Signal Processing Letters* (2025).
[73] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. 2019. Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering* 31, 12 (2019), 2346–2363. https://doi.org/10.1109/TKDE.2018.2876857
[74] Ning Lu, Guangquan Zhang, and Jie Lu. 2014. Concept drift detection via competence models. *Artificial Intelligence* 209 (2014), 11–28.
[75] Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Annual Conf. on Neural Information Processing Systems*. 4768–4777.
[76] David JC MacKay. 2003. *Information theory, inference and learning algorithms*. Cambridge university press.
[77] David Madras, James Atwood, and Alexander D'Amour. 2020. Detecting Extrapolation with Local Ensembles. In *International Conference on Learning Representations*.
[78] Debora Maia-Silva, Rohini Kumar, and Roshanak Nateghi. 2020. The critical role of humidity in modeling summer electricity demand across the United States. *Nature communications* 11, 1 (2020), 1–8.
[79] Hassan H Malik and John R Kender. 2008. Classification by pattern-based hierarchical clustering. In *From Local Patterns to Global Models Workshop, ECML/PKDD*. 1–18.
[80] Jose A Marengo, Tercio Ambrizzi, Naurinete Barreto, Ana Paula Cunha, Andrea M Ramos, Milagros Skansi, Jorge Molina Carpio, and Roberto Salinas. 2022. The heat wave of October 2020 in central South America. *International Journal of Climatology* 42, 4 (2022), 2281–2298.
[81] Sascha Marton, Stefan Lüdtke, Christian Bartelt, and Heiner Stuckenschmidt. 2024. GRANDE: Gradient-Based Decision Tree Ensembles for Tabular Data. In *The Twelfth International Conference on Learning Representations*.
[82] Charles T. Marx, Flavio Du Pin Calmon, and Berk Ustun. 2020. Predictive Multiplicity in Classification. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*. JMLR.org, Article 628, 10 pages.
[83] Song Mei and Andrea Montanari. 2019. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics* (2019).
[84] Leandro L Minku and Xin Yao. 2011. DDD: A new ensemble approach for dealing with concept drift. *IEEE transactions on knowledge and data engineering* 24, 4 (2011), 619–633.
[85] Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. 2012. A unifying view on dataset shift in classification. *Pattern Recognition* 45, 1 (2012), 521–530.
[86] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain generalization via invariant feature representation. In *International conference on machine learning*. PMLR, 10–18.
[87] Yilin Ning, Marcus Eng Hock Ong, Bibhas Chakraborty, Benjamin Alan Goldstein, Daniel Shu Wei Ting, Roger Vaughan, and Nan Liu. 2022. Shapley variable importance cloud for interpretable machine learning. *Patterns* 3, 4 (2022), 100452.
[88] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. 2022. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*. PMLR, 16888–16905.
[89] Manuel Olave, Vladislav Rajkovic, and Marko Bohanec. 1989. An application for admission in public school systems. *Expert Systems in Public Administration* 1 (1989), 145–160.
[90] ONS. 2018. Histórico Da Operação Instalada, Carga de Energia. *Operador Nacional do Sistema Elétrico, Rio de Janeiro* (2018).
[91] Guillermo Ortiz-Jiménez, Itamar Franco Salazar-Reque, Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. 2021. A neural anisotropic view of underspecification in deep learning. *CoRR* abs/2104.14372 (2021). arXiv:2104.14372
[92] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. 2010. Domain adaptation via transfer component analysis. *IEEE transactions on neural networks* 22, 2 (2010), 199–210.
[93] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, Inc., 6638–6648. http://papers.nips.cc/paper/7898-catboost-unbiased-boosting-with-categorical-features.pdf
[94] Sanqing Qu, Guang Chen, Jing Zhang, Zhijun Li, Wei He, and Dacheng Tao. 2022. Bmd: A general class-balanced multicentric dynamic prototype strategy for source-free domain adaptation. In *European conference on computer vision*. Springer, 165–182.
[95] Sergio Ramírez-Gallego, Bartosz Krawczyk, Salvador García, Michał Woźniak, and Francisco Herrera. 2017. A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing* 239 (2017), 39–57.
[96] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *AAAI Conf. on Artificial Intelligence*. 1527–1535.
[97] Hannah Ritchie, Edouard Mathieu, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Esteban Ortiz-Ospina, Joe Hasell, Bobbie Macdonald, Diana Beltekian, and Max Roser. 2020. Coronavirus Pandemic (COVID-19). *Our World in Data* (2020).
[98] Ando Saabas. 2014. Interpreting random forests. *Diving into data* 24 (2014).

[99] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. 2020. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems* 33 (2020), 11539–11551.
[100] Lesia Semenova, Cynthia Rudin, and Ronald Parr. 2022. On the Existence of Simpler Machine Learning Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1827–1858.
[101] Lloyd S Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games* 2, 28 (1953), 307–317.
[102] Galit Shmueli. 2010. To explain or to predict? *Statist. Sci.* 25, 3 (2010), 289–310.
[103] Ismael Santana Silva and Adriano Veloso. 2022. Automatic Model Evaluation using Feature Importance Patterns on Unlabeled Data. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.
[104] Elena Caires Silveira. 2020. Prediction of COVID-19 From Hemogram Results and Age Using Machine Learning. *Frontiers in Health Informatics* 9, 1 (2020), 39.
[105] Andrew AS Soltan, Samaneh Kouchaki, Tingting Zhu, Dani Kiyasseh, Thomas Taylor, Zaamin B Hussain, Tim Peto, Andrew J Brent, David W Eyre, and David A Clifton. 2020. Rapid triage for COVID-19 using routine clinical data for patients attending hospital: development and prospective validation of an artificial intelligence screening test. *The Lancet Digital Health* (2020).
[106] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. 2014. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international* 2014 (2014).
[107] W Nick Street, William H Wolberg, and Olvi L Mangasarian. 1993. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization*, Vol. 1905. International Society for Optics and Photonics, 861–870.
[108] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474* (2014).
[109] Guido van Capelleveen, Mannes Poel, Roland M Mueller, Dallas Thornton, and Jos van Hillegersberg. 2016. Outlier detection in healthcare fraud: A case study in the Medicaid dental domain. *International journal of accounting information systems* 21 (2016), 18–31.
[110] ADRIANO Veloso, PAULO Caramelli, KARINA BRAGA GOMES Borges, DANIELLA Araújo, GIANLUCCA Zuin, TIAGO HENRIQUE Alves, and NIVIO Ziviani. 2023. PROCESSO CENTRADO NO HUMANO PARA ELABORAçãO DE MODELOS BASEADOS EM APRENDIZADO DE MáQUINA E USOS. Brazil patent BR 102021015411-0 A8.
[111] Mark C Walters and Herbert T Abelson. 1996. Interpretation of the complete blood count. *Pediatric Clinics* 43, 3 (1996), 599–622.
[112] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*.
[113] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. 2022. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7201–7211.
[114] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. 2020. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566* (2020).
[115] Jason Woods, Nelson James, Eric Kozubal, Eric Bonnema, Kristin Brief, Liz Voeller, and Jessy Rivest. 2022. Humidity's impact on greenhouse gas emissions from air conditioning. *Joule* 6, 4 (2022), 726–741.
[116] Joseph T Wu, Kathy Leung, and Gabriel M Leung. 2020. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet* 395, 10225 (2020), 689–697.
[117] Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. 2022. Exploring the Whole Rashomon Set of Sparse Decision Trees. *arXiv preprint arXiv:2209.08040* (2022).
[118] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1Ddp1-Rb
[119] Marvin Zhang, Sergey Levine, and Chelsea Finn. 2022. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems* 35 (2022), 38629–38642.
[120] Weijia Zhang and Xiaofeng He. 2017. An anomaly detection method for medicare fraud detection. In *2017 IEEE International Conference on Big Knowledge (ICBK)*. IEEE, 309–314.
[121] Gianlucca Zuin. 2025. Code for "A 6 or a 9?": Ensemble Learning Through the Multiplicity of Performant Models and Explanations. (2025). https://doi.org/10.6084/m9.figshare.30081913
[122] Gianlucca Zuin, Daniella Araujo, Vinicius Ribeiro, Maria Gabriella Seiler, Wesley Heleno Prieto, Maria Carolina Pintão, Carolina dos Santos Lazari, Celso Francisco Hernandes Granato, and Adriano Veloso. 2022. Prediction of SARS-CoV-2-positivity from million-scale complete blood counts using machine learning. *Communications medicine* 2, 1 (2022), 1–12.
[123] Gianlucca Zuin, Rob Buechler, Tao Sun, Chad Zanocco, Daniella Castro, Adriano Veloso, and Ram Rajagopal. 2022. 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)*. 1–5.

[124] Gianlucca Zuin, Rob Buechler, Tao Sun, Chad Zanocco, Francisco Galuppo, Adriano Veloso, and Ram Rajagopal. 2023. Extreme event counterfactual analysis of electricity consumption in Brazil: Historical impacts and future outlook under climate change. *Energy* (2023), 128101.

[125] G. Zuin, L. Chaimowicz, and A. Veloso. 2020. Deep Learning Techniques for Explainable Resource Scales in Collectible Card Games. *IEEE Transactions on Games* (2020), 1–1. https://doi.org/10.1109/TG.2020.3030742

[126] Gianlucca Zuin, Felipe Marcelino, Lucas Borges, João Couto, Victor Jorge, Mychell Laurindo, Glaucio Barcelos, Marcio Cunha, Valdeci Alvarenga, Henrique Rodrigues, et al. 2021. Predicting Heating Sliver in Duplex Stainless Steels Manufacturing through Rashomon Sets. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[127] Gianlucca Zuin, Lucas Parreiras, Luiz Melo, Gabriel Barros, Humberto Lomeu, Batielhe Melo, Wesley Marini, Debora Lott, and Mateus De Souza. 2023. An ensemble approach for inconsistency detection in medical bills: A case study. In *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 573–578.

[128] Gianluca Zuin, Adriano Veloso, João Cândido Portinari, and Nivio Ziviani. 2020. Automatic Tag Recommendation for Painting Artworks Using Diachronic Descriptions. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[129] Gianlucca Lodron Zuin. 2023. *Ensemble Learning through Rashomon Sets.* PhD thesis. Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.