UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Gianlucca Lodron Zuin

Ensemble Learning through Rashomon Sets

Belo Horizonte
2023

Gianlucca Lodron Zuin

Ensemble Learning through Rashomon Sets

Final Version

Dissertation presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Advisor: Adriano Veloso

Belo Horizonte
2023

Zuin, Gianlucca Lodron.

Z94e Ensemble learning through Rashomon sets [recurso eletrônico] : / Gianlucca Lodron Zuin – 2023.
1 recurso online (161 f. il, color.) : pdf.

Orientador: Adriano Alonso Veloso.

Tese (Doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciências da Computação.
Referências: f.139-161

1. Computação – Teses. 2. Aprendizado do computador – Teses. 3. Efeito Rashomon – Teses. I. Veloso, Adriano Alonso. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Computação. III.Título.

CDU 519.6*82(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Ensemble Learning through Rashomon Sets

GIANLUCCA LODRON ZUIN

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. ADRIANO ALONSO VELOSO - Orientador
Departamento de Ciência da Computação - UFMG

PROF. WAGNER MEIRA JÚNIOR
Departamento de Ciência da Computação - UFMG

Nivio Ziviani

PROF. NIVIO ZIVIANI
Departamento de Ciência da Computação - UFMG

Prof. PAULO NAJBERG ORENSTEIN
Instituto de Matemática Pura e Aplicada - IMPA

Ram Rajagopal

Prof. RAM RAJAGOPAL
Stanford University

Prof. RAFAEL BORDINI
Faculdade de Informática - PUCRS

Belo Horizonte, 05 de janeiro de 2023.

Acknowledgments

I want to express my gratitude to my professors who guided me on this journey, for their valuable teachings and promising opportunities. In particular, I would like to thank Professor Adriano, which instructed me in algorithms and data structures back in 2010 and has taught me so much ever since, being my advisor during my Master's and this Doctorate.

To my parents, Romanelli and Elenice, not only for providing the conditions that allowed me to focus on my research, but also for always being great examples. I also thank you for all the help provided during the writing of this work, allowing me to take advantage of your vast experience. From my mother, I inherited the approach to Mathematics, and from my father and my uncle Ronaro, the path of Computer Science.

To my grandmother Leila, for her love, support, and encouragement throughout my academic life.

I want to thank my friends and my girlfriend Karen, for always providing me with good smiles and moments of joy. Their little distractions often allowed me to look at things from other perspectives and come up with new solutions.

Finally, to CAPES, FAPEMIG, and all the PPGCC staff. CAPES and FAPEMIG provided the financial support that allowed me to dedicate myself to the Doctorate program.

This work is also dedicated to those who supported and helped me during this journey. You provided me with the vital conditions for the making of this work and I believe that without you none of this would have been possible. To all of you, I am extremely grateful.

“*Essentially, all models are wrong, but some are useful.*”

(George Edward Pelham Box)

Resumo

Criar modelos a partir de observações e garantir a eficácia em novos dados é a essência do aprendizado de máquina. Portanto, estimar o erro de generalização de um modelo é um passo crucial. Apesar da existência de muitas métricas de desempenho que aproximam o poder de generalização, ainda é um desafio selecionar modelos que generalizem para dados futuros desconhecidos. Neste trabalho, investigamos como os modelos se comportam em conjuntos de dados que possuam diferentes funções geradoras, mas constituem tarefas correlatas. A principal motivação é estudar o Efeito Rashomon, que aparece sempre que o problema de aprendizagem admite um conjunto de soluções que apresentam desempenho semelhante. Muitos problemas do mundo real são caracterizados por múltiplas estruturas locais no espaço de dados e, como resultado, o problema de aprendizagem correspondente apresenta uma superfície de erro não convexa sem mínimo global óbvio, implicando assim uma multiplicidade de modelos performantes, cada um deles fornecendo uma explicação diferente. A literatura sugere este tipo de problema estar sujeito ao Efeito Rashomon. Por meio de um estudo empírico em diferentes conjuntos de dados, elaboramos uma estratégia focada na explicabilidade, especificamente na importância de variáveis. Nossa abordagem para lidar com o Efeito Rashomon é estratificar, durante o treinamento, modelos em grupos que sejam coerentes entre si ou contrastantes. A partir desses grupos, podemos selecionar modelos que aumentem a robustez das respostas em tempo de produção, sendo também capazes de medir possíveis desvios nos dados. Apresentamos ganhos de desempenho na maioria dos cenários avaliados ao criar um comitê de modelos e garantir que cada constituinte cubra um subespaço independente da solução. Validamos nossa abordagem em conjuntos de dados fechados e abertos, fornecendo intuições sobre possíveis aplicações ao analisar alguns estudos de caso do mundo real nos quais nosso método foi empregado com sucesso. Não apenas nossa abordagem provou ser superior ao estado-da-arte a comitês baseados em árvores, com ganhos em AUC de até 0,20+, mas os constituintes são altamente explicáveis e permitem a integração de humanos no processo de tomada de decisão do modelo, assim os tornando mais eficientes.

Palavras-chave: Rashomon Effect, Ensemble Learning, Data Drift

Abstract

Creating models from previous observations and ensuring effectiveness on new data is the essence of machine learning. Therefore, estimating the generalization error of a trained model is a crucial step. Despite the existence of many capacity measures that approximate the generalization power of trained models, it is still challenging to select models that generalize to future data. In this work, we investigate how models perform in datasets that have different underlying generator functions but constitute co-related tasks. The key motivation is to study the Rashomon Effect, which appears whenever the learning problem admits a set of models that all perform roughly equally well. Many real-world problems are characterized by multiple local structures in the data space and, as a result, the corresponding learning problem has a non-convex error surface with no obvious global minimum, thus implying a multiplicity of performant models, each of them providing a different explanation, which literature suggests to being subject to the Rashomon Effect. Through an empirical study across different datasets, we devise a strategy focusing primarily on model explainability (i.e., feature importance). Our approach to deal with the Rashomon Effect is to stratify, during training, models into groups that are either coherent or contrasting. From these Rashomon groups, we can select models that increase the robustness of the production responses along with means to gauge data drift. We present performance gains on most of the evaluated scenarios by locating these models and creating an ensemble guaranteeing that each constituent covers an independent solution sub-space. We validate our approach by performing a series of experiments in both closed and open-source benchmark suites and give insights into the possible applications by analyzing real-world case studies in which our framework was employed with success. Not only does our approach prove to be superior to state-of-the-art tree-based ensembling techniques, with gains in AUC of up to .20+, but the constituent models are highly explainable and allow for the integration of humans into the decision-making pipeline, thus empowering them.

Keywords: Rashomon Effect, Ensemble Learning, Data Drift

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Akira Kurosawa (1910-1998) was a Japanese film director, screenwriter, and producer who directed 30 films in a career spanning 57 years. He is regarded as one of the most influential filmmakers in the history of cinema [Davis et al., 2015, Prince, 1999]. One of his greatest movies was entitled *Rashomon* [Kurosawa, 1950], which premiered in 1950 in Tokyo and became the winner of the Golden Lion at the 12th International Venice Film Festival in 1951.

In the film, Kurosawa deals with the impossibility of objective truth by portraying a mysterious crime reported by four witnesses who contradict themselves from different points of view. A terrible crime occurs in a forest: a samurai was murdered by a bandit and his wife raped in front of him before he was killed - these are supposedly the facts from which four witnesses are heard in a court. The only sure truth in this story is the fact that the samurai is dead. Everything else is uncertain. Any conclusion about what transpired in the forest is impossible to assert. This happens because the versions relayed by the witness are antagonistic and divergent. Each of the four witnesses tells a story that is either the least compromising or best favors themselves. The narrative ends with the testimonies, without any manifestation from the judges. Given the difficulty of searching for objective truth based on biased reports exhibited by the witnesses, the job of reaching a verdict is left entirely the responsibility of the viewer.

Kurosawa's construction impressed philosophers. In tribute to the movie, the *Rashomon Effect* was coined as the difficulty of detecting the truth of a fact through the testimonies of several witnesses as a relationship between epistemology and subjectivity. That is the effect in which the same objective epistemological fact causes different interpretations by the subjective perspective launched by each of the observers [Dorland, 2016]. Although the movie reached western audiences several years ago in 1951, the term is still employed up to this date. Justice Applegarth [2020] pointed out in a recent case in the Supreme Court of Queensland:

> *"The Rashomon effect describes how parties describe an event in a different and contradictory manner, which reflects their subjective interpretation and self-interested advocacy, rather than objective truth. The Rashomon effect is*

> *evident when the event is the outcome of litigation. One should not be surprised when both parties claim to have won the case."*

When studying the machine learning and statistics fields, one often comes across different models that can fit data with statistically similar performance. In this scenario, models used in an analysis may employ a completely distinctive set of factors from another, in such a way that it is not possible to draw a correlation between the models, aside from the fact both achieve similar predictive capability. This phenomenon was termed as the *Rashomon effect of statistics* by Breiman [2001a]. A common consequence is that attempting to induce a single model that encompasses all possible factors often leads to underperfomance [Pujari and Gupta, 2012].

Within the Rashomon Effect concept, Fisher et al. [2019] analyzes the set of models that contain accuracy close to the optimal model. From this set, he formally defines the concept of *Rashomon Set*, this being the subspace of the universe of models that summarize the range of effective prediction strategies that an optimal analyst might choose. Semenova and Rudin [2019] delve deeper into the theme of the Rashomon effect in machine learning, giving pertinent definitions about the generalization of the Rashomon Set as well as its format and volume. In particular, it is explored in which situations it is possible to obtain a sample of the model space such that the properties related to Rashomon in this subspace are similar to those of the sample universe.

The standard approach for model selection, adopted both in industry and research, is cross-validation. Although it usually produces robust risk estimation, it has been shown to fail for some problems depending on the goal of model selection [Arlot and Celisse, 2010], and the obtained measure of empirical risk on a test set might not directly translate to good performance in real-world applications, posing a major challenge for successful machine learning [D'Amour et al., 2020]. Although it is hard to understand the generalization power of machine learning models, a common observation is that empirical risk is significantly affected when different models perform indistinguishably well on the test set [Hinns et al., 2021].

The main problem arises when the selected model faces data drawn from a different distribution during production. The guarantees established by cross-validation do not hold for out-of-distribution data leading to unpredictable model performance and thus, the held-out performance is not a reliable risk estimation anymore. Analyzing models under additional axes, besides the empirical risk, could allow for the selection of models that increase the robustness of predictions. By definition, Rashomon set for some problem consists of the subset of all risk-equivalent functions, which are all plausible candidates for model selection. Cross-validation gives guarantees that all of these should perform similarly with high confidence on data distributions that match those of the training data. A simple approach asses new data distribution could be the evaluation of each model's prediction. However, deploying the whole Rashomon set might be intractable. Further, there could also exist functions that are so similar that changes in data distribution affect them equally.

Consider two scenarios. In the first we perform some transformation over the validation data to induce a different distribution. We will call the resulting dataset our *'transformed test'* data. The second consists of a co-related dataset to the one used in training, but that was collected either at another point in time or location. We will call this dataset our *'production'* data. In real-world applications, this could be data presented to a model once it is deployed. We can use insights regarding model behavior on this transformed test to extrapolate for possible distribution divergences during production. We may select only the models from the Rashomon set that diverge under the transformed test, thus building an ensemble and drastically reducing the number of deployed models. If, during production time, the ensemble's constituents diverge on their predictions, this should be indicative of the unreliability of output. In fact, inducing an ensemble from the performant models of the Rashomon set is promissing, as the utilization of a diverse set of robust learning algorithms has been demonstrated to be a more effective approach of ensemble learning compared to utilizing techniques that seek to reduce the complexity of the models in the interest of promoting diversity [Gashler et al., 2008].

We view diversity among individual models as crucial for gaining a broad understanding of any phenomenon. We assume that problems are not tied to a single causative factor, and that causative factors may vary depending on factors that might not be directly intuitive. Indeed, the Rashomon effect in statistics also referred to as "*the multiplicity of good models*", suggests the existence of multiple potential explanations for a given problem, all consistent with the data. To encourage diversity and identify patterns, we group models based on the similarity of their explanations. Ideally, this leads to dense groups in which models assigned to the same cluster share common explanatory factors, while dissimilarity is expressed in disjoint clusters. For each group, we select the most distinct models under the proposed transformed test, resulting in an ensemble that is diverse in its constituents, incorporates high-performing models, summarizes the entire Rashomon set and solution space, and allows for an approximation of a risk metric under new data distributions based on constituent agreement. We coin this idea as the Rashomon Ensemble. Our approach can be summarized by the following steps:

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

2. Compute the explanation vector of the sampled models and their pair-wise similarity.

3. Perturbate a held-out test data through some data transformation.

4. Compute the pair-wise distance in the transformed test set.

5. Split the Rashamon set into subgroups given models explanation vectors and distances.

6. Select a set of models with contrasting explanations and divergent predictions on the transformed data.

7. Build an ensemble and evaluate agreement to estimate reliability.

We validate our approach on a set of public datasets for reproducibility and demonstrate its robustness in simulated scenarios. We also collect four datasets to validate our hypothesis regarding model behavior under scenarios where the data generation function might be different. Our results show that Rashomon ensembles consistently outperform state-of-the-art ensemble learning approaches if the Rashomon set is large enough. When exposed to data drift, our approach remained the performant one in most evaluated scenarios providing further evidence of its reliability. We proceed to employ the Rashomon ensembles in three real-world applications partnered with various institutions and study the impact of our approach in these case studies.

1.1 Contribution

Our main premise is that explicability can be employed to induce diversity in ensembles. From this, we consider two hypotheses. The first is that we can estimate the reliability of an ensemble by exploiting the fact that some models might behave similarly only when data is drawn from the same distribution as the one seen in training. And the second is that we can find these models by looking at their divergent explanations. If these hypotheses are true, then we could verify at the production stage the output of any given instance. If the models disagree, then that would imply that data is drawn from a new unknown distribution, and we cannot trust predictions. Thus, we can establish an estimation of a production risk metric by the prediction distance from these explainability-diverse models. We empirically demonstrate these hypotheses and that we can elect suitable ensemble constituents after splitting the Rashomon set and performing a perturbation of the training data. We validate this finding in distinct problems in which we verify the existence of large Rashomon spaces. The Rashomon Effect guarantees that all models selected are performant and coherent with the data.

We tackle the identification of contrasting models to serve as ensemble constituent candidates. In complex problems, data is inherently composed of several local structures and sub-populations which might lead to poor performance by attempting to induce a single model from all sub-populations. We argue that by analyzing the Rashomon set and dividing it into subspaces is a preferable approach, and present a technique for the division of the Solution-space. By combining individuals from each of these populations, we investigate the possibility of building ensembles with highly explainable and diverse members that can answer specific parts of the problem. Further, since each constituent encompasses a different explanation for the target phenomenon, the ensemble output is directly tied to the trustworthiness of prediction. If after deployment on the real-world constituents agree, this serves as a strong indicator that the data distribution matches the one seen during training and all cross-validation guarantees hold. On the other hand, if the constituents disagree, then those proprieties cannot be trusted. As such, we also present a new strategy to ensemble learning coherent with this theory, deemed Rashomon ensembles. The most straightforward technique for combining the outputs of the constituent individuals is simple voting, albeit we also investigate stacking. A desirable characteristic of such a simple method is that the returned output of the ensemble is a direct measure of agreement and thus, an estimation of the Rashomon ensemble reliability under new data distributions.

Applied to real-world problems tackled in this work, our method presented considerable improvements to the respective industry practices. After translating model explanations to process rules, we observed a reduction in over 50% in the occurrence of heating slivers in Duplex stainless steel plates, the problem addressed in Chapter 5. We were also able to verify that Brazil is unlikely to handle its own energetic demand by 2070, 20 years prior to expert projections, while also measuring the impact on energetic consumption of different extreme events, described in Chapter 6. Finally, from the proposed experiments regarding COVID-19 automated diagnosis presented in Chapter 7, we observed a significant gap in the literature regarding virus biases in medical machine learning literature. These case studies provide empyrical evidence on the robustness of Rashomon ensemble learning.

1.2 Thesis Statement

In many situations, the data is inherently composed of several local structures and sub-populations. The traditional all-in-one approach considers the use of all data at once to induce a single model. Assuming that each local structure could be viewed as a different view of the same data, it is more difficult for an algorithm to minimize the error by considering the information of all views, in many cases contrasting. Further, assessing the risk of deploying a single model under data distributions that differ from training is challenging. This thesis aims to show, based on evidence, that in these situations it is advantageous to make use of the concept of these local structures for induction of models that are more robust and consistent with the data. We argue that each local structure can be mapped to a partition of the solution space and, by exploiting model explanation techniques, we can elect different underlying explanations for the studied phenomenon. We argue that by locating the performant models within each partition and then performing an ensembling approach, we can obtain a general view of the problem such that each ensemble constituent is explainable, covers an independent sub-space of the problem, and is resilient in the presence of data drift by providing an estimation of prediction reliability under unknown data distributions.

1.3 Thesis Structure

The remainder of the thesis is organized as follows: Chapter 2 presents a discussion of relevant related work and gives some literature background. Chapter 3 describes our approach and the insights that led to the proposed ensembling technique, followed by a series of empirical experiments in Chapter 4. Complementing this discussion, Chapters 5 to 7 describe real-world case studies in which we were able to successfully identify a large Rashomon space and learn an ensemble. Finally, Chapter 8 presents our conclusions and a summary of the study carried out, as well as the directions for future work.

Chapter 2

Background

One of the core inspirations of this work arises from the insight that a dataset might be heterogeneous. 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 represented by a data bias. This could lead to different areas of the solution space being able to similarly approximate the target label, giving rise to multiple models with comparable performance and contrasting explanations and thus inducing the Rashamon effect. If this is true, it would be more suitable if local behavior was represented by a local model, which can be incorporated into an ensemble [Zuin et al., 2021]. Sampling multiple local minima allow for an approximation of the global objective while also expanding the representation space [Dietterich, 2000]. One of the advantages of this approach is the ability evaluate the ensemble under different contexts. Since all constituents present similar performance under the train data distribution, a divergence in their behavior could be an indicator of anomalities in data, such as data drift.

2.1 Data Drift

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 be able to handle new concepts as they arrive, properly tuning itself to new data distributions. The main challenge consists in the fact that, as data drift toward these new concepts, it negatively impacts the accuracy of the models that are learned based on past training instances [Gonçalves et al., 2014]. Therefore, early data drift identification and adaptation are key aspects of such systems. Lu et al. [2019] 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 one method and another are tied to stages 3 and 4. Concerning stage 3, two of the big 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 DDD [Minku and Yao, 2011], which also establishes warning levels for the error rate to identify when the model should be retrained with data of this new concept or updated with new incoming instances. There are many other error-based methods but, as stated by Lu et al. [2019], DDD is perhaps the most-referenced method. Under their framework, other methods can be summarized by changes to some stage of drift detection to DDD, such as employing another hypothesis testing [Frias-Blanco et al., 2014] or changing some detail of the evaluated metric [Baena-Garcıa et al., 2006].

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 [Kifer et al., 2004]. 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. But there remains a problem concerning actual and virtual drift.

Let T be the train distribution of the source data, and U be some unknown distribution from another dataset. Candela et al. [2009] defines data drift as a change in the joint distribution of features. That is:

$$P(x_t, y_t) \neq P(x_u, y_u) \tag{2.1}$$

Probably approximately correct learning relies on the independent and identically distributed assumption between data distributions to estimate the empirical risk of a learning function. If we verify data drift, we cannot guarantee that the empirical risk is close to the real risk.

We can decompose $P(x, y) = P(x) \times P(y|x)$. Thus, if we verify data drift, we could assume that it might come from two sources. We can observe a change in $P(x)$ (covariate drift), or a change in $P(y|x)$ (concept drift). As stated by Moreno-Torres et al. [2012], covariate drift is tied to the distribution of a variable, while concept drift implies that the relationship between the target and predictor changes between datasets. Finally, it's possible that both $P(x)$ and $P(y|x)$ present significant differences from the original distributions, which we define 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 [Lu et al., 2014].

The decomposition of Equation 2.1 presents the sources of data drift, them being covariate and concept drift. Covariate drift is often called virtual drift due to drift in $P(x)$ not affecting the decision boundary of models [Ramírez-Gallego et al., 2017]. Learning a new model when presented with 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 shift under new data. It is important to highlight that the aforementioned approaches to drift detection are well suited specifically in 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. And sliding window data-based methods depend on the notion of temporal relationships. Further, the knowledge of novel instances' labels is necessary to differentiate between dual and virtual drift, which might not be possible in scenarios outside of online learning.

2.2 Feature Importance

Regardless of the algorithm of choice, understanding the model predictions, and giving an explanation of how the model arrived at the decision, are challenging tasks. Both legal and ethical reasons gave rise to the field of Explainable AI (XAI) research to address these challenges [Holzinger et al., 2018, Shneiderman, 2020]. A particular research subtopic is that of Human-Centered XAI (HCXAI) discussed by Ehsan and Riedl [2020], in which we place the human decision-makers at the core of the algorithmic pipeline. This sort of approach helps in building trust on the end model prediction [Weitz et al., 2019]. As such, Xu [2019] suggests that any HCXAI professional should not only strive to provide an explainable and comprehensible model, but also a tool that is both useful and usable by the greater public. As such, one of our main goals in this thesis is to provide a mechanism that can empower human beings through the lenses of explainable AI, as previously achieved in our previous work in Zuin and Veloso [2019] and Zuin et al. [2020]. Regarding the more general XAI research, there is an extensive literature in the field of measuring variable importance, in particular about tree-based models which tend to be highly explainable. One of the most commonly employed metrics is called Gini importance and was first introduced by Breiman [2001b] alongside the definition of Random Forests.

The Gini importance is a special case of a Mean Decrease Impurity (MDI), in which we add up the weighted impurity decreases for all nodes on a tree and, in the case of Random Forest, average out each feature's impurity decrease across all trees. When applied to the Gini coefficient [Gastwirth, 1972], the MDI approach produces the Gini importance result. The Gini coefficient measures inequality between frequency distributions, that is, the difference between the hypothetical straight line depicting perfect equality and the actual curve depicted by the sampled probability distribution. Another common metric employed alongside MDI is entropy or information gain. One of the key algorithms to build decision trees, the ID3 devised by Quinlan [1986] employs a top-down, greedy search through the space of possible branches and computes the information gained by each feature to decide the best split point. Thus, it is no surprise that this should have a somewhat direct relationship with feature importance in trees and both these metrics are present in the scikit-learn python package [Pedregosa et al., 2011b] that we employ in this thesis.

Given the plethora of methods, it is not obvious how to compare one feature attribution method to another, Lundberg et al. [2020] proposes two proprieties that should be desirable in any feature or variable importance method: consistency and accuracy. Whenever we change a model such that it relies more on a feature, then the attributed importance for that feature should not decrease. If consistency fails to hold, then importance does not translate to model reliance on a given feature. Further, the sum of all the feature importances should sum up to the total importance of the model. If accuracy fails to hold, then it is uncertain how attributions of each feature combine to represent the output of the whole model. Any attempt at normalization might jeopardize the consistency of results. Thus, if any of these characteristics cannot be guaranteed, then we also cannot properly compare different approaches.

One such method that attempts to address these issues is the work of Lundberg and Lee [2017]. Shapley Additive Explanations (or simply SHAP) is the usage of Shapley values to interpret a target model. We represent how model x' explains the data as a $d-$dimensional vector $E(x') = e_1, e_2, \ldots, e_d$ showing which features are contributing most to the model's prediction. The Shapley value is a concept in cooperative game theory [Shapley, 1953]. In each game, a unique distribution of the rewards generated by the cooperation of all players given is provided.

Let N be a set of n players in a cooperative game, S denote a coalition of players, and ν be a characteristic function over S. That is, $\nu(S)$ denotes the worth of a coalition S and describes the total expected sum of payoffs that the members of S obtain by cooperation. Adding player n_i to an existing coalition S increases the expected payoff by $\nu(S \cup \{n_i\}) - \nu(S)$. Since there are $n!$ possible ways to line up the n players and the player n_i must be preceded by all the members of S and followed by remaining players in

N, there are $|S|!(n-1-|S|)!$ lineups in which player n_i joins the existing coalition S. If we sum its contribution over all lineups in which n_i joins S and over all possible existing coalitions S that it might join, we get its total contribution over all possible lineups of N. The Shapley value $\varphi_{n_i}(\nu)$ of player n_i is the average of its total contribution in the cooperative game (ν, N). The Shapley value $\varphi_{n_i}(\nu)$ in the cooperative instance (ν, N) is the average of its total contribution over all possible scenarios:

$$\varphi_{n_i}(\nu) = \sum_{S \subseteq N \setminus \{n_i\}} = \frac{1}{n!} \; |S|!(n-1-|S|)! \; (\nu(S \cup \{n_i\}) - \nu(S))$$

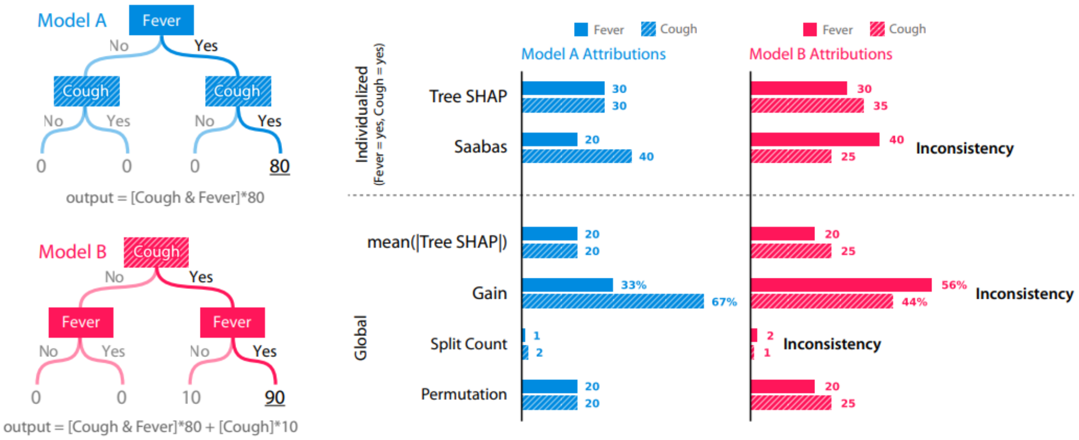
in which $v(S \cup \{n_i\}) - v(S)$ is called the marginal contribution.

Thus, in each iteration, a unique distribution of the rewards generated by the cooperation of all members given all possible coalitions is provided, giving each feature's contribution to the explanation. To interpret the target model, all features are players in a cooperative game represented by the trained predictive model cooperating to predict a given task. Each feature's aggregate payoff, its reward, is its actual prediction minus the result from the explanation model. The impact of each feature can, therefore, be found by calculating its Shapley value. A key important aspect is that this approach is model agnostic. Any learning algorithm, from simple linear regression models to complex deep networks, can be explained through Shapley values however, their exact computation is an NP-hard problem as it involves averaging the results of all possible $N!$ permutations of coalitions. SHAP proposes several approximation methods. Feature independence and model linearity are two optional assumptions that simplify the computation of the expected values. Aside from GINI and Shap, there are many other feature attribution methods [Breiman et al., 1984, Chen and Guestrin, 2016, Ribeiro et al., 2016, 2018, Saabas, 2014], but SHAP is the only method with the three desirable properties as pictured in Figure 2.1:

- Local accuracy: the explanations are truthfully explaining the model.
- Missingness: missing features have no attributed impact on the model 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.

The local accuracy axiom states that the value assigned to a player should equal the player's marginal contribution to the coalition if the player were to join or leave the coalition. The missingness axiom ensures that the value assigned to a player should not depend on the presence or absence of other players in the coalition. The consistency axiom requires that the value assigned to a player should remain unchanged if the player joins or leaves different coalitions. These axioms work together to ensure that the Shapley value

Figure 2.1: Two examples of decision trees that demonstrate inconsistencies in the Saabas, gain, and split count attribution methods.



Source: Lundberg and Lee [2017]

accurately reflects the contribution of each player to the coalition and is not affected by external factors or the presence of other players.Shapley [1953] proved that if these three axioms are to hold, then the only solution to the proposed problem of fairly distributing credit importance is through the Shapley values.

As stated by Lundberg and Lee [2017], the feature importance values from the gain, split count, and Saabas methods are all inconsistent. Applied to the field of Machine Learning, this means that under these alternative feature importance methods, a model can change such that it relies more on a given feature, yet the importance estimate assigned to that feature decreases. In the example from Figure 2.1, the Cough feature has a larger impact on Model B than Model A but is attributed less importance in Model B. The global attributions represent the overall importance of a feature in the model. It is important to highlight that the consistency guarantee of Shapley values is what allows us to use it as a direct measure of model reliance, which is one of the key concepts behind our Rashomon ensembles.

2.3 Ensemble Learning

Regarding ensemble learning and searching for partitions in data, Grosskreutz [2008] propose splitting dataset lines 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 a problem where there is no target variable. However, one can employ an equivalent technique regardless of this fact, similar to Malik and Kender [2008] and Knobbe and Valkonet [2009]. 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

Another possibility is the one presented by Dembczyński et al. [2008], focusing 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. [2020], it is possible that 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. However, boosting remains one of the state-of-the-art techniques concerning ensemble learning.

The main idea behind boosting is using an ensemble of weak learners that can be, somehow, combined to generate a stronger one. This idea was first proposed by Kearns [1988] as the *Hypothesis Boosting Problem*. He states that there might be an efficient algorithm that could convert poor hypotheses, like weak learners which are slightly better than a random guesser, into a single very good hypothesis. One approach is filtering the observations, thus modifying the distribution of examples in such a way as to force the weak learning algorithm to focus on the harder-to-learn parts of the distribution [Schapire, 1990].

Therefore, boosting consists of the usage of the weak learning method several times to get a succession of hypotheses. Each one is focused on learning to handle the remaining difficult observations in which the previous learner struggled. Predictors are not made independently, but sequentially and each learns from the mistakes of the previous predictors. This in turn leads to observations having an unequal probability of appearing in subsequent models and the ones with the highest error appear the most. Gradient boosting machines (GBM) is an example of a boosting algorithm that originated from the observation of Breiman [1997]: boosting can be interpreted as an optimization algorithm over a suitable loss function.

Let y be the actual values of the output variable, i be an iteration of the gradient boosting algorithm, and $F_i(x)$ be the output of the proposed model at time i. The gradient boosting algorithm improves $F_i(x)$ by constructing a new model that adds an estimator h to provide a better model, which leads to $F_{i+1}(x) = F_i(x) + h(x)$. A perfect h would imply in $h(x) = y - F_i(x)$. Therefore, the gradient boosting approach will attempt to fit h to the residual loss. However, in classification and ranking problems, residuals $y - F(x)$ for a given model are the negative gradients concerning $F(x)$. Therefore, gradient boosting is a gradient descent algorithm for combining and training weak learners. A common learner used is random forests.

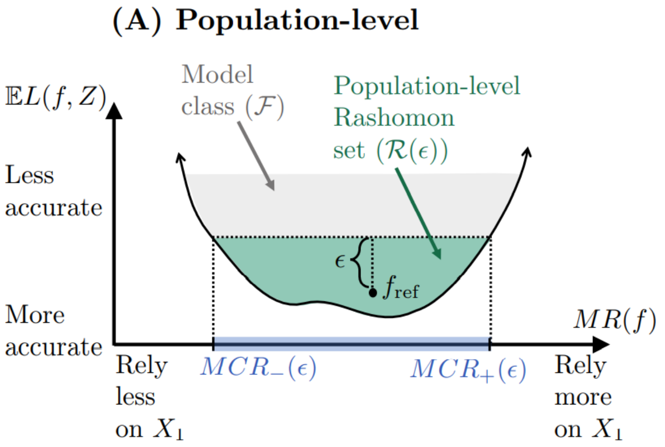
In this work we explore using XGBoost, which improves upon the original GBM [Chen and Guestrin, 2016]; LightGBM, a both faster and more efficient implementation of GBM by Ke et al. [2017b]; and Catboost, the state-of-the-art in decision tree-based boosting [Prokhorenkova et al., 2018]. For instance, it allows trees to be greedily created from sub-samples of the training dataset. This leads to a reduction in the correlation between the trees and prevents over-fitness. This variation of boosting is called stochastic gradient boosting. The main drawback lies in its sensitivity to outliers since every constituent is dependent on the errors of the predecessors in the ensembling pipeline. Another disadvantage is its scalability and parallelization due to this inner dependence, as as such we diverge from boosting for our Rashomon learning approach.

2.4 Rashomon Effect

According to D'Amour et al. [2020], a learning pipeline selects a prediction $f(X)$ from a model space $\mathcal{F}$ by minimizing the predictive risk $R_Z(f) := E_{(X,Y)\sim Z}[L(f(X), Y)]$ validating that f achieves low expected risk on a second identically distributed $Z = [Y, X]$. This validation provides a statistical guarantee of model performance on unseen data and, as such, we say that in this scenario the model is *specified*. A pipeline is *underspecified* if there are many predictors f that achieve a similar predictive risk, encompassing a set of equivalent near-optimal predictors. When these encode different biases, we can expect different generalization behavior on distributions that differ from Z. This notion is closely related to the *Rashomon effect of statistics* of Breiman [2001a], also known as the multiplicity of performant models, and which can be exploited to obtain insights from the explored problem.

The Rashomon set represents the study of a set of close-to-optimal models that share similar performance due to the Rashomon Effect. In Fisher et al. [2019] definition, we need a comparison to some key reference model, which will be denoted as f_{ref}. This

Figure 2.2: Hypothetical ϵ-Rashomon set within a model class $\mathcal{F}$. The y-axis represents the loss of each model and the x-axis the model's reliance on X_1.



Source: Fisher et al. [2019]

f_{ref} can be derived from expert knowledge such as, for example, a flowchart used to predict injury severity in a hospital's emergency room, or from another quantitative decision rule that is currently implemented in practice. This prespecified reference model will serve as a baseline performance. Thus, if we establish ϵ as the maximum accepted error about f_{ref} to consider a model as part of a subset compromised by f_{ref}, we can denote the ϵ-Rashomon set as:

$$R(\mathcal{F}, \epsilon) := \{f \in \mathcal{F} : E[L(f, Z)] \leq E[L(f_{ref}, Z)] + \epsilon\} \tag{2.2}$$

where E denotes expectations with respect to the population distribution, L is some nonnegative loss function. 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 a feature that model f_{ref} rely upon to reach a prediction. This reliance metric has a direct relationship with the explanation of the model. We can expect models that rely too heavily on x_1 to be prone to high variance, leading to low performance. Likewise, models that rely too laxly on x_1 are prone to high bias, also leading to low performance. The 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. Figure 2.2 illustrates a hypothetical Rashomon set R(ϵ), within a model class $\mathcal{F}$. The range of all possible MR values inside this class gives rise to the notion of Model Class Reliance (MCR), shown in blue, and helps us define a minimum ($MCR_-(\epsilon)$) and maximum ($MCR_+(\epsilon)$) value of MR to render a model f to be within the class of models defined by f_{ref}, ϵ and X. These models compromise the set of performant models that also share a similar reliance on the predictor variables as the reference model f_{ref}.

Semenova and Rudin [2019] extend the definition of a Rashomon set by defining the anchored Rashomon set. Given a threshold γ that restricts the empirical risk of a model concerning a loss function σ, we denote the anchored-Rashomon set $\hat{R}(\mathcal{F}', \gamma)$ as the subset of models with expected loss no more than γ, defined by $\hat{R}(\mathcal{F}', \gamma) := \{f \in \mathcal{F}' : \hat{L}(f) \leq \gamma\}$. There are two key distinctions between Semenova's and Fisher's definitions. The first is that a threshold is employed to define the Rashomon set instead of a prespecified reference model, which in turn makes it independent of the choice of an anchor model. The second is that it measures directly the empirical Rashomon set instead of making assumptions regarding the true Rashomon set concerning empirical observations.

As a theorem, let $\mathcal{F}_1$ and $\mathcal{F}_2$ be hypothesis spaces such that $\mathcal{F}_1 \subset \mathcal{F}_2$. For instance, $\mathcal{F}_1$ could be the less complex empyrical hypothesis space, and $\mathcal{F}_2$ be the true complete hypothesis space. Further, let the expected loss σ be bounded by b such that $\sigma(f_2, z) \in [0, b] \forall f_2 \in \mathcal{F}_2, \forall z \in Z$ with, again, $Z = [YX]$. We can define an optimal function $f_2^* \in argmin_{f_2 \in \mathcal{F}_2} L(f_2)$. Thus, f_2^* is the model with the smallest loss among all models contained within $\mathcal{F}_2$. If we assume that the true Rashomon set is large enough to include a function $f_1' \in \mathcal{F}_1$ such that $f_1' \in R(\mathcal{F}_2, \gamma)$ is also true then, for any $\epsilon > 0$ with probability at least $1 - \epsilon$ concerning the random draw of data:

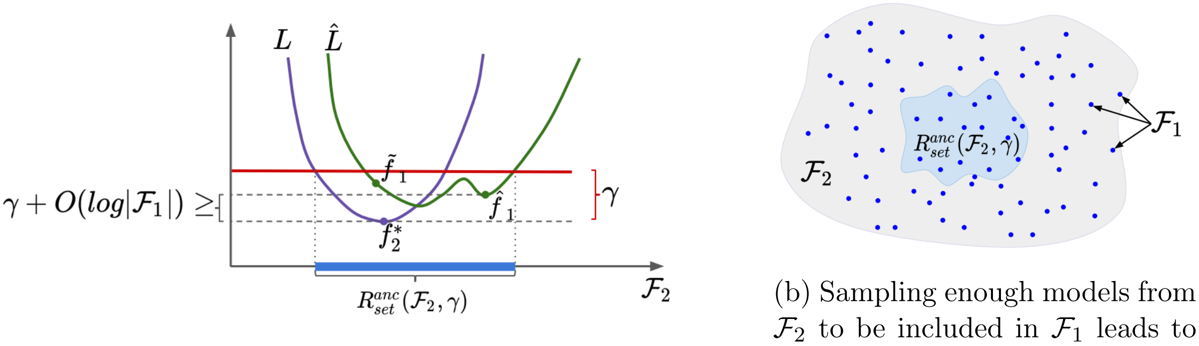
$$|L(f_2^*) - \hat{L}(\hat{f}_1)| \leq \gamma + 2b\sqrt{\frac{log|\mathcal{F}_1| + log2 - log\epsilon}{2n}} \tag{2.3}$$

where $\hat{f}_1$ is the model with the lowest loss among all models within $\mathcal{F}_1$.

This implies that the difference between the smallest empyrical error found in $\mathcal{F}_1$ is close to the smallest true error found in $\mathcal{F}_2$ and we can approximate an optimal model for $\mathcal{F}_2$ with the best empirical model within $\mathcal{F}_1$. Another finding of Semenova and Rudin [2019] is that we can create $\mathcal{F}_1$ by random sampling of $\mathcal{F}_2$. If we sample sufficiently many models from $\mathcal{F}_2$, with a high probability there will be a model from $\mathcal{F}_1$ that will be within the Rashomon set of $\mathcal{F}_2$. If $R(\mathcal{F}_1, \gamma)$ is large enough to include a function $f_1' \in \mathcal{F}_1$ such that $f_1' \in R(\mathcal{F}_2, \gamma)$, then the difference between the losses of the best models found within $R(\mathcal{F}_1, \gamma)$ and $R(\mathcal{F}_2, \gamma)$ is small. Figure 2.3 illustrates both of these ideas. Semenova and Rudin [2019] provides a Theorem with these previous statements, reproduced as follows:

> ***Lemma**: For any models $f, f' \in \mathcal{F}$ that are in the true anchored Rashomon set, we have $|L(f) - L(f')| \leq \gamma$.*
>
> ***Proof**: Consider two models f and f' from the true anchored Rashomon set. Let $L(f) = \gamma'$ and $L(f') = \gamma''$. Then if $\gamma' > \gamma'' : L(f) - L(f') = \gamma' - \gamma'' \leq \gamma$, otherwise $L(f') - L(f) = \gamma'' - \gamma' \leq \gamma'' \leq \gamma$. Combining these inequalities, we get the statement of the lemma.*
>
> model $\tilde{f}_1 \in \mathcal{F}_1$ such that $\tilde{f}_1 \in R^{anc}_{set}(\mathcal{F}_2, \gamma)$. In that case, for any $\epsilon > 0$ with probability at least $1 - \epsilon$ with respect to the random draw of data:

Figure 2.3: Relationship between different hypotheses spaces Rashomon sets.


L
$\hat{L}$
$\tilde{f}_1$
$\hat{f}_1$
f_2^*
$\gamma + O(log|\mathcal{F}_1|) \geq$
γ
$R_{set}^{anc}(\mathcal{F}_2, \gamma)$
$\mathcal{F}_2$
$\mathcal{F}_2$
$R_{set}^{anc}(\mathcal{F}_2, \gamma)$
$\mathcal{F}_1$

(a) Risks of $\mathcal{F}_1$ and $\mathcal{F}_2$ are close if there exists a model $\tilde{f}_1$ in the intersection of $\mathcal{F}_1$ and $R(\mathcal{F}_2, \gamma)$ if $\mathcal{F}_1 \subset \mathcal{F}_2$.

(b) Sampling enough models from $\mathcal{F}_2$ to be included in $\mathcal{F}_1$ leads to high probability of a model from $\mathcal{F}_1$ to be in $R(\mathcal{F}_2, \gamma)$.

Source: Semenova and Rudin [2019].

$$|L(f_2^*) - \hat{L}(\hat{f}_1)| \leq \gamma + 2b\sqrt{\frac{log|\mathcal{F}_1| + log2 - log\epsilon}{2n}}$$

***Proof of Theorem**: We apply the union bound and Hoeffding's inequality to the Lemma. The result is that with probability at least $1-\epsilon$ for every $f_1 \in \mathcal{F}_1$ we have, for a finite hypothesis space $\mathcal{F}_1$:*

$$|L(f_1) - \hat{L}(f_1)| \leq 2b\sqrt{\frac{log|\mathcal{F}_1| + log2/\epsilon}{2n}}.$$

Combining this Occam's razor bound with the definition of $f_2^* \in argmin_{f\in\mathcal{F}_2} L(f)$ we get that, under the same conditions:

$$L(f_2^*) \leq L(\hat{f}_1) \leq \hat{L}(\hat{f}_1) + 2b\sqrt{\frac{log|\mathcal{F}_1| + log2/\epsilon}{2n}}.$$

By assumption of the theorem, there exists a function $\tilde{f}_1 \in \mathcal{F}_1$ such that $\tilde{f}_1 \in R_{set}^{anc}(\mathcal{F}_2, \gamma)$. Since f_2^* is an optimal model, then $f_2^* \in R_{set}^{anc}(\mathcal{F}_2, \gamma)$ is one as well. From the lemma $|L(f_2^*) - L(\tilde{f}_1)| \leq \gamma$, which implies $L(\tilde{f}_1) \leq L(f_2^*) + \gamma$. Given that $\hat{f}_1 \in argmin_{f\in\mathcal{F}_1} \hat{L}(f)$, and using the equation, we get that with probability at least $1-\epsilon$, we have:

$$\hat{L}(\hat{f}_1) \leq \hat{L}(\tilde{f}_1) \leq L(\tilde{f}_1) + 2b\sqrt{\frac{log|\mathcal{F}_1| + log2/\epsilon}{2n}} \leq L(f_2^*) + \gamma + 2b\sqrt{\frac{log|\mathcal{F}_1| + log2/\epsilon}{2n}}$$

Combining the previous two equations we obtain:

$$|L(f_2^*) - \hat{L}(\hat{f}_1)| \leq \gamma + 2b\sqrt{\frac{log|\mathcal{F}_1| + log2/\epsilon}{2n}}$$

[Semenova and Rudin, 2019]

The theorem provides a bound on the difference between the expected value of the loss function of two models: $\hat{f}_1$ in the hypothesis space $\mathcal{F}_1$ and f_2^* in the hypothesis space $\mathcal{F}_2$. The loss function L is evaluated on the Rashomon set. The bound is dependent on the presented lemma pertaining to the difference between the losses of the two models $L(f_1)$ and $L(f')$, which in the theorem represent $L(f_2^*)$ and $\hat{L}(\hat{f}_1)$, and also on the size of the hypothesis space $log|F_1|$ and the tolerance ϵ. The theorem states that with a high probability, the expected loss of the model $\hat{f}_1 \in \mathcal{F}_1$ will be close to the expected loss of the optimal model $f_2^* \in \mathcal{F}_2$.

Occam's Razor bounds and Hoeffding's inequality are used in the theorem to provide a probabilistic bound on the difference between the expected value of the loss functions. Occam's Razor bounds help ensure that the model is not over-complicated, by penalizing complex models. In this theorem, the Occam's Razor bound provides a term proportional to the square root of $log|\mathcal{F}_1|$, which becomes smaller as the size of the hypothesis space $\mathcal{F}_1$ becomes smaller, thus becoming more restrictive. Hoeffding's inequality is a concentration inequality that provides a bound on the deviation of the sum of independent random variables from its expected value. In this theorem, Hoeffding's inequality is used to provide a term that becomes smaller as the number of samples increases or as the error tolerance ϵ decreases. By combining these two bounds, the theorem provides a probabilistic bound on the difference between the expected value of the loss function of two models, penalizing both the complexity of the models and the deviation from the expected loss value. This in turn leads to the following proposition of Fisher et al. [2019]: for a loss σ bounded by b and for any $\epsilon > 0$, with probability at least $1 - e^{-2n(\epsilon/b)^2}$ for the random draw of data, if $f \in \hat{R}(\mathcal{F}, \gamma)$ then $f \in R(\mathcal{F}, \gamma + \epsilon)$.

From these definitions, many works have exploited the Rashomon effect to gain insights into the solution space. Marx et al. [2020] explore 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 [2021] search for an entire collection of plausible models via a forward selection approach and resampling the training dataset to account for uncertainty. Dong and Rudin [2020] introduces the notion of variable importance cloud mapping every variable to its importance for the Rashomon set, and experimenting on criminal justice, marketing data, and image classification tasks, while Ning et al. [2022] performs a similar approach but 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.

[2022] and Danna et al. [2007] both look for the set of near-optimal solutions for integer linear programs while Xin et al. [2022] restrict their analysis of the Rashomon set to Decision Trees. However, to the author's knowledge, building an ensemble from the Rashomon set is a novel idea.

Chapter 3

Method

We consider a supervised learning scenario and formulate a classification model as a function $f(X, Y; \theta)$ parameterized by θ that maps inputs $x_i \in X$ to labels $y_i \in Y$. During cross-validation, we train models on data D_{train} coming from a distribution T. We can estimate the predictive risk of each function by employing additional data D_{test} correspondingly coming from T and evaluating $f_n \in F$ on this independent and identically distributed data. Usually, the most straightforward way to achieve this is to draw data D and hold out a set of instances selected completely at random, guaranteeing that $[D, D_{train}, D_{test}] \sim T$, $D_{train} \cup D_{test} = D$ and $D_{train} \cap D_{test} = \emptyset$. The standard model selection step consists in selecting the function that minimizes the empirical predictive risk. If future data follow the same distribution T, we obtain a guarantee of performance. These guarantees do not hold for other distributions such as when drift happens.

Let T be the train distribution of the source data, and U be some unknown distribution from another dataset. Candela et al. [2009] defines data drift as a change in the joint distribution of features. That is:

$$P(x_t, y_t) \neq P(x_u, y_u) \tag{3.1}$$

Probably approximately correct learning relies on the independent and identically distributed assumption between data distributions to estimate the empirical risk of a learning function. If we verify data drift, we cannot guarantee that the empirical risk is close to the real risk. That is if $U \neq T$, then the risk found during cross-validation may be inaccurate and model behavior becomes unpredictable.

We can decompose $P(x, y) = P(x) \times P(y|x)$. Thus, if we verify data drift, we could assume that it might come from two sources. We can observe a change in $P(x)$ (covariate drift), or a change in $P(y|x)$ (concept drift). As stated by Moreno-Torres et al. [2012], covariate drift is tied to the distribution of a variable, while concept drift implies that the relationship between the target and predictor changes between datasets. Finally, it's possible that both $P(x)$ and $P(y|x)$ present significant differences from the original distributions, which we define 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 [Lu et al., 2014].

Our main objective is building a diverse ensemble compromised of different and contrasting explanations for the same problem. As a secondary objective, we would like to estimate the reliability of our predictions under uncertainty due to the presence of an unknown data distribution U, which may contain drift when compared to the training data distribution T. We start our investigation by understanding how a model can behave when the differences between one execution and another may be only minor. Under this framework, θ encompasses any choices made during the training procedure that lead to virtually similar models possessing contrasting performances. We then introduce drift to the test data and once again evaluate its effects on each model.

Our hypothesis is three-fold and centered around the Rashomon effect. First, that individual model behavior under this new data distribution correlates to behavior under unseen distributions. Second, that electing models from the Rashomon set that disagree under this new test data enable the induction of an ensemble that tends to agree under distributions similar to training and disagree on other situations. And finally, that these learned insights are exploitable to real-world applications, addressed in the later Chapters.

Learning a model from the data space requires the minimization of an objective function $f(x)$. Instead of simply mixing multiple different structures into a single model x and minimizing $f(x)$, we can sample the model space by minimizing different functions $f(x')$, such that $x' \subseteq x$ and $|x'| < |x|$ [Zuin et al., 2020]. When considering Rashomon sets, optimizing inside this subspace is similar to optimizing in the complete model space if the subsample is large. In the absence of expert knowledge regarding the problem or a baseline, the loss obtained by a model trained on the whole set X should prove a suitable value for the stipulation of the Rashomon set. As the model space $\mathcal{F}'$ may contain models with competing explanations about their decisions, we want to build an ensemble exploiting two concepts:

- The concept of diversity between individual models. We recognize diversity as a central element to getting a more general understanding of any phenomenon. We assume that problems are not tied to a single causative factor, and that causative factors may vary depending on factors that might not be directly intuitive. To promote diversity while finding patterns, we cluster models in $\mathcal{F}'$ based on the distance between their explanation vectors (i.e., SHAP values). Ideally, this creates numerous groups of models that are internally dense and also separated from the rest of the models in terms of their explanatory factors, that is, within each cluster, the explanatory factors are similar, while factors within disjoint clusters are dissimilar.

- The concept of stability between model explanation and empirical predictions [Shmueli, 2010]. This is also tied to Occam Learning [Blumer et al., 1987]. We define a configuration of clusters as stable if models within the same cluster are associated with the same explanatory factors and perform similar predictions. Achieving cluster stability is challenging, as models that perform similar predictions can be associated with very different explanatory factors.

To assess prediction-explanation stability, we cluster the model space based on the distance between the explanation vector associated with each model, and then we project the clusters into the prediction space. This enables us to locate different Rashomon subgroups inside the Rashomon set and select models from each subspace. If we evaluate one constituent model at a time, the remaining constituents of the ensemble serve as hint models to address new data distribution. If it agrees with the remainder of the ensemble, this is indicative of prediction stability. But to study the Rashomon set for a given problem, we need to sample models from the complete model space. See Algorithm 1 for the pseudo-code regarding our ensemble learning approach, in which each step is further described in this chapter.

3.1 Deriving an Ensemble of Models from the Rashomon Set

We assume a factorial combinatorial space encompassed by all feature combinations constrained to a single learning algorithm. A key question is defining how many models to sample to guarantee the diversity of the Rashomon set. We assume that for a model to be considered within a specific Rashomon subset, it needs to contain a nonempty set of key features K that characterize the evaluated subspace. That is, there exists a region in the complete model space characterized by the nonempty F feature set that show complex correlations among a specific set of K features and the target label and the same correlations are not necessarily so strongly observed in other regions of the data space thus inducing a Rashomon subspace. This corroborates with the notion of Model Class Reliance and its relationship with the Rashomon set proposed by Fisher et al. [2019]. The complete model space is characterized by models from size $s = 1$ to F. If we also consider the $\emptyset$ model to be a part of the complete model space, then there are ${}_FC_0 + {}_FC_1 + ... + {}_FC_F$ models. From the binomial theorem:

$${}_FC_0 + {}_FC_1 + ... + {}_FC_F = \sum_{s=0}^{F} {}_FC_s = 2^F \tag{3.2}$$

For any model with less than $|K|$ features, intuitively it cannot contain all features from K. Furthermore, $|K| \approx |F|$ implies that nearly all features must be present for a model to be part of the Rashomon subset. In this scenario, it is unlikely that there exist multiple Rashomon subsets as the Rashomon ratio will be small. We limit our scope to problems in which $|K| << |F|$. For simplicity, we shall disregard our equations models with less than K features. The number of models containing K is:

$$\sum_{s=0}^{F} \frac{K!(F-K)!}{s!((F-K)-(s-|K|))!} = \frac{2^F}{{}_FC_K} \tag{3.3}$$

Combining the results of Equations 3.2 and 3.3 we obtain:

$$\frac{\text{models containing } K}{\text{total models}} = \frac{\frac{2^F}{{}_FC_K}}{2^F} = \frac{1}{{}_FC_K} \tag{3.4}$$

If we sample an arbitrary model from the complete model space, the probability of this model not containing K is $({}_FC_K - 1)/{}_FC_K$. From Equation 3.4, if we wish to guarantee that a given pattern is present with α probability we need to sample at least:

$$\eta = \frac{ln(1-\alpha)}{ln(1-\frac{1}{{}_FC_K)}} \tag{3.5}$$

For $F = 75$, $K = 3$ and $\alpha = 0.90$ one needs to sample at least 155 481 models. It is important to highlight that since a subset containing K span up to $|F| - |K|$ smaller subsets with $|K| + 1$ key features, this equation addresses the limit regarding the most specific patterns able to be found by sampling.

3.2 Time Complexity

Due to the large sample size, the time complexity of the algorithm needs to be addressed. In the experiments highlighted in our case studies we consider Decision Trees as base models for ensemble constituents, as they possess the advantage of being human-understandable, are not ensembles themselves, and allow for a fair comparison of the Rashomon ensemble to other tree-based state-of-the-art algorithms such as Gradient Boosted Machines or Random Forests, which are ensembles of Decision Trees themselves. As such, we will provide the complexity of running our proposed approach employing Decision Trees as base models.

Let I be the number of instances on a dataset, F be the number of features, and D be the maximum tree depth. The time complexity of training and explaining a tree is $O(log(F)ID + D^2)$ Lundberg and Lee [2017]. Since we sample T trees to evaluate the Rashomon space with both $log(F)$ and D being small constants, and since other steps present negligible complexity in comparison to the sampling stage, this results in $O(TI)$ complexity. That is, the time to train and explain T decision trees that employ $log(F)$

features on the I sized data. However, it is extremely unlikely that all sampled models will lie within the Rashomon set, thus not all sampled models will need to be explained.

The main advantage of employing a Rashomon set lies in severely reducing the number of models to be evaluated. For instance, any model with a loss close to random guessing is unlikely to present itself as a useful constituent. In real-world scenarios, only a small fraction of the sampled models will need to be explained. We can then compare the outputs of each base decision tree on a controlled test set and evaluate the agreement of constituents to establish the ensemble final prediction. The agreement of the ensemble is used as a proxy for empyrical uncertainty estimation. If the constituents, all of them independent and encompassing different biases, disagree even though during training they mostly agreed, then this should be an indicator of uncertainty. But to induce these independent constituents, we also need to adrees how to split the Rashomon set to uncover the different data biases regions. That is, the different regions of the Rashomon set that show complex correlations among specific sets of features and the target label and which are not necessarily so strongly observed in other regions.

3.3 Splitting the Rashomon Set

We represent how model f' explains a phenomenon as a d-dimensional vector $S(f') = [e_1; e_2; ...; e_d]$ showing which features $[x_1, x_2, ...x_d]$ are driving the model's prediction. If among a pair of models, the importance given to some feature varies wildly then this is an indicator that behavior in production may likewise vary. We wish to split the Rashomon set into cluster given the vector representation of each model within. We used K-Means to induce this division, finding a suitable number of clusters by maximizing the silhouette value. The silhouette is a measure of how similar a model is to its cluster (cohesion) compared to other clusters (separation) and ranges from -1 to $+1$, where a high value indicates that the model is well matched to its cluster and poorly matched to neighboring clusters. If the clustering presents a high average silhouette value, then its configuration is appropriate. In out experiments, we were able to determine that the groupings are divided primarily by which features compose their models. 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 leads to concise and well-divided clusters.

3.4 Kullback-Leibler Divergence and Unknown Data Distributions

Consider two decision tree models f_p and f_q training optimizing the cross-entropy function on the same data coming from the distribution T but that, due to some arbitrary parameterization of θ, differ from one another in their first split of features x_1 and x_2 while achieving the same empirical risk on a held-out test set. This leads to the importance of these features on f_a and f_b differing even though both equally fit the source data, implying the existence of multiple solutions for the same problem. Since $S(f_p) \neq S(f_q)$ due to the first split, we have no guarantee nor insight into their behavior on data drawn from an unknown distribution U. We could compute the empirical risk of models f_p and f_q on U to assert that the empirical risk remains small, but if data from U is unlabeled we cannot perform this estimation. But we can compare the probabilities returned from these two similar models to indirectly measure their risk.

Let P be the probability distributions returned from a model f_p, and we wish to compute a metric that estimates the risk of selecting it in production. Further, let Q be the probability distribution from a model f_q that ideally behaves similarly to f_p. As shown by MacKay and Mac Kay [2003], we can compute the error of P from Q by the cross-entropy between P and Q as:

$$H(P, Q) = H(P) + D_{KL}(P||Q) \tag{3.6}$$

Since the entropy $H(P)$ is inherent to f_p regardless of the choice of f_q, we can omit it from our calculation and instead focus on the Kullback-Leibler (KL) divergence D_{KL}, representing the expected excess surprise from using Q to approximate P. During training, we can restrain the choice of f_q to only models that achieve a statistically equal empyrical risk to that of f_p, that is, the Rashomon effect. We can also compute the KL divergence between both of these models under the test data distribution T'. In our experiments ee empirically verified small KL for the held-out test set for models within the Rashomon set.

We can then compute the KL divergence between these models under the unknown distribution U. If P and Q agree (i.e. low KL divergence) then we have a strong indicator that U should be similar to T. However, if P and Q are contrasting (i.e. high KL divergence), then it might be the case that U differs from T, and the returned predictions cannot be trusted. It can even be the case that the evaluated point is an outlier, whose feature domain lies outside the one seen during training. The main drawback of employing 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 opt to employ the Jensen-Shannon distance defined by Endres and Schindelin [2003] as our metric of choice for a measure of risk. Let $M = (P+Q)/2$, the Jensen-Shannon distance (JSD) is the total divergence to the average distribution:

$$JSD(P, Q) = \sqrt{\frac{1}{2}D_{KL}(P, M) + \frac{1}{2}D_{KL}(Q, M)} \tag{3.7}$$

3.5 Searching for Optimal Constituents

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 f_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 $y_i = f(x_i + \epsilon_i)$ and $\epsilon_i \sim N(0, \sigma^2)$. We can then select models that have contrasting explanations and predictions. There are many possible data transformations. We opt for the addition of gaussian noise due to its simplicity and ease of computing the 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 presents two desirable properties: the first being that any vertex is reachable from the $[\emptyset]$ model, the second being that there exists a topological ordering, an ordering of all vertices into a sequence such that for every edge, the start vertex occurs earlier in the sequence than the ending vertex of the edge for any feature set path. These properties imply a partial ordering of the graph starting from the root node, which allows us to search it in an orderly manner. It has been shown that this modeling approach is effective for the task at hand [Zuin et al., 2021, 2022a].

We can, for example, apply the A* algorithm [Hart et al., 1968] employing as 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 Rashamon subgroups clusteroids. We hypothesize that there exists a set of optimal feature expansions that lead to the most performant models for each specific base task. After sampling models and clustering them by their explanatory factors, we can build a single graph for each cluster and prune it. If after sampling a large number of models we do not verify the existence of a feature in any of the models about a specific grouping, we can prune all vertexes concerning this feature from the cluster's graph. This allows us to search the $F!$ combinatorial space of feature subsets to select the best-performing specialized models and build the Rashomon ensemble. Algorithm 1 presents the pseudo-code of our complete approach.

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

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

Algorithm 1: Rashomon ensemble algorithm.

3.6 Related Work

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 [Madras et al., 2020] or learning a scoring rule that captures ambiguity in targets [Lakkaraju et al., 2017, Lakshminarayanan et al., 2017]. However, most recent research on this topic has been mainly focused on Neural Networks and how they learn intermediary features. More specifically, the state-of-the-art approaches to Out-of-Distribution (OoD) detection enrich the intermediate feature space beyond what would ordinarily be learned via only supervised learning, such as encouraging a model to learn as many high-level task-agnostic semantic features as possible [Winkens et al., 2020] or employing an additionally labeled outlier dataset during training [Hendrycks et al., 2019]. When one cannot look at the intermediate feature space, most of the mentioned approaches fail. As mentioned by Chen et al. [2021], this sort of approach possesses two drawbacks: the first is that models trained to identify OoD may fail to cover the whole data distribution. And the second is that 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 their explanatory factors [Lundberg and Lee, 2017]. A second key idea is to exploit the Rashomon Effect to look for models with similar performance during training. Both of these propositions enable an explanation of the risk metric by assigning importance to the factors leading to each model decision and comparing both. Further, our approach is algorithm-agnostic, and reproducible in any model that handles tabular data. We can therefore summarize three pivotal points underlying our approach: understanding that production data may fall in out-of-distribution data; the multiplicity of performant models; and the explanatory factors behind a model decision. Finally, we present in this section some other relevant definitions to understand the context and project decisions behind our approach:

Underspecification: Underspecification in deep learning arises when models achieve similar in-sample performance but present divergent behaviors in out-of-sample data. This is a problem when some of these models perform significantly worse when employed in production and thus present a challenge for proper model selection [D'Amour et al., 2020]. Although underspecification literature has been focusing on the emergence of this phenomenon in deep neural networks, in which underspecification mainly arises from the elevated number of optimized parameters [Bui Thi Mai, 2021, Ortiz-Jiménez et al., 2021], Mei and Montanari [2019] state that this phenomenon is common to any machine learning pipeline. Damour et al. observed that repeating a training process can generate many models of identical test performance but significantly different behaviors, even when performing only minor perturbations such as selecting a different random seed. This in turn differentiates each created model into small arbitrary learning decisions and, even though these differences are usually considered minor, the consequence is varying degrees of performance seen in the real world. As such, underspecification is closely tied to the Rashomon effect.

Rashomon Effect: Fisher et al. [2019] analyzes the set of models with accuracy close to the optimal model. From this set, they formally defined the concept of *Rashomon Set*, this being the subspace of the universe of models that summarize the range of effective prediction

strategies that an optimal analyst might choose. Semenova and Rudin [2019] delve deeper into the theme of the Rashomon Effect in machine learning, giving pertinent definitions about the generalization of the Rashomon Set as well as its format and volume. In particular, it is explored in which situations it is possible to obtain a sample of the model space such that the properties related to Rashomon in this subspace are similar to those of the sample universe.

As discussed previously, the Rashomon set thus compromises a study of a set of close-to-optimal models that share similar explanations and performance due to the Rashomon Effect. For that, we need a comparison to some key reference model, which will be denoted as f_{ref}. Fisher suggests that f_{ref} can be derived from expert knowledge or from some quantitative decision rule that is implemented in practice. This prespecified reference model will serve as a baseline performance which, when coupled with a maximum error margin ϵ, estabilish the Rashomon set $R(\mathcal{F}, \epsilon)$ of Equation 2.2.

Rashomon ratio: Semenova and Rudin [2019] also explore the shape and volume of the Rashomon set. In their study, they define the Rashomon ratio, that is, the fraction of models inside the Rashomon set derived from the complete model space. This can be computed as

$$R_{ratio} = \frac{|R|}{|H|} \tag{3.8}$$

in which R represents the Rashomon space and H is the complete hypothesis space. The exact computation of the Rashomon ratio requires evaluating all possible $h \in H$, which is untractable. We can approximate the $\hat{R}_{ratio}$ by random sampling models from the complete model space and checking the number of sample models that lie within the Rashomon set. Most of our proposed approach relies on sampling the model space and evaluating the empirical Rashomon set which, if large enough, holds guarantees of similarities to the real Rashomon set.

Model Explanation: Instead of the model reliance metric proposed by Fisher et al., another possible approach is the one presented in Lundberg and Lee [2017]. Shapley

Additive Explanations (or simply SHAP), is the usage of Shapley values to interpret a prediction model. We represent how model f' explains the data as a $d-$dimensional vector $E(f') = e_1, e_2, \dots, e_d$ showing which features are contributing most to the model's prediction. The Shapley value is a concept in cooperative game theory by Shapley [1953]. In each game, a unique distribution of the rewards generated by the cooperation of all players given is provided. There are many other feature attribution methods [Breiman et al., 1984, Ribeiro et al., 2018, Saabas, 2014] but, as highlighted by Hinns et al. [2021], both sound mathematical foundation and ease of implementation make SHAP ideal for underspecification identification. Further, SHAP is the only method with the three desirable properties:

- Local accuracy: the explanations are truthfully explaining 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.

Chapter 4

Exploring the Rashomon Set

We assess the statistical significance of our measurements through a pairwise t-test with p-value ≤ 0.05 and 5-fold cross-validation. No hyperparameter tuning was performed in any of the algorithms employed, opting to keep their default values across all datasets. We evaluate the performance of both classical and state-of-the-art algorithms for tabular data, namely Random Forests [Breiman, 2001b], LightGBM [Ke et al., 2017a], XGBoost [Chen and Guestrin, 2016] and Multi-layer Perceptron Neural Networks [Murtagh, 1991] with a single hidden layer of 100 neurons. For the LightGBM and XGboost models, we set the subsample probability to 50% to both avoid overfitting given the size of the evaluated datasets and introduce some stochastic behavior. We hold out 20% data from all training datasets to perform our evaluations, leaving the remaining 80% for training employing the SMOTE [Chawla et al., 2002] algorithm to upsample the minority class. All models were trained to optimize the binary cross-entropy function.

Five datasets were employed in our exploratory stage, one consisting of a classical open-source dataset that has been employed in numerous studies, thus allowing for easy reproducibility of results. The remaining four consist of data from pairs of co-related tasks that were collected in different locations. Table 4.1 summarizes the mean AUROC of each algorithm on each dataset. As expected, the performance variance on data from the same distribution is smaller than those on co-related ones.

WDBC: The Wisconsin Breast cancer dataset is composed of 569 diagnosis of breast mass usually associated with breast cancer alongside 32 features from a digitalized image of a fine needle aspirate (FNA) [Street et al., 1993]. Each feature describes the characteristics of cell nuclei present in a breast image. The problem is formulated as a binary classification task, where the end goal is to predict the presence of malignant tumor cells.

COVID: The COVID-19 Data Sharing/BR is an initiative of the São Paulo Research Foundation (FAPESP) in collaboration with a variety of local hospitals to publish open COVID-19 data to contribute to and foster research [FAPESP, 2020]. Pseudonymized data regarding clinical and laboratory exams, as well as hospitalization information, is available. We collected data from two key institutions in Brazil, the *Beneficência-Portuguesa hospital* (HBP), consisting of 91 648 exams, and from the *Sírio-Libanês hospital* (HSL), consisting of 37 643 exams, both of which share a similar set of biomarkers that we employ as features. The problem is formulated as a binary classification task, where the end goal is to predict the death prognosis of a patient 20 days before the event. We propose training on HBP data and evaluating HSL as our production stage representing the unknown U distribution that might be subject to drift. After filtering non-hospitalized patients, the training dataset consists of exams from 453 individuals hospitalized on HBP, and the production dataset consists of exams from 4 018 individuals hospitalized on HSL.

Alzheimer Disease: These datasets comprise data from patients who may suffer from Alzheimer's Disease symptoms. Each patient is represented using features such as gender, age, education level, and laboratory test results. The outcome indicates whether a patient is diagnosed with Alzheimer's Disease or not. The data come from two different hospital departments, namely: Geriatrics and Neurology. It is important to highlight that each department receives patients with different socio-economic characteristics, and some groups might be either under or over-represented in one of the departments. We propose training on geriatric data and evaluation on Neurology as our production stage representing the unknown U distribution that might be subject to drift. The neurology department includes patients that do not classify as geriatric, and thus ensures some degree of divergence. The training dataset consists of exams from 154 individuals admitted to the geriatrics department and the production dataset consists of exams from 166 individuals admitted to the neurology department.

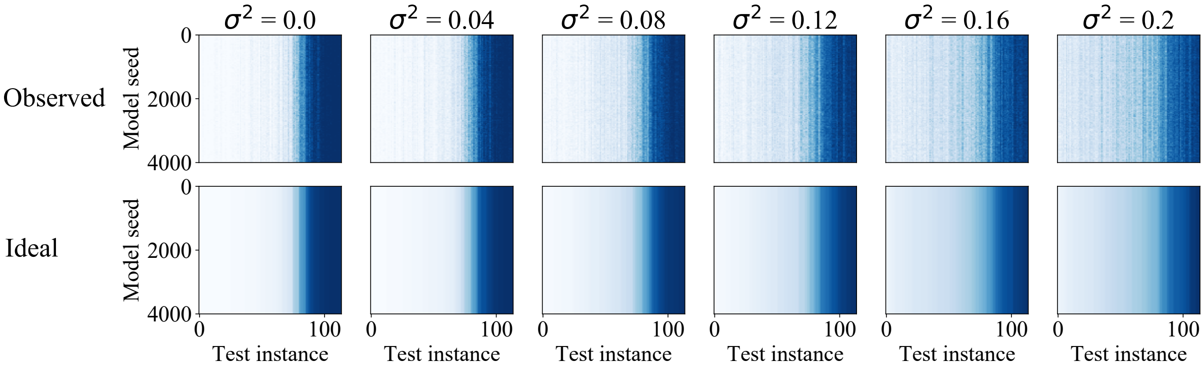
Table 4.1: AUROC performance of models on each dataset.

	WDBC	COVID-19		Alzheimer	
Algorithm		Beneficência Portuguesa	Sírio Libanês	Neurology	Geriatrics
Random Forest	.985 (±.007)	.929 (±.015)	.903 (±.030)	.914 (±.031)	.891 (±.040)
LightGBM	.982 (±.006)	.943 (±.004)	.901 (±.013)	.907 (±.023)	.812 (±.032)
XGBoost	.975 (±.009)	.935 (±.009)	.907 (±.020)	.878 (±.031)	.848 (±.039)
MLP	.996 (±.001)	.988 (±.002)	.962 (±.014)	.703 (±.041)	.677 (±.027)

4.1 Understanding Model Behavior

In our first set of experiments, we wish to understand how models differ from one another when performing only minor changes. The only hyperparameter changed from one model to another is its random seed and we trained all models on the same data. Thus, we study the models within this extremely concise Rashomon set. We elect the WDBC dataset for this set of experiments due to the vast literature explaining its intricacies. We progressively introduce Gaussian noise to the validation set, from $\sigma^2 = 0$ (no-noise) to $\sigma^2 = 0.2$ (high noise). Figure 4.1 illustrates each model's response to the increasing noise over its inputs. Ideally, we would expect the models' behavior to be indistinguishable from one another since they have seen the same training samples. In practice, each model responds differently to the modified input.

Figure 4.1: WDBC: effect of introducing increasing amounts of noise to the input features. For all subplots, each row represents a model and each column is a test instance while color illustrates the returned probability. Columns are ordered by the mean probability of all models.

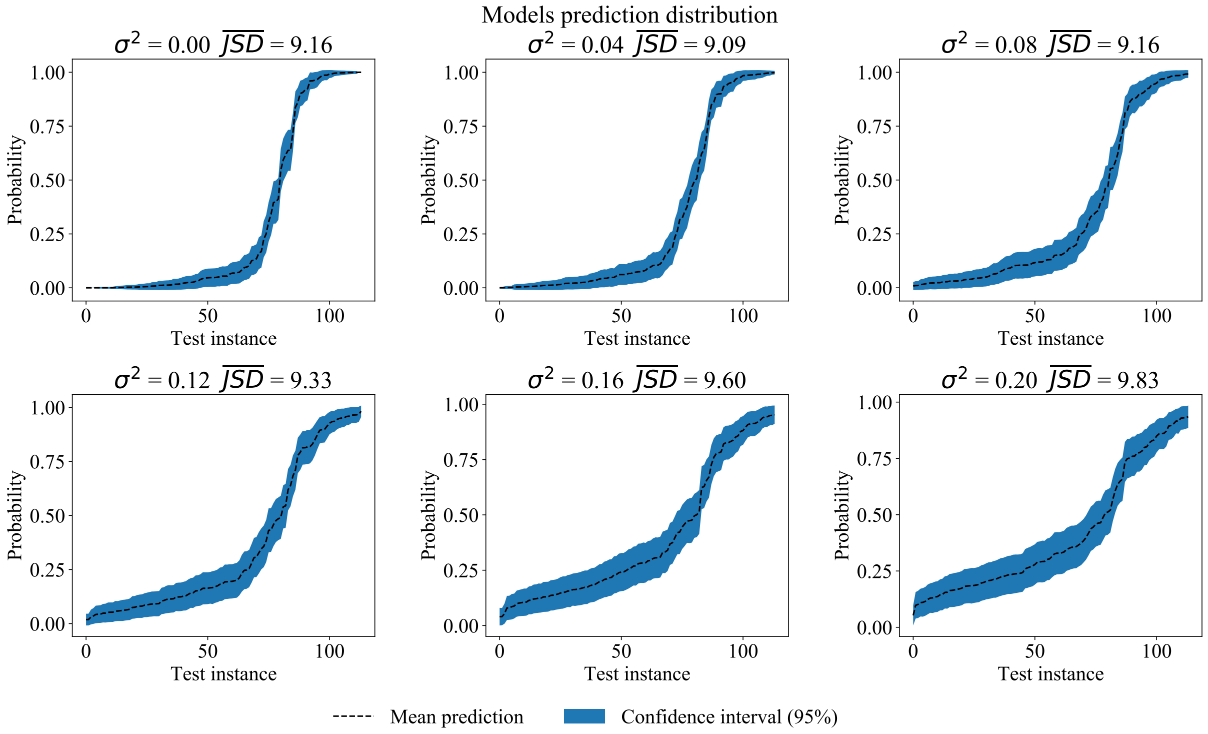


We verify that some models' predictions are more impaired by noise than others, implying that each model's reliance on the feature set differs. The more significant the amount of noise introduced, as shown in Figure 4.2, the higher the variance between models' predictions. This result corroborates with the underspecification observations made by D'amour et al. in which the differences in each model decision-making process, which at first are considered minor, can lead to varying effects. Under this controlled scenario, large amounts of noise introduced resulted in less than 10% in returned probabilities variation. This was not enough to change the predicted class for most instances, retaining comparable performance between models.

4.1.1 Divergence in Production

In our second set of experiments, we explore the behavior of models when trained on one dataset and evaluated on the second dataset of a co-related task. Namely, training on the Beneficência-Portuguesa COVID-19 dataset and evaluating on the Sírio-Libanês

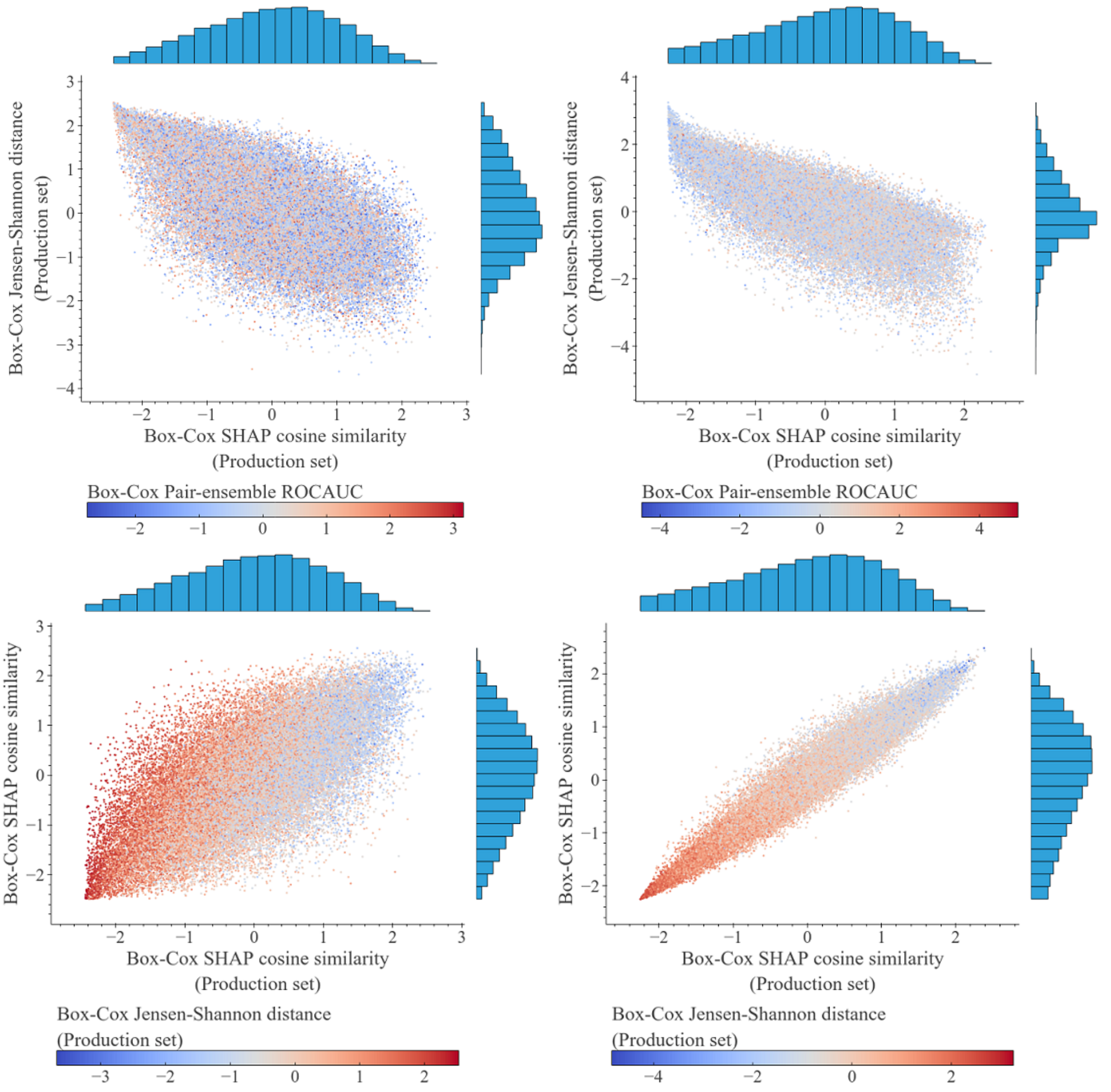
Figure 4.2: WDBC: effect of introducing noise to the model's input features. 5000 models are evaluated and the mean predicted probability for each test instance is computed. The greater the amount of gaussian noise, the greater the divergence around the mean average probability.



dataset, and training on the Neurology Alzheimer's dataset and evaluating on the Geriatrics Neurology dataset. This scenario represents what might happen when deploying a model. Like the previous experiments with the WDBC, we also separate 20% of each training dataset and introduce Gaussian ($\sigma^2 = 0.2$) noise to it, thus simulating a possible synthetic deployment dataset. The main hypothesis we wish to answer is whether we can employ some strategy from the data of this simulated dataset that could help us select a good deployment model. Further, since we already verified that each model can have a varying response to a new data generation function, we compute all pair-wise differences from 5 000 models with their only difference being the random seed, thus representing 12 497 500 point-wise comparisons. This enables us to understand the relationship between all models within this proposed Rashomon subset without being anchored to a single reference model. In practice, 100 models were enough to verify our results, but we sample 5 000 models for better visualization of the found patterns and behaviors. Box-Cox normalization was employed for clarity.

Figure 4.3 illustrates the comparison of models in the production and training stage in the COVID-19 pair datasets. We employ the cosine similarity between SHAP vectors as a means to compute the relatedness in the explanatory factors of each model while employing the Jensen-Shannon distance to measure the divergence in returned probability outputs. Figure 4.4 illustrates how the explanatory factors relate to themselves on the Alzheimer pair datasets. We also explore other state-of-the-art models on this same problem and observe a similar pattern across all of them, regardless of the problem or algorithm of choice.

Figure 4.3: COVID: Relationship between explanation, returned probabilities, and performance. We cannot observe a correlation between performance and either explanation or the returned probabilities themselves, but we can draw a relation from how similar the explanatory factors are in training and production and the probability vectors.



(a) Random Forest.

(b) LightGBM.

Figure 4.5, on the other hand, shows how the predictions themselves compare in test and production on the COVID datasets. Although there seems to be a small correlation in the Jensen-Shannon distance in the two scenarios, there is high variability across the whole prediction spectrum. We can however verify a relationship between the

Figure 4.4: Alzheimer: Relationship between SHAP and returned probabilities.

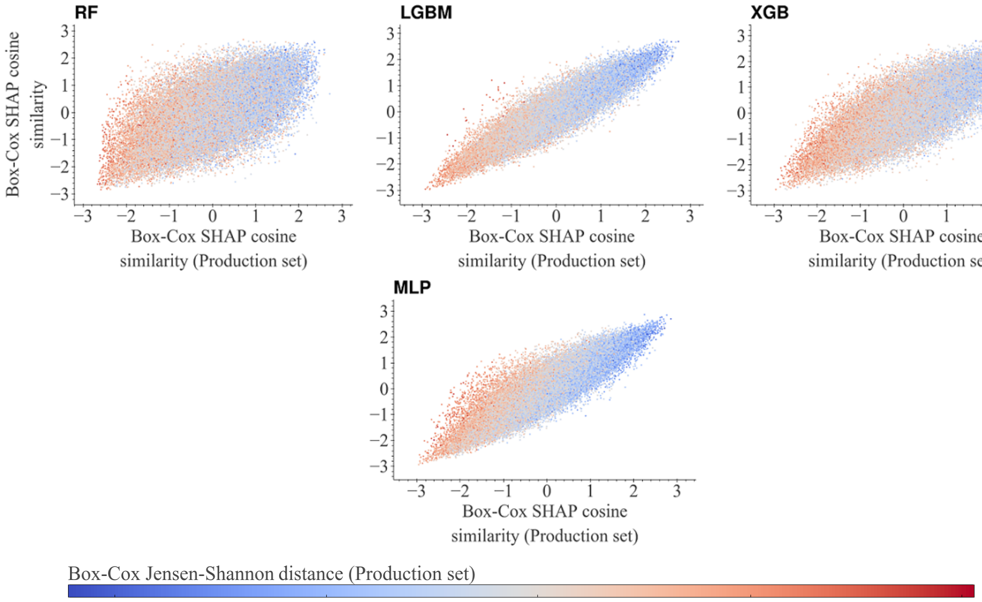
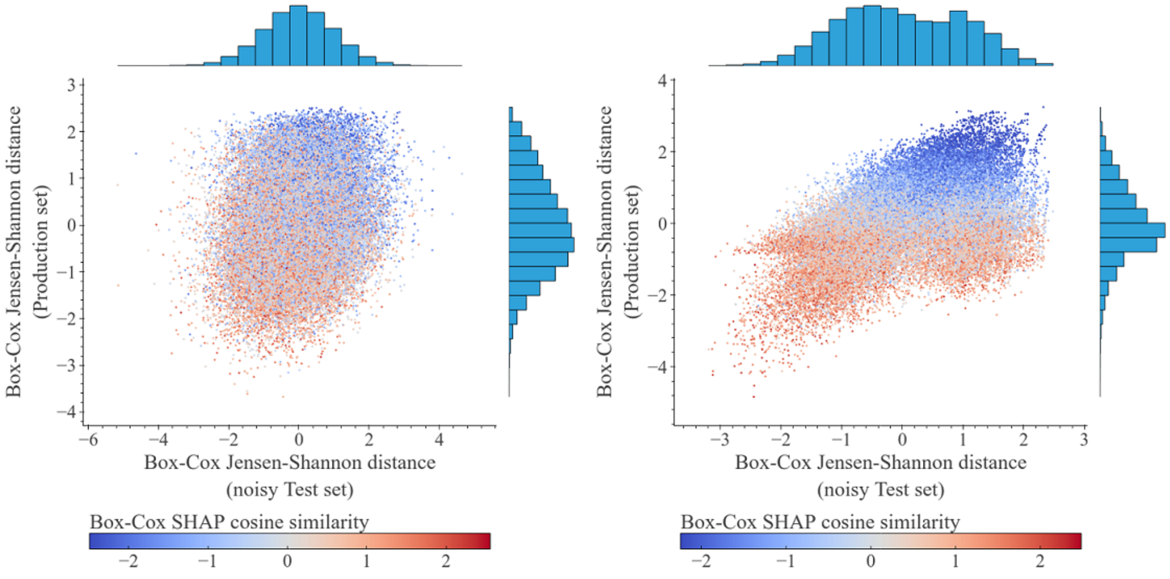


Figure 4.5: COVID - Relationship between the returned probabilities in production and the noisy test simulation.



(a) Random Forest.

(b) LightGBM.

probabilities vectors and the similarity of the explanatory factors, much like in Figures 4.3 and 4.4. We also reproduce this analysis on the Alzheimer's datasets, summarized in Figure 4.6.

Combining the results from our two previous experiments, we propose stratifying the models into two distinct sets and observing their behavior on production. The first one is comprised of pairs of models which have high similarity in both the predictions and the explanatory factors in the noisy simulated test set. We select only the pairs above the 95th percentile in the SHAP cosine similarity order and below

Figure 4.6: Alzheimer- Relationship between the returned probabilities in production and the noisy test simulation.

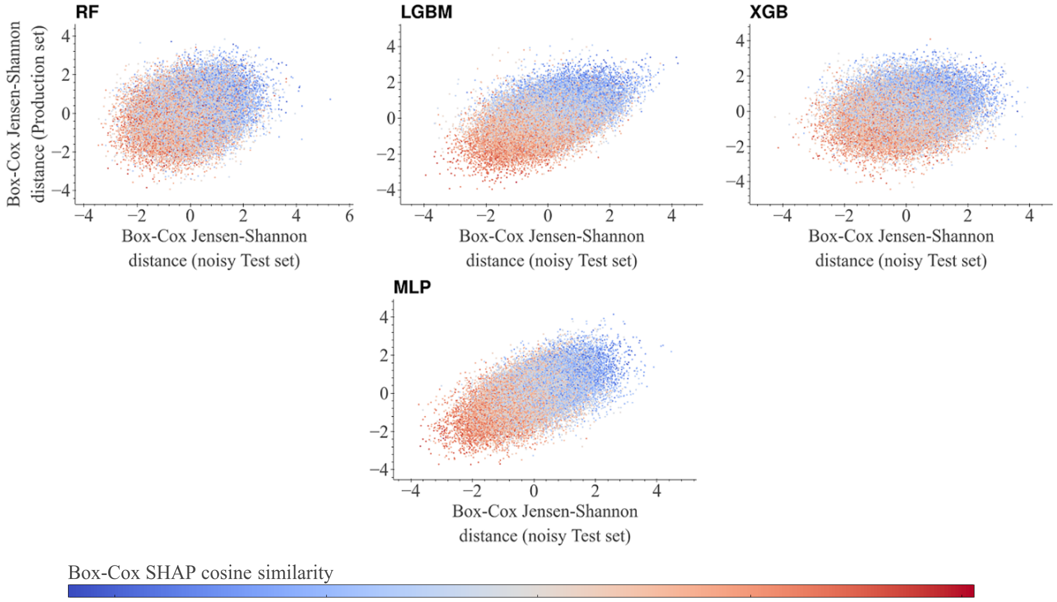
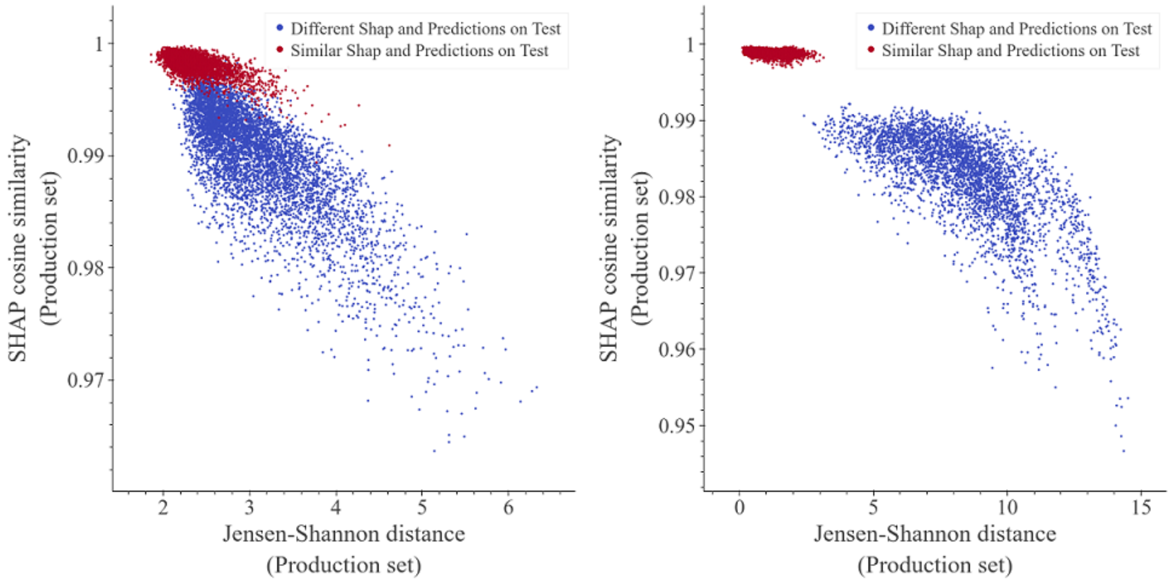


Figure 4.7: COVID: Rashomon subspaces characterized by high similarity in both explanation and probabilities (95th percentile and above) and low similarity in explanation and probabilities (5th percentile and below).



(a) Random Forest.

(b) LightGBM.

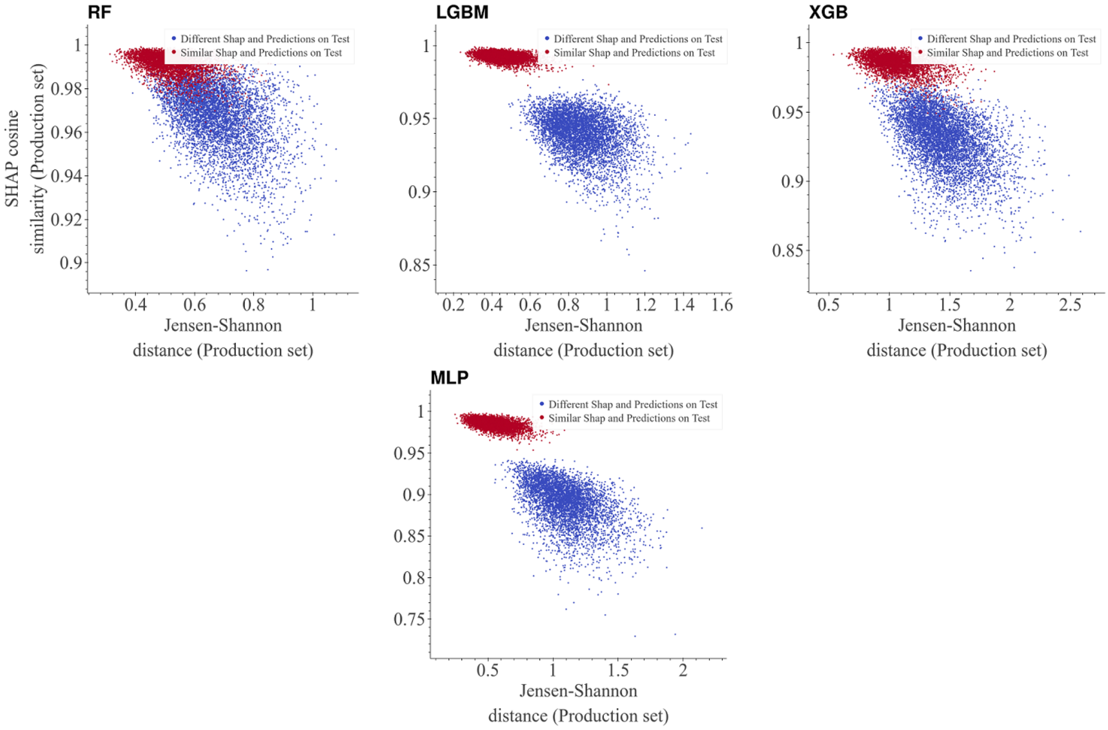
the 5th percentile in the Jensen-Shannon distance ordering. We hypothesize that these models should likewise behave similarly in production. The second group is composed of pair of models which have a SHAP cosine similarity below the 5th percentile, and a Jensen-Shannon distance above the 95th percentile. Analogously, we hypothesize that these models should diverge from one another in production. Figures 4.7 and 4.8 illustrate these two groupings in the COVID and Alzheimer's datasets respectively.

4.1.2 Choosing Models for Production

Looking at the standard cross-validation performance, all models reach statistically equal performance and empirical risk on the held-out test set. This can be seen in our first set of experiments where we do not introduce any Gaussian Noise to the input features. In this scenario, the variability in model predictions is small. Further, even under small perturbations, such as $\sigma^2 = 0.04$, this pattern can still be seen and the returned probabilities don't change much. We verify that the explanatory factors of each model are not similar, implying that the decision-making process differs from one model to another. We can expect outputs to diverge under other data distributions. For values of σ^2 above 0.1, we verify an increase in the confidence interval margin for the returned probabilities, thus signifying diverging outputs. We can also draw a correlation between the distance from the initial data distribution T and the inter-model prediction variability. This key result motivates all our remaining experiments.

In our production setting experiments, we saw no relationship between either probability or explanation and an increase in performance. As such, we cannot recommend any pair for production if this is the only parameter of interest. However, we can observe

Figure 4.8: Alzheimer: Rashomon subspaces characterized by high similarity in both explanation and probabilities (95th percentile and above) and low similarity in explanation and probabilities (5th percentile and below). Similar patterns to the ones observed on the COVID dataset can be seen.

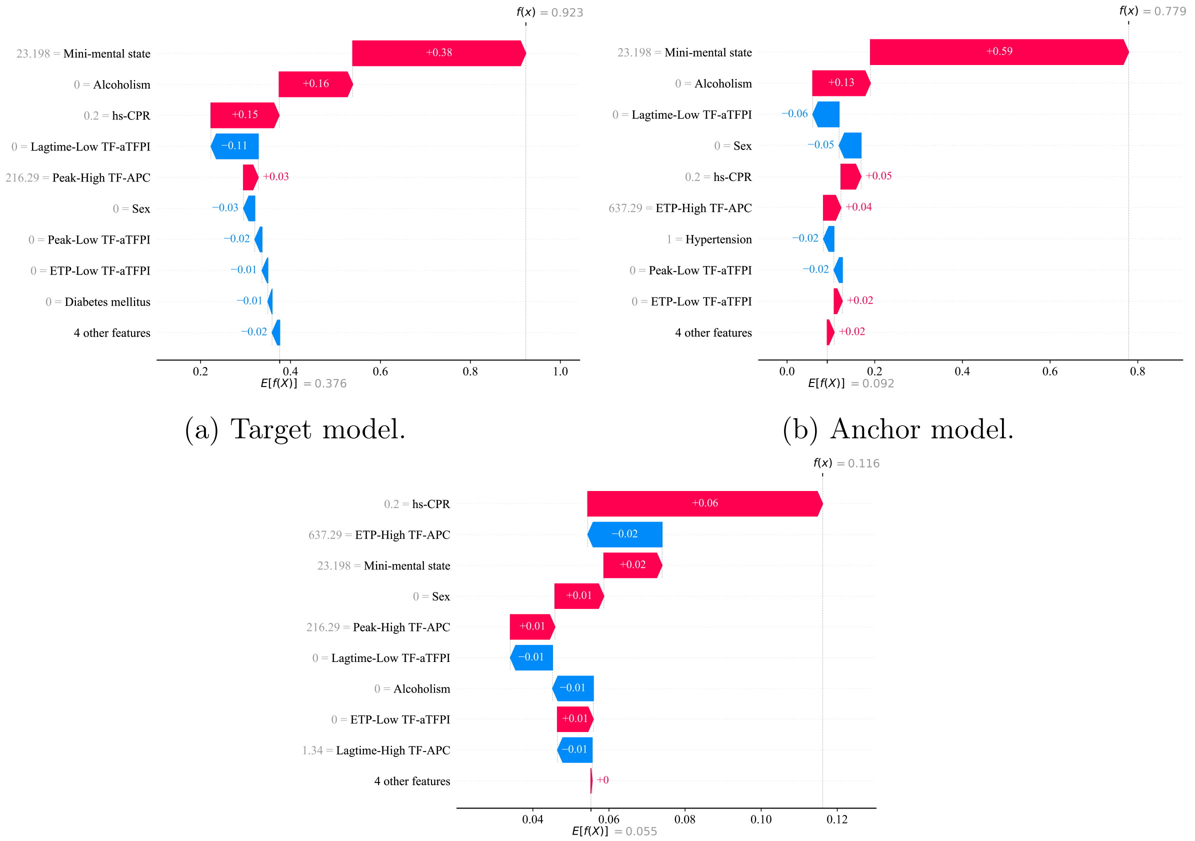


a direct relationship between the explanation and the probabilities. The more similar the explanatory factors of the two given models, the more similar the returned probabilities tend to be. There is also a direct relationship between the explanatory factors in the test and production. The similarity in returned probabilities allows us to further stratify models in those that behave similarly in the proposed setting from those that do not. The same cannot be said when comparing the similarity of probabilities in isolation. We do verify a weak relationship between outputs in production and under the simulated noise scenario, but not as strong as the explanation itself. Looking at the intercept between probabilities and explanations leads to some emerging patterns. This finding is found in both COVID-19 and Alzheimers experiments and across all algorithms.

We can split the model pairs into two relevant groups given their behavior. Those with contrasting explanations and predictions, and those with compounding ones. We can expect pairs of models that behave similarly to one another under the presence of noise to remain to do so under other distributions. Models that differ under the presence of noise, on the other hand, might prove more useful. We know that these models do not diverge in predictions when test and train distributions match. Thus, we can use this knowledge to derive a risk assessment. If we employ both of these models at production and they disagree, we have a strong indicator that the new instances seen lie outside the train distribution. Under this scenario, the cross-validation risk guarantees cease to hold and we cannot trust the prediction. Analogously, if the models agree then the distributions should be similar. The threshold of how close the predictions should be to validate that the distributions match is a work in progress. We suggest that this limit should be evaluated in a case-by-case scenario, depending on the application and with the aid of a domain expert. Using the same framework described so far, we can explain why the models diverge.

Let F be a set of f features, S denote a coalition of features, and ℓ be a characteristic function over the loss function L. That is, $\ell(S)$ denotes the worth of a coalition S and describes the total expected loss that the members of S obtain. If we consider a pair of models to be a black box and the Jensen-Shannon distance between the pair predictions as a loss function, then computing the Shapley values of the Jensen-Shannon distance concerning both model's input features is akin to finding the optimal payoff for $\ell(S)$ following the Shapley axioms. We can encapsule the output of each model, creating a function $\ell(S)$ that expresses the Jensen-Shannon distance between any two models recieving as input F. This enables SHAP to explain the distance between individual model's predictions given the data, and is similar to what was proposed by Lundberg et al. [2020], monitoring a model's loss employing SHAP and observing the feature importance. Figure 4.9 illustrates an example of employing the SHAP algorithm to explain the difference in model predictions for a given instance. Because of complex relationships that arise when training a model, the explanation of prediction difference is not the same as the difference in prediction explanation.

Figure 4.9: SHAP explanations for a sample target and anchor model, alongside the explanation for the difference between predictions. Single production instance for the Alzheimer dataset, LightGBM with random seeds 114 and 3452.



(a) Target model.

(b) Anchor model.

(c) Jensen-Shannon.

4.2 Learning Rashomon Ensembles

In the presence of problems with many possible contrasting or competing explanations, employing the Rashomon sets as a method for obtaining ensemble constituents can be useful. Even in the absence of such structures, diversity is a desirable characteristic for any ensemble as it allows the end model to cover a wider region of the solution space. To support this statement and to verify whether Rashomon sets provide a suitable tool for model space partitioning, we propose splitting the Rashomon space into clusters, grouped by the explainability vectors of each model, and creating ensembles composed solely of models located close to centers of each Rashomon subgroup. We consider the K-Means algorithm to induce clusters, performing silhouette to obtain optimal K. We expand upon our previous datasets by considering a series of open-source datasets, acquired from the UCI machine learning repository [Asuncion and Newman, 2007] and the OpenML database [Bischl et al., 2017]. Following is a brief description of each problem in our benchmark suite:

APS Failure: this is the dataset used for the 2016 IDA Industrial Challenge [Costa and Nascimento, 2016]. It consists of data collected from heavy Scania trucks in everyday usage. The problem is formulated as a binary classification task consisting of component failures for a specific component of the APS system

Diabetes readmission: this dataset was submitted on behalf of the Center for Clinical and Translational Research, Virginia Commonwealth University [Strack et al., 2014], representing 10 years of clinical care at 130 US hospitals and integrated delivery networks. The problem is formulated as a binary classification task predicting 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 [Aha and Kibler, 1988]. From several indicators such as age or pain profiles, 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 is an artificial dataset containing data points grouped in 32 clusters placed on the vertices of a five-dimensional hypercube and randomly labeled +1 or -1, being one of five datasets used in the NIPS 2003 feature selection challenge [Guyon et al., 2004]. The problem is formulated as a binary classification task separating 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) [Bock et al., 2004]). The problem is formulated as a binary classification task discriminating 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 [Olave et al., 1989]. The final decision depended on three subproblems: the occupation of the parents and child's nursery, family structure and financial standing, and the social and health picture of the family. 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 [Fisman et al., 2006]. During the events, the attendees would have a four-minute "first date" with every other participant of the opposite sex. At the end of their four minutes, participants were asked if they would like to see their date again. The problem was formulated as a binary classification task in which, given each participant's questionary responses and characteristics, the goal is to predict whether both participants would like to date each other again.

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 [Street et al., 1993]. The problem was formulated as a binary classification task, in which the end goal is to predict the presence of malignant tumor cells.

Wine quality: this dataset is composed of chemical analysis of wines grown in the same region in Italy but derived from three different cultivars [Aeberhard et al., 1992]. The analysis determined the quantities of 13 constituents found in each of the three types of wines. The end goal is to predict the wine quality score, ranging from 0.0 to 8.0, given its chemical characteristics. 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.

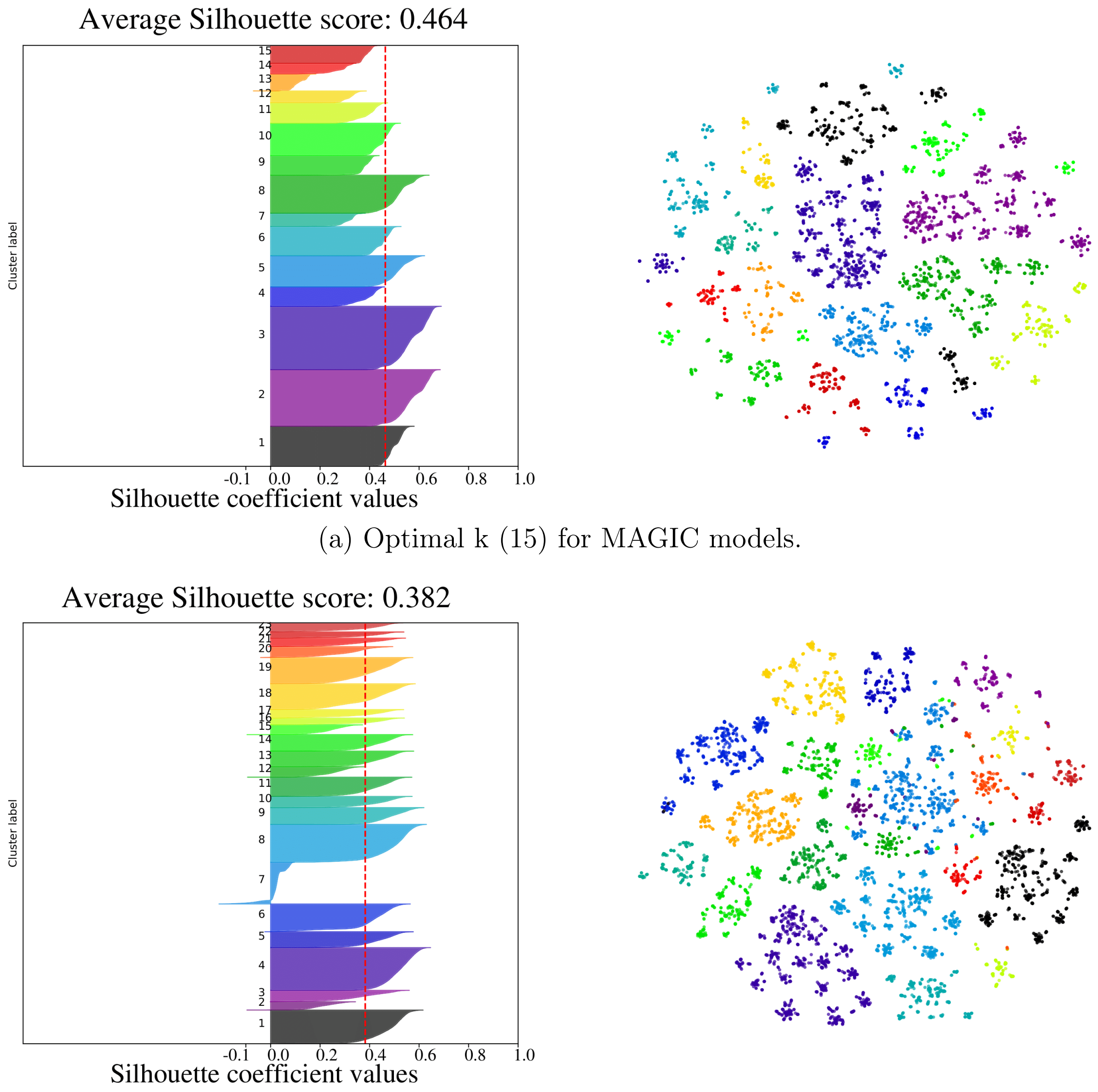
Table 4.2 summarizes our comparison between our approach and classic and state-of-the-art algorithms. In our experiments, we sample 100 000 decision trees to guarantee a minimum subset diversity and train a meta-model to combine constituent outputs in a stacking ensemble. Figures 4.10a and 4.10b illustrate some Rashomon subspaces.

Table 4.2: Benchmark suite results.

Benchmark			Baseline Algorithm							Rashomon	
Dataset	Instances	Features	Decision Tree	AdaBoost	GBM	Random Forest	XGBoost	LGBM	CatBoost	Ensemble	Ratio size
APS Failure	76000	172	.866	.824	.806	.869	.835	.853	.888	**.911**	12.4%
Diabetes	101766	1691	.544	.614	.609	.599	.615	.616	**.619**	**.618**	17.4%
Heart	303	171	.748	.787	.793	.826	.796	.830	.834	**.839**	50.3%
MADELON	2000	502	.764	.598	.782	.694	.828	.832	**.852**	*.746*	< 0.5%
MAGIC	19020	102	.808	.830	.812	**.857**	.837	.850	.850	.848	19.4%
Nursery	12630	784	.999	**.999**	.964	**.999**	.991	**.999**	**.999**	**.999**	83.2%
Speeddating	8378	123	.650	**.673**	.619	.630	.639	.642	.668	*.632*	< 0.5%
WDBC	569	903	.949	**.973**	.956	.967	.963	.967	**.974**	**.974**	21.5%
Wine	4898	13	.762	.722	.723	.802	.755	.764	.782	**.805**	8.9%

The poor performance of our approach on the Speeddating 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. As such, our following experiment focuses on comparing MAGIC to the remaining datasets, as a means

Figure 4.10: TSNE reduction of the Rashomon space and optimal silhouette for the Rashomon subgroups.



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

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

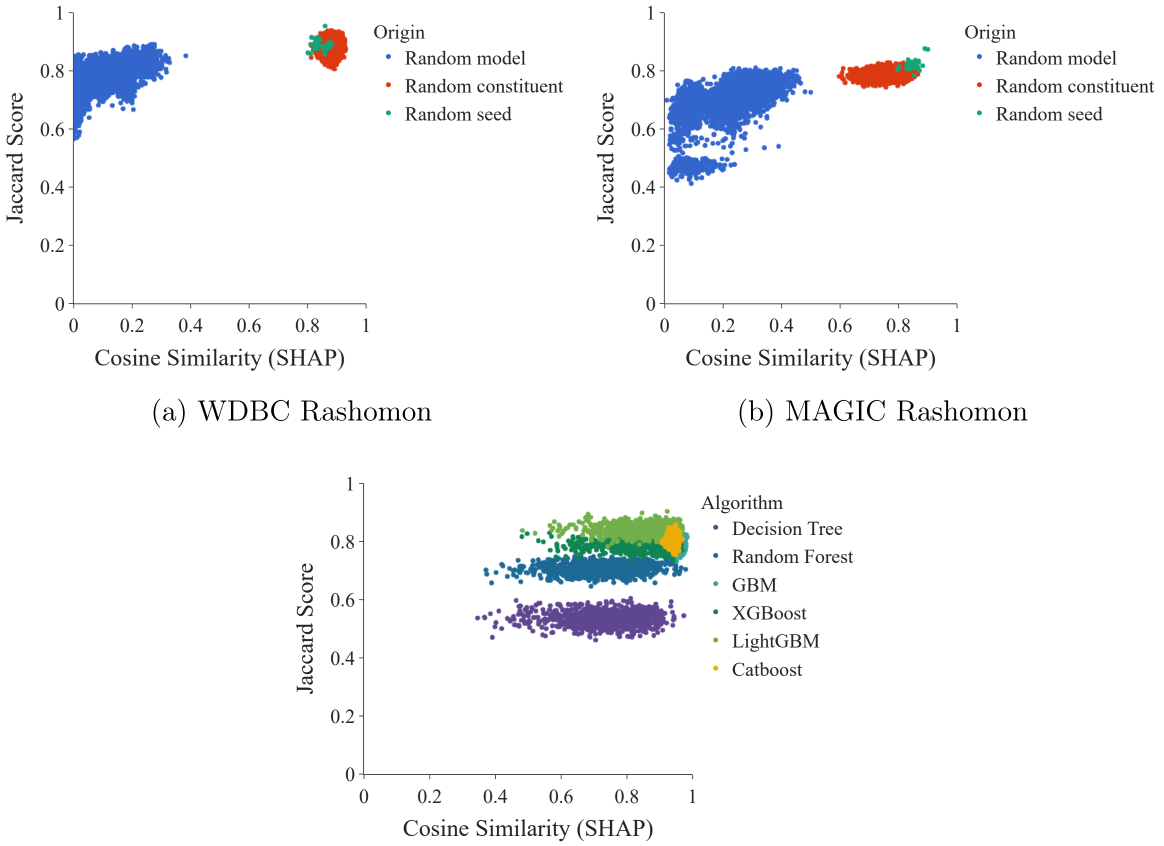
to both comprehend what makes this problem unique and grasp a better understanding of our Rashomon ensemble learning technique. Taking into consideration the time complexity to perform each experiment, we evaluate the following scenarios in Figure 4.11:

I) Assess the choice of the random seed, thus replicating the whole Rashomon pipeline (30 runs, green points).

II) Maintain the Rashomon sets found previously, but select models from each cluster to represent ensemble constituents (10,000 runs, red points).

III) Ignore the Rashomon subgroups, selecting models from the whole Rashomon set to represent ensemble constituents. (10,000 runs, blue points).

Figure 4.11: Similarity to a reference model found from running the Rashomon pipeline, filtering models with statistically worse performance.



(a) WDBC Rashomon

(b) MAGIC Rashomon

(c) MAGIC other pipelines)

There is a significant overhead of running the whole proposed pipeline in comparison to arbitrarily selecting a model from the Rashomon pool to pose as a constituent. For this simple reason, we do not perform the same amount of runs for each scenario. As discussed by D'Amour, only observing the performance of these models poses an ineffective way to judge underspecification. We propose an alternative view in Figure 4.11 comparing the models of each scenario with the one previously found (performance on Table 4.2). For better visualization, we filter all models of III with statistically inferior performance to the said reference model. Our visualization scheme relies on a joint observation of the Jaccard Index and SHAP values similarity between the reference model and the ones found in each scenario. The Jaccard metric is a useful way of assessing the degree of agreement between two prediction sets given the relation between their intersection and union sizes. This provides an interpretation of the minor differences between predictions that are overlooked by standard performance metrics. The comparison between the Shapley vectors, in turn, shows us the robustness of the explanations and whether the explicatory factors remain unchanged under the stochastic behavior of algorithms

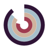
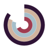
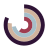
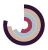
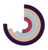
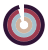
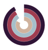
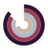
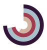
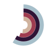
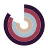
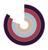
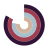
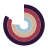
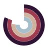
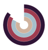
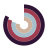
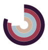
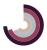
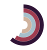
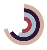
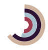
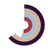
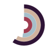
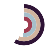
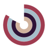
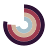
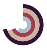
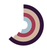
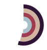
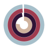
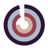
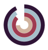
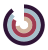
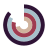
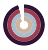
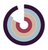
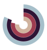
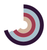
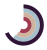
When exploring drift, we verified that Group III's population represents a sample of several ensembles that can be reached by direct optimizations over their constituents. On the opposite extreme, Group I represent the ensembles found following our proposed pipeline. It is important to remark that our approach depends on sampling the extremely complex model space. Thus, it is highly unlikely that the clusteroids found in each repetition are the same. However, the high Jaccard coefficient associated with the high cosine similarity between the SHAP vectors provides a shred of strong evidence that the centroids found in each repetition are contiguous, resulting in similar clusteroids that lead to similar ensembles. Finally, Group II represents 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 sparser point cloud. Groups I and II were more cohesive and concentrated over high values of similarity with the reference model. When we consider that all models have a statistically equal or higher performance than the reference model, it is reasonable to conclude that the pipeline involving Rashomon sets reduces the impact of data drift while retaining concise predictions. When further exploring phenomenons akin to underspecification by both introducing Gaussian noise and shuffling feature values, the robustness of Rashomon ensembles becomes evident. In most explored scenarios, our approach remained the performant model even when considering the MAGIC dataset in which Rashomon ensembles had slightly worse original performance than other ensembling approaches.

4.2.1 Performance as Generator Function Diverges

Addressing the concerns of out-of-distribution data, we evaluated the performance impact of models on distributions that diverge from the trained one. We considered two scenarios, the first being the addition of Gaussian noise with increasing values of σ^2 as a means to simulate data drift. In the second scenario, we shuffled the values of features in such a way that their distribution stays the same, but any correlation to the target variable is lost. This second scenario aims to evaluate how heavily models rely on core key features and whether they can extrapolate from global information rather than local aspects, which bears similarities to the notion of Model Reliance from Fisher et al. [2019]. Table 4.3 presents a comparative analysis of each approach in the form of ring plots ordered by performance and whose radius represents the respective mean AUROC after 30 repetitions.

Table 4.3: Performance loss comparison between Random Forest ■, LightGBM ■, CatBoost ■ and Rashomon ensembles ■.

	Data Drift (σ^2)					Data Shuffle (n)				
	0.4	0.8	1.2	1.6	2.0	10%	30%	50%	70%	90%
APS										
Heart										
MAGIC										
Nursery										
WDBC										

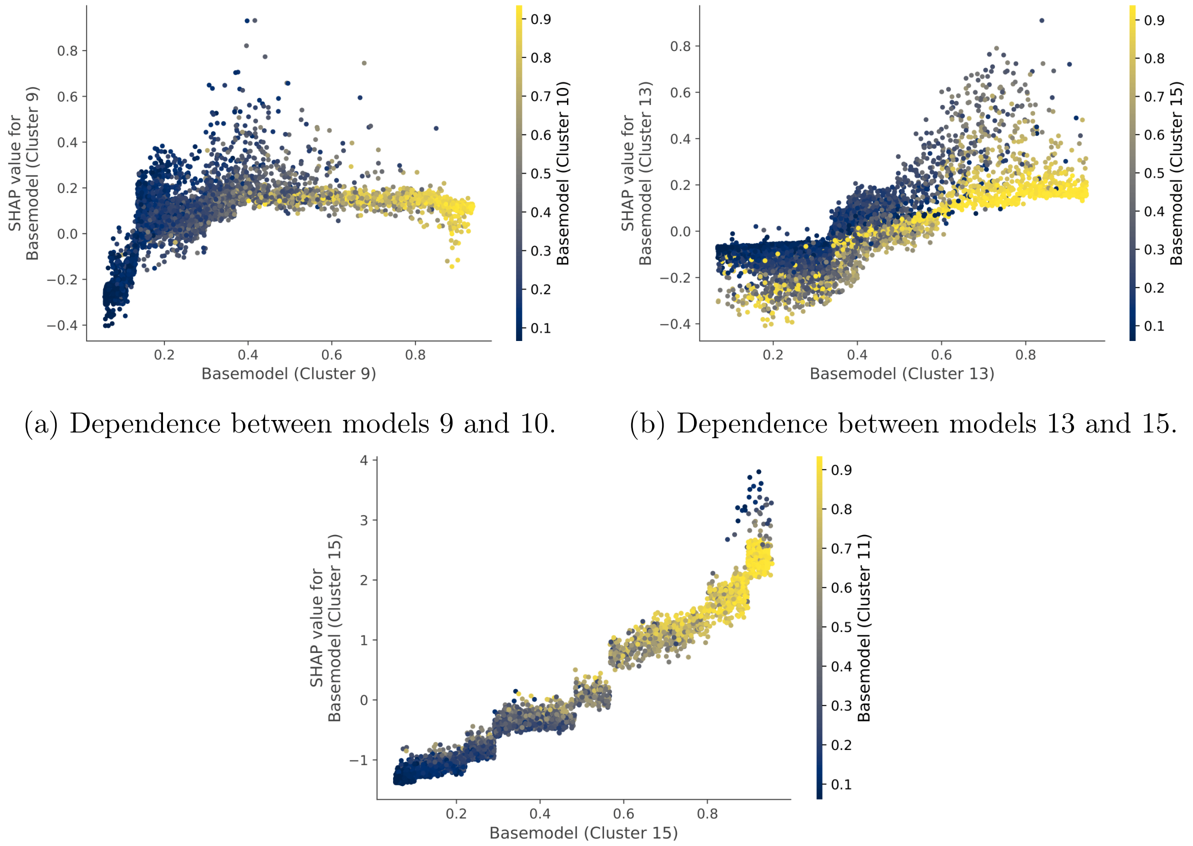
4.2.2 Intra-model Associations

When further investigating the relationships learned, a variety of interesting patterns can be observed. For instance, in Figure 4.12a we observe that the ensemble learns to employ the output of the 9th constituent model to give mostly positive predictions, with nearly all points above the 0.2 probability threshold presenting a positive SHAP. We also verify that models 9 and 10 provide a complementary view of the problem, as we can observe that higher prediction values of 9 and likewise associated with high prediction values of 10. Figure 4.12b on the other hand shows that models 13 and 15 are mostly contrasting. Whenever there is disagreement, the relative importance of model 13 increases. Similarly, model 13 presents Shapley values close to zero when both models agree. These give rise to a concentration of yellow points on the lower side of the distribution. Figure 4.12c illustrates that models 15 and 11 are mostly complementary, except for a distinct set of points. This pattern suggests that model 15 is specialized in solving these instances, for which model 11 gives a low positive probability and which model 15 presents both high Shapley and prediction values.

4.2.3 Ensemble Behavior under Train and Production Divergence

We look back to the COVID-19 and Alzheimer dataset pairs for our last set of experiments. Once again, to ensure fairness to the previously evaluated ensemble algo-

Figure 4.12: Dependence between relevant base models in the MAGIC Rashomon ensemble.



(a) Dependence between models 9 and 10.

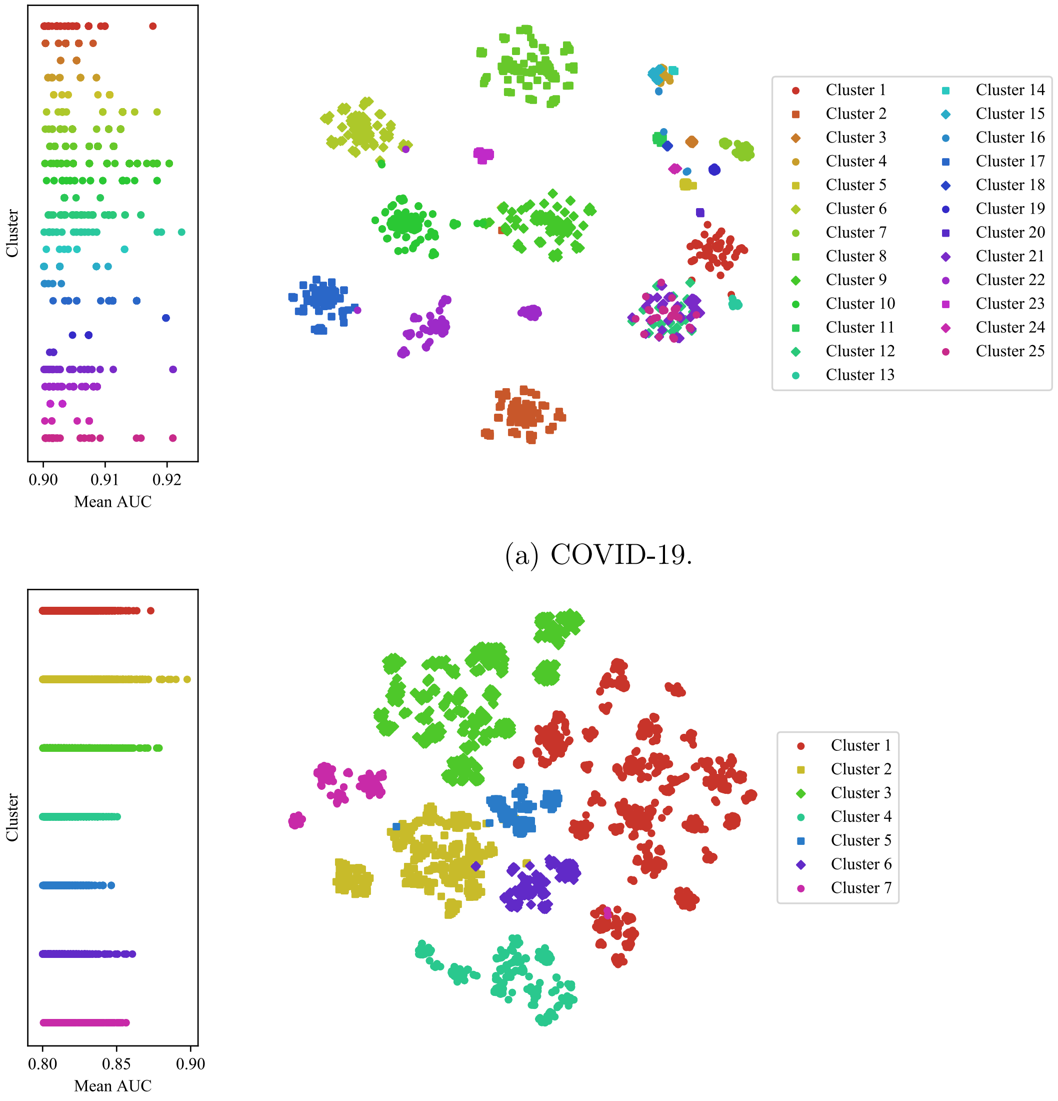
(b) Dependence between models 13 and 15.

(c) Dependence between models 15 and 11.

rithms, we consider decision trees as base constituents. Since there is no strong baseline in the literature, we used the all-in-one approach to provide a reference f_{ref} and ϵ value. The average AUROC values obtained by the all-in-one model were 0.90 and 0.81 for the two dataset pairs respectively, which were used as a performance threshold to consider a model minimally performant and establish the Rashomon set. This resulted in a sub-space containing 2 554 and 6 251 models out of the original 100 000 sample, thus presenting a Rashomon ratio of 2.5% and 6.2%. Figure 4.13 illustrates the Rashomon subspaces found after clustering, as well as the performance of each constituent. We verified that all clusters have a strong completely connected component. That is a set of key features that are present in all models encompassed within the same cluster. This matches our previous hypothesis. When evaluating model performance, we observe no strong relationship between group labeling and individual model loss, thus suggesting the multiplicity of feasible explanations, the Rashomon Effect.

We proceed to search for constituents inside each Rashomon subgroup. We consider two ensembling techniques, one in which we learn a meta-model that combines the outputs of each base model (Stacking, denoted by S) and another where we return as agreement ratio of the predicted classes by each base model (Voting, denoted by V). The advantage

Figure 4.13: TSNE visualization of the Rashomon space of each problem.



(a) COVID-19.

(b) Alzheimers'.

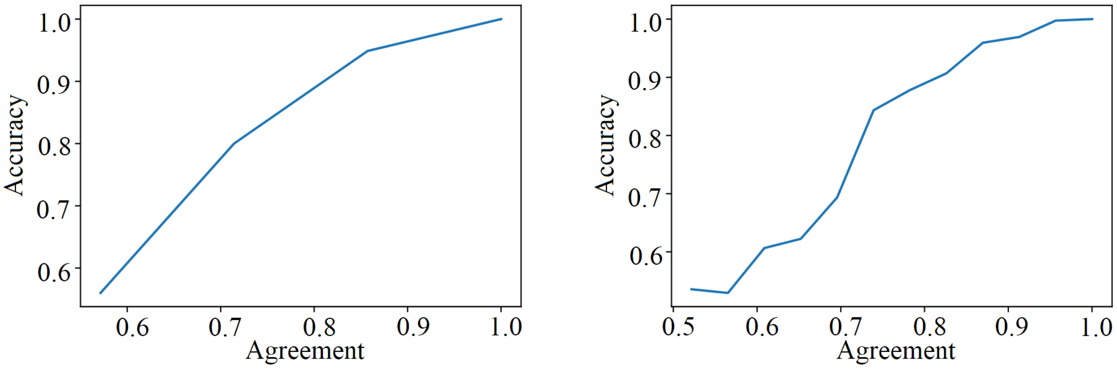
of the first approach is that the meta-model can learn to optimally combine base model outputs. For instance, if some model outputs a low probability for a given class, the meta-model can learn to ignore the said prediction. The main drawback is that this meta-model is prone to some of the same biases from the base models, and we likewise have no guarantee of performance on distributions that diverge from the train data.

Under the voting approach, since each model is trained on different inductive biases, we can expect noise in individual base model output to be smoothed out. Further, the ensemble output is a direct measure of prediction reliability. The agreement ratio can be seen as the returned probability by the ensemble. Probabilities close to the upper bounds imply that nearly all models agreed on the same predictions. Likewise, probabilities close to the lower bounds imply a large disagreement. This is a desirable property, as it allows for quick expert judging without the need for extra analysis.

Our final set of experiments aims to understand the relationship between model agreement and confidence. The main hypothesis behind the proposed ensemble approach is that experts can only trust the provided predictions if models agree. Figure 4.14 illustrates the comparison between ensemble accuracy and constituent agreement, and we can infer a direct relationship between these two metrics. Since we formulated the problem as a binary classification task, the agreement can never fall below 50%. We verify performance close to random guessing when near this threshold. We also observe accuracy close to 100% when all constituents agree. We can compute the calibrated prediction reliability from these curves.

Figure 4.15 illustrate the performance of each base model and the ensembles on the COVID and Alzheimer dataset. As expected, all constituents present statistically equal performance under the training dataset. However, once presented with new data, their behavior becomes erratic. Comparing the two proposed ensemble approaches, their performances are comparable on the train data, but voting outperforms stacking on new distributions. It is worth mentioning that both approaches always presented performance superior to the mean on the constituents. Voting was able to outperform all base models under all evaluated scenarios, as well as the state-of-the-art methods presented in Table 4.1.

Figure 4.14: Relationship between Rashomon Ensemble accuracy and intra-constituent agreement. As we hypothesized, there exists a direct correlation between ensemble performance and agreement. When constituents agree, accuracy lies close to 1 implying a similarity to training and that predictions are trustable. When constituents disagree, the observed instance likely diverges from what was learned resulting in untrustable predictions. The ensemble degenerates to random guessing with accuracy close to 0.5.

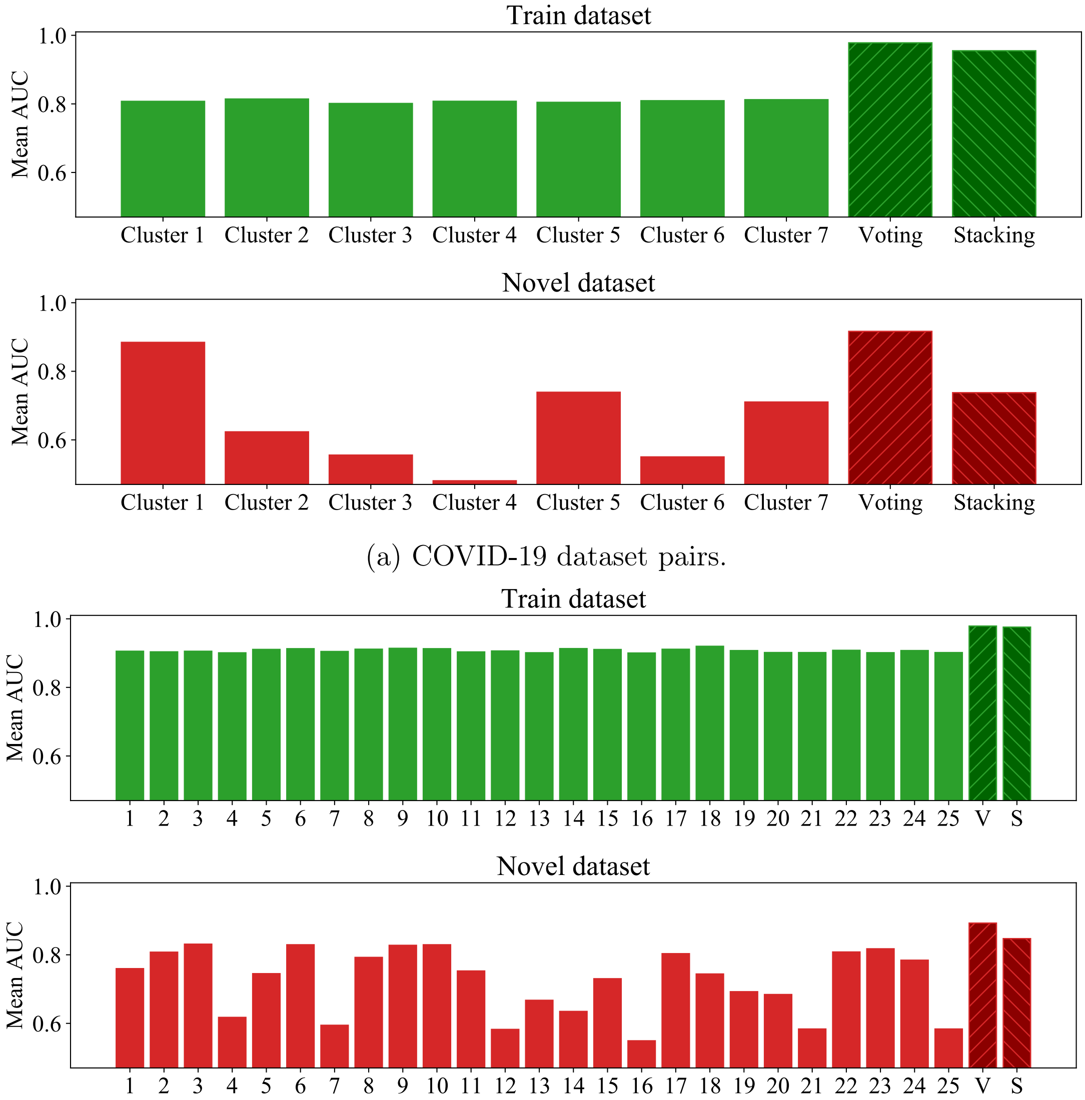


(a) COVID-19 production dataset.

(b) Alzheimer's production dataset.

A potential solution to mitigate drift is to utilize multiple models in production, however, this selection process can be challenging as we have outlined certain limitations in the choice of reference models. In many cases, the data comprises multiple local structures and sub-populations. In these circumstances, it is beneficial to make use of local structures for the induction of models that are more reliable and in line with the data. We contend that each local structure can be correlated with domains and, utilizing model explanation

Figure 4.15: Comparison of model performances across evaluated datasets. Each constituent model is represented by the Cluster from which it hailed. In the train datasets, we can observe that all constituents model behave similarly. On the novel datasets, under the unknown U distribution, performance becomes unpredictable. However, in our empirical experiments, we verify that the voting approach outperforms the best constituent model, thus suggesting itself as a suitable technique to smooth individual model erratic behavior.



(a) COVID-19 dataset pairs.

(b) Alzheimer's' dataset pairs.

techniques, we can distinguish contrasting plausible explanations for the studied problem, in accordance with the Rashomon Effect. Our approach demonstrated consistent gains in AUROC compared to other tree-based ensemble techniques in scenarios where such multiple local structures are expected. In situations where the generator function at production time may differ from that seen during training, we observed a direct correlation between accuracy and model consensus. That is, if the models concur, the accuracy is high, and ensenble predictions can be trusted. Conversely, if the agreement between models is low, the accuracy is also low and their predictions should not be trusted, which was our intended outcome and demonstrates the robustness of our approach. All these characteristics are valuable in real-world industrial scenarios, as described in the following chapters.

Chapter 5

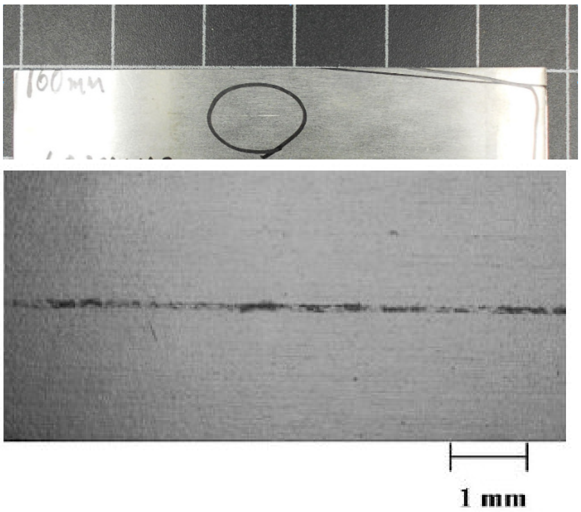
Case Study: Surface Defects in Stainless Steel Manufacturing

Initial applications for duplex stainless steel materials were almost exclusively heat exchanger tubing, particularly in corrosive cooling water services, and shafting or forgings. Currently, these steel materials have a great variety of potential applications as discussed by Davis and Committee [1994], Gunn [1997]. In particular, duplex stainless steels are finding increasing use in the offshore industry, primarily because they often offer an economical combination of strength and corrosion resistance. The duplex stainless steel most commonly used today in deepwater exploration includes those with ≈22% chromium (Cr), which usually also contains more molybdenum (Mo) and nitrogen (N). Other elements that may be included in this steel material are nickel (Ni), copper (Cu), manganese (Mn), phosphorus (P), sulfur (S), boron (B), niobium (Nb), and silicon (Si). These elements must occur in certain specified ranges to properly characterize the finished material as duplex stainless steel [Davison and Redmond, 1990].

The quality of duplex stainless steel is often threatened by the presence of surface defects, specially slivers, which are postulated to originate during the solidification stages of the casting process [Stradomska et al., 2009]. Such defects are elongated in the direction of lamination, with a usual length of 70mm, and they are generally concentrated closer to the edges of the steel plates [Thomas, 2006], as shown in Figure 5.1. Slivers increase production costs as they remain undetected in intermediate processing stages, being observed only during the final inspection of the finished product. Once slivers are observed the defective plate is usually discarded, and thus the design of new steel materials that are less susceptible to sliver formation is of paramount importance.

With the expansion of the sub-salt oil exploration segment, there is an increasing need for ultra-resistant stainless steel materials. These novel steel materials are imposed on high-quality standards to withstand the extremely adverse deepwater environmental conditions, and a major challenge for achieving the required quality standards is slivering. To determine the relationship between sliver formation and factors associated with the production of duplex stainless steel, in partnership with *APERAM South America* we created a dataset containing the chemical compositions and metallurgical process variables of 122 duplex stainless steel production runs, from which 71 presented at least one defective plate. This corresponds to a dataset with nearly 500 stainless steel plates for studying the slivering problem. Nevertheless, according to Barbosa et al. [2007], determining the causative factors associated with sliver formation is not trivial, because slivers can be associated with either combination of process variables or chemical compositions, and these variables are obtained at different steps of the steelmaking process.

Figure 5.1: Top – Duplex stainless steel plate presenting slivers near its edge. Bottom – A zoomed image of the slivers.



Intuitively, if different data points (i.e., steel plates) are associated with different local structures in the data, then we would expect each structure to be better described by a different model. In this case, instead of modeling the data using the standard all-in-one approach which fits all the available factors (or features) into a single model, we wish to obtain models associated with contrasting possible causes. In high dimensionality problems, such as the prediction of heating sliver, attempting to perform direct inference of the possible explanations is ineffective, as shown in our experiments by the poor performance of all-in-one models. The main hypothesis of this work is that we can isolate the various explanations by considering only subgroups of correlated features. We believe that models with similar responses and similar feature importance distributions are likewise associated with similar effects. During our analysis, we found a strong link between features and model predictions, showing that some features are tailored for detecting different sliver formation mechanisms and cover a specific region of the defect space, as presented in our work [Zuin et al., 2021].

While there are several guidelines regarding which components are related to sliver formation in duplex stainless steel [Chai and Kangas, 2014], to the best of our knowledge, an in-depth analysis using a unique dataset of scale and considering the entire spectrum of chemical components and process variables has never been done. Further, the novelty of this work lies in the framework proposed to learn multiple contrasting explanations. The specific contributions of this case study are two-fold:

- Among our main results, we emphasize that our predictive models are empirically accurate for estimating whether an arbitrary steel plate will be defective, with an AUC score that ranges from 0.78 to 0.85. This can be considered a relevant result, as the ability to anticipate defects right on the first steps of the steelmaking process is of great importance for reducing operational costs.

- Our novel methodology employs a large number of predictive models to find a diverse set of patterns that are associated with non-defective plates.

5.1 Defects and Data

Steelmaking is the process of producing steel from iron ore and scrap. The process involves removing impurities such as nitrogen, silicon, phosphorus, sulfur, and excess carbon from the raw iron, and adding alloying elements such as manganese, nickel, chromium, and vanadium to produce different grades of steel [Deo and Boom, 1993]. As illustrated in Figure 5.2, the process can be divided into seven main steps:

Initial Casting: in the first step of the steelmaking process we have primordial ferrous materials coming from two sources: pig iron, a product obtained directly from the iron ore transformed into blast furnaces which have roughly 5% carbon in its composition; and recycled scrap, materials obtained from either external acquisitions or the scrapping of other materials provided from the production line. The resulting material constitutes a mixture of liquid pig iron and scrap.

Adjusting Steel Composition: The second step involves melting, purifying, and alloying operations carried out at approximately 1 600°C (2 900°F) in molten conditions. Various chemical reactions are initiated, either in sequence or simultaneously, to arrive at the specified chemical compositions of the duplex stainless steel. The chemical composition adjustment process is done by adding or removing certain elements and/or manipulating the temperature and pressure and production environment.

AOD (Argon-Oxygen-Decarburization): the start of the melting process consists of taking the mixture to the AOD, where a significant reduction of the partial pressure of the system takes place. Alongside an intense blow of oxygen, a drastic reduction of the carbon content of the mixture takes place. Tons of substances that have a chemical affinity with the carbon are complemented, also to reduce the carbon content. Relevant chemical data is mainly collected in this stage through spectrometers. Two relevant samplings are associated with the chemical composition in AOD. The first and most immediate is made of the gases emitted during the process. The second sampling is done through the use of spectrometers and allows the generation of more detailed data about its chemical composition.

VOD (Vacuum-Oxygen-Decarburization): during the VOD stage, the reduction of the partial pressure of the system is much more intense, reaching what is considered absolute vacuum. These conditions allow a more controlled reduction of carbon content. While in the AOD stage the additions are made in the order of tons, in the VOD stage additions are done in pounds. To be carried out accurately, the VOD operations depend on the ability to measure the composition of the stream of gas leaving the furnace throughout the process. Sampling systems based on mass spectrometers provide precise, representative, and real-time analysis of the chemical composition of the steel being produced.

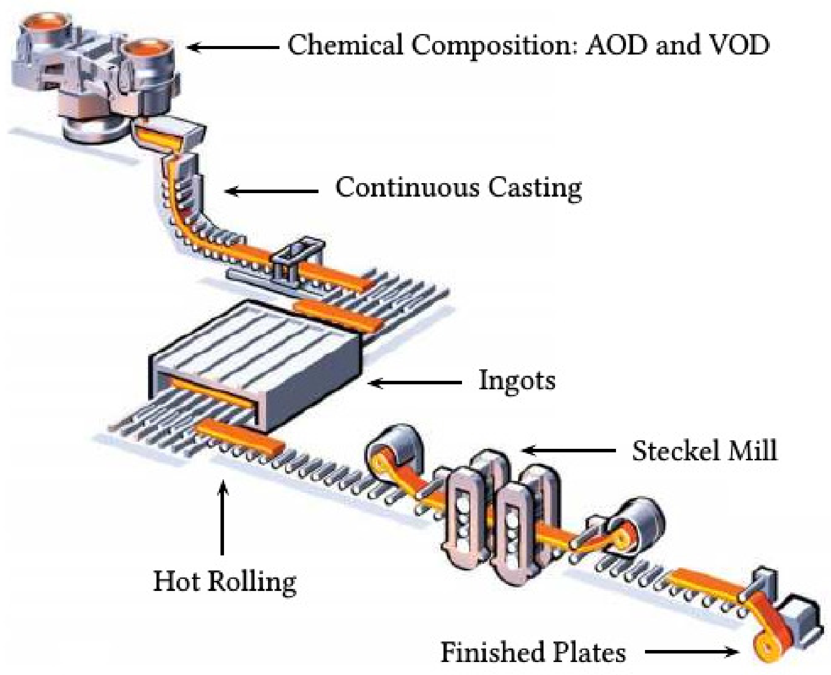
Continuous Casting: the refined steel is reheated in a pan oven before being taken to continuous casting. More specifically, refined steel is taken to cooling molds and transformed into semi-finished and solidified plates. From this moment, the steel ingots can be rolled according to the specifications of each customer. During the solidifying moment, a pair of scissors cut the ingots and it is already possible to observe the steel in lengths useful for the remaining processes.

Steckel Mill Hot Rolling: steel, during the process of solidification, is mechanically conformed and transformed into duplex stainless steel products used by the transformation industry, such as heavy and thin plates and coils. Specifically, the duplex stainless steel is conformed into plates and coils by hot rolling using a Steckel mill, which is a reversing mill with a heated coiler at each end, and the two coilers are used to feed the material through the mill. The material is fed back and forth through the mill until a precise thickness across the full width is reached, as well as consistent flatness. Steel thickness is drastically reduced while the length is expanded.

Visual Inspection after rolling, duplex stainless steel plates are moved to inspection facilities. The identification and registration of any defects that the plates may have, as well as their location and extent, are made by trained experts [Neogi et al., 2014, Zhao et al., 2017]. In particular, this step is where the occurrence of the sliver defect is reported. Experts can verify the occurrence of defects only after the whole process is completed, thus expressing the value of a predictive approach to defect formation.

Our working dataset consists of the chemical composition of duplex stainless steel plates and hot rolling process variables measured during the steelmaking process. Spectrometers were used to assess the relative abundances (%) for each element in a given plate. Each element has a particular spectrum pattern, and thus peaks in the spectra are associated with specific elements, based on comparison with reference sample results. In total, we have 20 chemical elements that are measured in different stages of the steelmaking process. We also consider the ratios between these elements as features, thus extending the evaluated feature space to 220 attributes. The hot rolling variables are composed of 1 160 temporal series related to each aspect of the steelmaking process. Since we wish to evaluate their importance alongside the chemical features, we discretize the data by cal-

Figure 5.2: The steelmaking process.



culating the momentum of each variable in 30-second spans accessing averages, kurtosis, variances, and skewness. After filtering non-actionable variables, we obtained 11 488 hot rolling features.

The performance of machine learning methods is heavily dependent on the choice of features on which they are applied [Forman et al., 2003]. For this reason, much of the current effort in deploying such algorithms goes into the design of preprocessing pipelines and data transformations that result in a representation of data that can support effective machine learning [Forman et al., 2003, LeCun et al., 2015, Leiner et al., 2019]. The process of using available features to create additional ones to improve model performance is often called 'feature engineering', a predominantly human-intensive and time-consuming step that is central to the data science workflow. It is a complex exercise, performed in an iterative manner with trial and error, and mostly driven by domain knowledge [Gada et al., 2021]. Recently, many studies have shown the benefits of automatizing this process by creating candidate features in a domain-independent and data-driven manner followed by an effective method of feature selection. This way it is possible not only to improve model correctness but also to discover powerful new features and processes that could be additional candidates for domain-specific studies [Gada et al., 2021, Kaul et al., 2017, Sumonja et al., 2019]. We avoid potential spurious correlations by confirming that all selected features present a strictly non-zero impact on model output after n-fold cross-validation.

Data was collected during the entire year of 2018 and led to a dataset composed of 122 steelmaking runs, each with 4 duplex stainless steel plates on average, and with 11 708 possible (chemical and hot rolling) features. There were 71 defective runs (due to slivering) with the remaining 51 being non-defective ones. Although the chemical compositions are approximately the same for all plates in a given run, the hot rolling features vary constantly throughout the process. This led to a dataset composed of 499 duplex plates. The proposed task corresponds to solving the binary classification problem regarding the formation of slivers given both chemical and hot-rolling features.

5.2 Building a Rashomon Ensemble

As described in Chapter 3, the first step towards building a Rashomon Ensemble consists of sampling the complete model space. Due to the nature of this dataset, simple random sampling features would lead to process features overshadowing the importance of chemical ones. Because chemical features are less numerous than hot-rolling ones, we modified the selection probabilities to guarantee that a model is equally likely to choose either a chemical or hot-rolling feature. Otherwise, our model sampling approach would be biased towards solely hot-rolling causes. We also had two constraints for algorithm selection. The first one is that we aimed to discover *pure* models, that is, models with no concurrent explanations introduced into the stand-alone models. The second constraint is that we needed a human-understandable model to derive a set of rules and standards to be employed, as the project's end goal is heating sliver prevention rather than real-time prediction. To achieve both of these objectives, we opted out of ensembling approaches and employed Decision Trees [Pedregosa et al., 2011a].

We sampled 75 000 models for each possible feature set size, until no significant gain in performance could be verified by including additional features into a model. To evaluate performance, we utilized the standard AUC (area under the ROC curve) measure [Fawcett, 2006, Hanley and McNeil, 1982] with five-fold cross-validation. There exists diminishing gains in performance for models with more than 15 features, as seen in Figure 5.3. We limited our experiments to this threshold, which led to a sample total of 1 049 999 models.

The next step consisted in establishing the Rashomon set in terms of a reference model. Since there is no strong baseline in the literature, we used the all-in-one approach to provide a baseline comparison. The all-in-one employs the same implementation of the sampled models but produces only a single model composed of all available features. The average AUC value obtained by the all-in-one model was 0.62, which was used as a performance threshold to consider a model minimally performant and establish the Rashomon set. This resulted in a sampled model space $\mathcal{H}'$ containing 63 374 models out

Figure 5.3: AUC values of sampled models. Dashed lines represent an increase in set size.

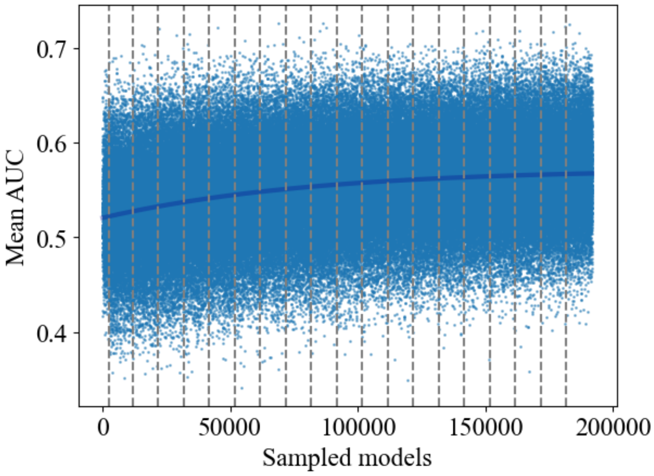
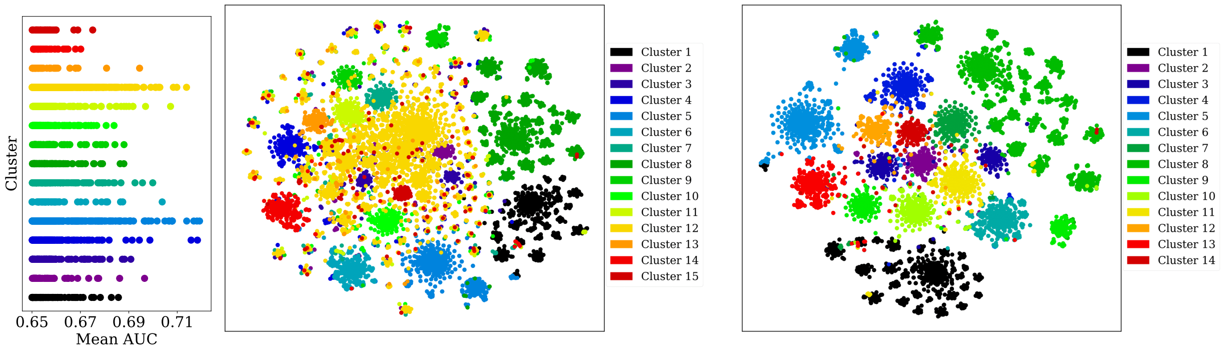


Figure 5.4: T-SNE visualization of the sampled model space $\mathcal{H}'$. Each point represents a model $\mathbf{x}'$. Models are placed according to the defect explanations assigned to each steel plate so that models that possess similar SHAP values are placed next to each other in space (see Section 4.4). The color indicates the cluster for which the model was assigned. N = 2 500 models.



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

of the original 1 049 999 (6.04% of the models perform better than the all-in-one model). For better visualization, we show only 4% models in Figure 5.4a, which sums up to 2 500 points. Later in this chapter, we evaluate both state-of-the-art and classical models for our ensembling baseline and present a performance comparison with and without feature selection to reduce input space dimensionality.

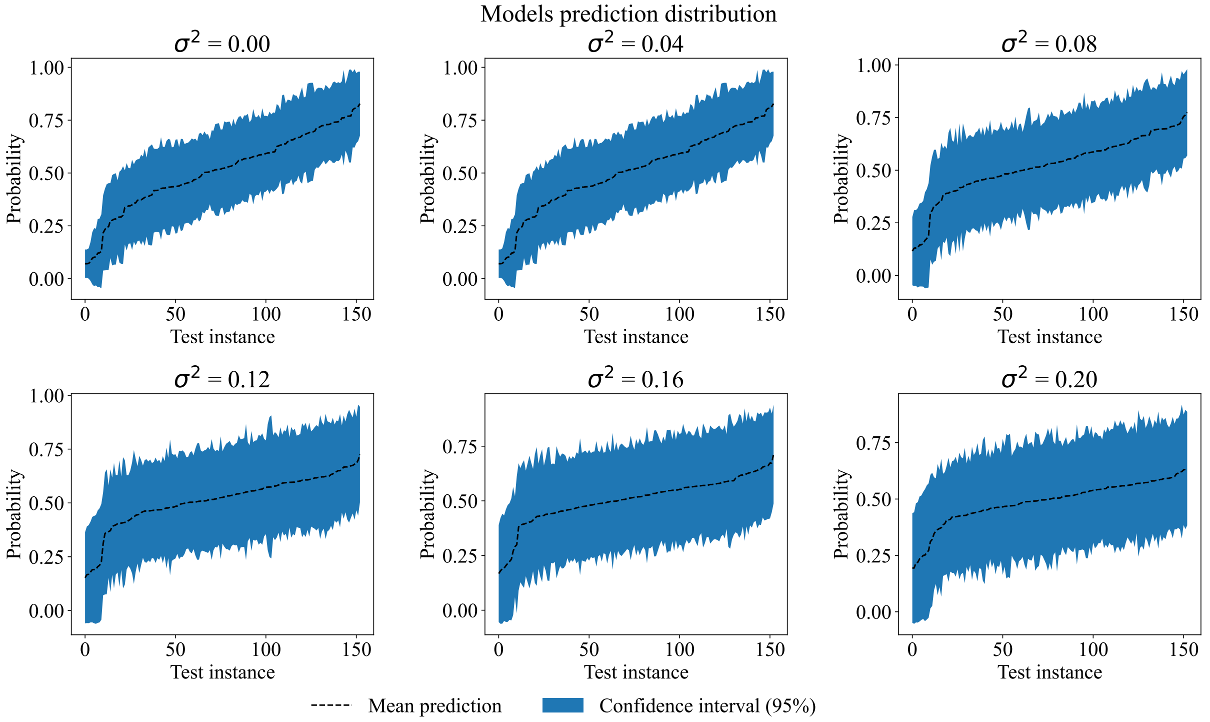
We assumed a learning scenario in which models can be mapped to explanation domains (C), enabling us to learn specific models for each domain. The constituent selection step of the Rashomon ensemble consists in searching the model space. However, to reduce search space, we only considered models associated with each domain $c \in C$ alongside a feature relationship graph. We believe two features to be related if they co-occur in the same performant model and both have a statistically significant impact on model explanation. This entails only complementary or contrasting features being related to one another. Organizing models under explanations, such as the division seen in Figure 5.4, enable us to acquire feature relationship graphs for each specific grouping and feature vectors $\mathbf{x_c} \subseteq \mathbf{x}$. The dual problem of learning distinct explanation models consisted in finding optimal paths in the feature relationship graphs. We solved this task by minimizing different functions $f(\mathbf{x'_c})$, such that $\mathbf{x'_c} \subseteq \mathbf{x_c}$ and $|\mathbf{x'_c}| \ll |\mathbf{x_c}|$ and, unlike in the model sampling approach, selecting features to compose each model $\mathbf{x'_c}$ in accordance to their relationship. Section 5.3 describes with more detail our searching approach for the heating sliver problem.

Once models are selected, it is important to verify their suitability as Rashomon constituents. Our main hypothesis is that by exploiting models that behave differently under data drift and that encompass diverse explanations, we can build a more robust ensemble. Since our proposed clustering of the Rashomon space guaranteed diversity in explanations, a pivotal next step was verifying drift response empirically. In Figure 5.5, we introduced increasing amounts of gaussian noise to the normalized features and observed each constituent's returned probability distributions under each scenario. We verified a direct relationship between noise and the confidence interval, thus signifying that models become more divergent under drift and ensemble reliability decreases, as intended. This suggested that our constituent selection was appropriate and that we could employ them in production. In this work, we did not have access to a separate dataset that could be subject to drift. However, our model was deployed into the APERAM South America and used to predict heating slivers during 2019's production. By actively changing production rules according to our ensemble's predictions and explanations, APERAM verified a decrease of over 50% in the occurrence of heating slivers.

5.3 Experiments and Results

After performing model sampling and explanation grouping, we obtain feature relationship graphs related to each perceived explanation. The next step consists of finding optimal traversing paths. In terms of action space, we consider the neighborhood of a model as those that are reachable by the aggregation of a co-related feature. We experiment using the A* algorithm employing as a heuristic the AUC of the current model. We hypothesize that there exists a set of optimal feature expansions that lead to the performant model, and exploiting models of solid performance is a reasonable approach. To further investigate this possibility, we also modify the algorithm to attempt to follow a backbone-like structure. Instead of considering all of a model's features to

Figure 5.5: Heating sliver: effect of introducing noise to ensemble constituents' input features.

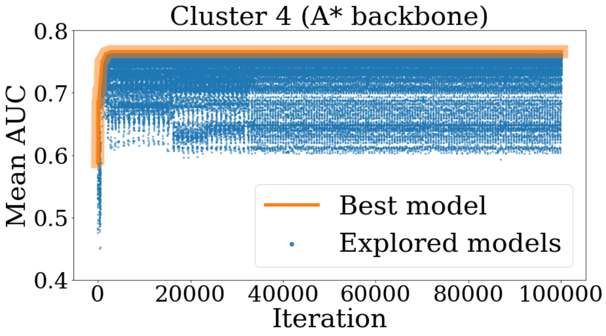


compute its neighborhood, we only access co-relation to the last added feature. Once the search is no longer able to follow a backbone, the algorithm traces back to the root exploring new backbone chains. We also experiment with Monte Carlo simulations as a means to introduce a non-stochastic behavior to the search exploring otherwise neglected areas of the search space. Given that Monte Carlo Tree Seach (MCTS) often copes with opening moves, we also judge the impact of briefly running A* before early simulations.

5.3.1 Optimal Expansion Paths

These strategies can only be accomplished if feature relationships translate to connected graphs. During our analysis, we verify that not only graphs are connected, but also that nearly all share a similar trait: the existence of a small completely connected component that is adjacent to all the remaining vertexes. We call this structure keynodes of the respective feature relationship graphs. Further, the keynode of a given graph is not present in any of the remaining relationship graphs. The only exception to this rule is the sparse cluster 12 illustrated in Figure 5.4b, in which we encounter the presence of

Figure 5.6: Typical search patterns found while running the A* algorithm. N = 100 000 models.



most keynodes and we do not observe the pattern of a strongly connected component with edges to all remaining vertexes. This phenomenon implies that any model that attempts to mix multiple keynodes also mixes up multiple causes, providing a split in credit and a distinct SHAP pattern. We believe that the keynode importance is so distinct that this sparse cluster contains all models that 'do not fit' any of the remaining patterns. When filtering these models, we obtain concise clusters with little to no overlap as evidenced in Figure 5.4c

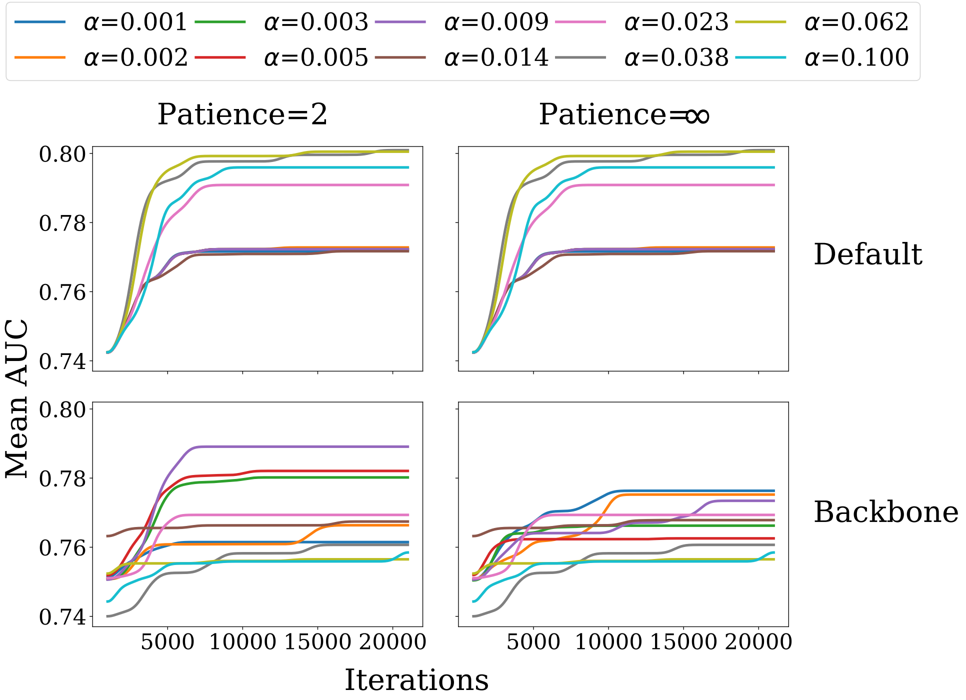
When observing the evolution of the best models found by the A* algorithm, we verify premature convergence. The algorithm quickly finds a configuration of high-performance features and tends to explore only minor modifications. In particular, it focuses on the addition of null-impact features that maintain AUC. This behavior generates the repetitive patterns of AUC noticed in the graphs of Figure 5.6. We propose two methods to remedy this problem. The first involves altering the standard A* algorithm to promote a more extensive exploration stage before reaching convergence. The second involves reducing the solution space by considering expansion using whole groups of related features rather than unitary ones.

5.3.2 Halting Null Expansions

Consider a space of solutions with features A-Z, and the optimal model is ABYZ. Among the possible feature pool, there should exist a group that does not affect the performance of models positively or negatively. Perhaps because they are not relevant or because they partially explain a phenomenon that is already covered by another feature. For this example, let's consider that these encompass features D-G. There may be models such as ABC that are superior to ABY, ABZ, ABCY, and ABCZ. Thus, throughout the search, the algorithm will prioritize the ABC model and, consequently, find ABC {D-G} models with similar performance to ABC. In the next iterations, the algorithm will focus on ABCD, ABCE . . . ABCG, ABCDE, ABCDF, and so on until all possible combinations involving the D-G features are found. Due to the high degree in the action graph of the heating sliver problem, the search algorithms are prone to fall into this type of situation and are unable to explore promising solution paths in a feasible time. This mainly happens because the search algorithms were not able to explore enough of the model space before attempting the exploitation of the best models.

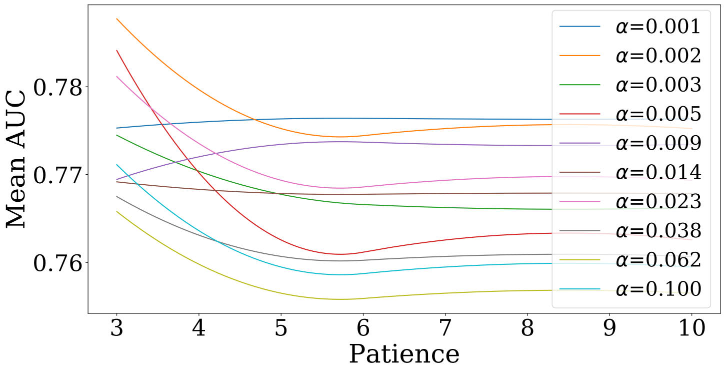
A simple method to attempt to halt null expansions is introducing an improved patience threshold. In the case of the default A*, if a model is not able to be improved upon after a certain number of iterations, then we should not attempt further expansions. For backbone A*, this patience threshold tells the algorithm when to stop pursuing a specific backbone chain and rather focus on other possible paths. A second approach to further improve exploration is modifying the fitness function to introduce a penalty to large models. This promotes the expansion of models with few features and enables the search algorithms to find a more diverse set of base models before beginning exploitation. Figure 5.7 illustrates this experiment. The patience threshold and the size penalty work toward the same goal of promoting exploration. However, there is an inherent tradeoff in their concurrent deployment, as both tend to hinder the expansion of large models. To find appropriate values for these approaches, we analyze how they interact with one another as pictured in Figure 5.8.

Figure 5.7: Effects of different α size penalties. N = 200 000 models.



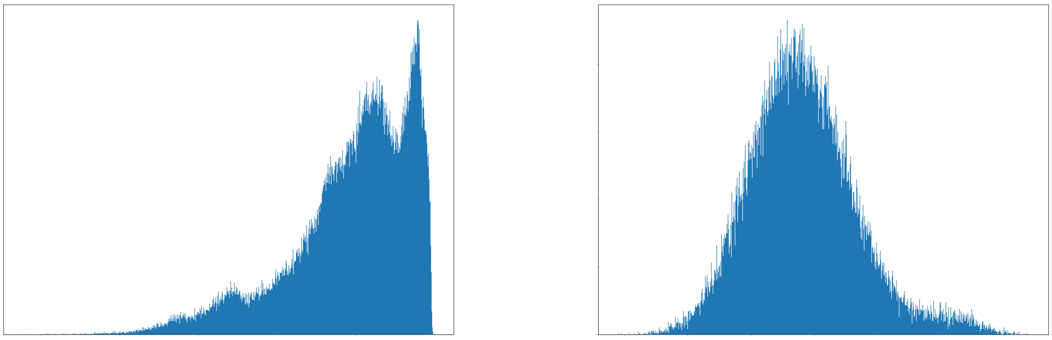
Search in the action graph is a challenging task due to its structure. However,

Figure 5.8: Effects of different α relative to patience threshold. N = 200 000 models.



the feature relationship graph provides useful insights regarding promising paths. Since edges represent features that co-occur in performant models, we should be able to encounter highly connected components that most likely contain features that should be evaluated together. We employ the NBNE algorithm from Pimentel et al. [2018] to compute neighborhood-based node embeddings and collapse nodes that have a high cosine similarity between vectors. At first, we observe a bimodal distribution with a prominent set of highly similar nodes but after collapsing, we obtain a bell-shaped curve with even distributions, as pictured in Figure 5.9.

Figure 5.9: Cosine similarity distribution between nodes.



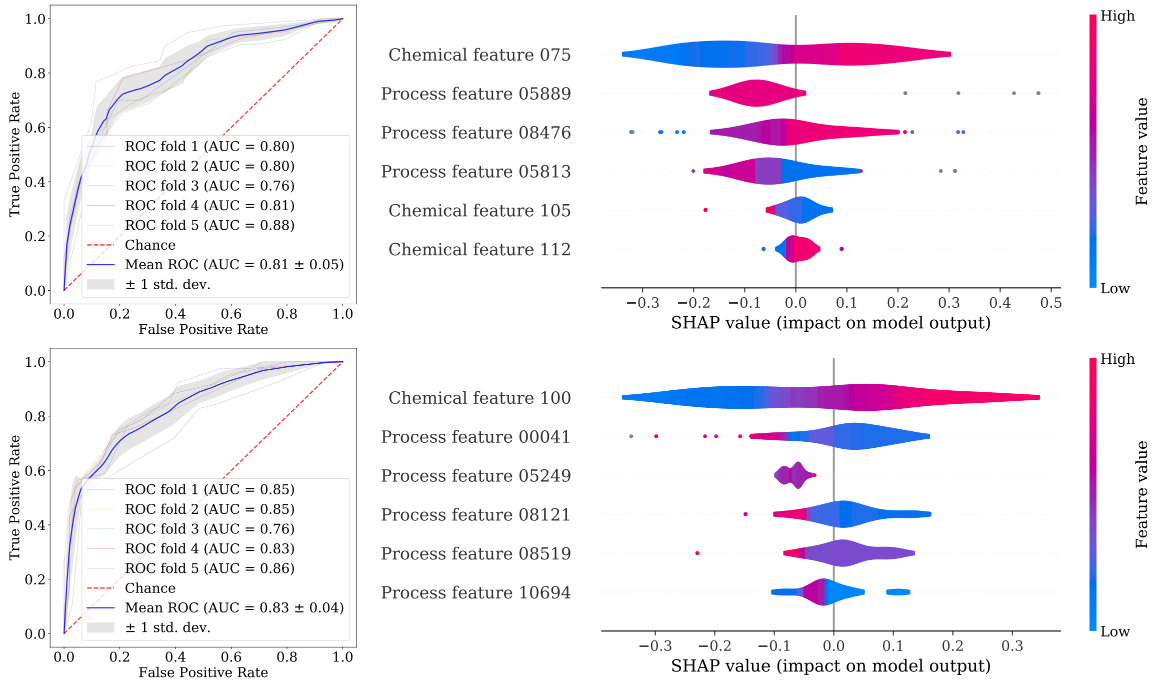
(a) Before collapsing.

(b) After collapsing.

5.4 Discussion

Model interpretability is one of the main requirements for the design of novel steel plates. An interpretable model needs to shed some light on the rationale behind the prediction. Figure 5.10 (Left) shows ROC curves for some performant models. Figure 5.10 (Right) shows SHAP summary plots. These plots show the SHAP values of each feature

Figure 5.10: Representative models for some clusters. Left – ROC curve showing the performance of the model. Right – The corresponding SHAP summary plot shows an overview of which features are most important for the model. N = 499 steel plates.



in all steel plates. In each plot, features are sorted by the sum of SHAP value magnitudes, with colors representing feature value. Curiously, some elements seem to be very relevant, as they occur very often in models within different clusters. In general, however, summary plots show that the features in models within different clusters are rather distinct. Overall, the predominant behavior is that specific groups of features are associated with distinct solution spaces and no clear relationship between feature sets and performance can be seen.

When accessing AUC performance among clusters, we can see that performant models are scattered through the model space, indicating that there are models with similar performances being assigned to different clusters. That is, models associated with different explanatory factors and different preferences can obtain similar performance numbers. There is no pattern in the distribution of clusters by AUC. Indeed, Pearson's, Spearman's and Kendal's correlation between the clusters and AUC are inferior to 0.1. When searching the model space, we were able to considerably improve upon the best models previously found by the model sampling approach. Further, our proposed pipeline presents superior results to the alternatives.

We explore the usage of various methods to search for the best models within each cluster. We found that any technique, even a greedy search, is superior to an all-in-one model expecting it to learn which features are the most relevant. This phenomenon happens in support of our original hypothesis and can be associated with data dimensionality. In our experiments, we also found that attempting to force a backbone-like structure does not appear to be effective. However, due to the high dimensionality and the degree of the feature relationship graphs, search algorithms appear to regularly get stuck in an exploitation step by assessing only small modifications of the same performant model, leading to visible repetitive patterns through iterations. The introduction of hyper-parameters that penalize this type of behavior proved to be useful, improving the exploratory stage of the algorithms.

We were able to verify that performing MCTS can be beneficial in some situations. We believe that the stochastic factor present in Monte Carlo simulations might help the algorithm explore different model spaces. To keep results comparable, we only allowed MCTS to perform as many simulations as the number of models explored by A*. It might be the case that we should allow the algorithm to run longer and achieve convergence, as these extra simulations should lead to a more reliable approximation. Nevertheless, these results provide an argument that there might be sequences of sub-optimal expansions that, when combined, lead to an overall better model.

Once we find the optimal models within each cluster, we can combine them into an ensemble to give a final prediction. We employed a simple voting ensemble in which each of the constituents shares the same importance. Figure 5.11 illustrates the results of an ensemble consisting of both the best models found (A* Ensemble) and one where we simply use the clusteroid model within each cluster. We evaluate our method against other state-of-the-art tree-based ensemble techniques as well as classic algorithms from the literature. Plotting the results of such algorithms against our approach would prove to be an unfair comparison as we already discussed that all-in-one approaches can have below-par performance. Therefore, we also consider selecting only the features with the highest shap values as well as applying the Boruta [Kursa and Rudnicki, 2010] feature selection strategy.

Figure 5.11: 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. N = 499 steel plates.

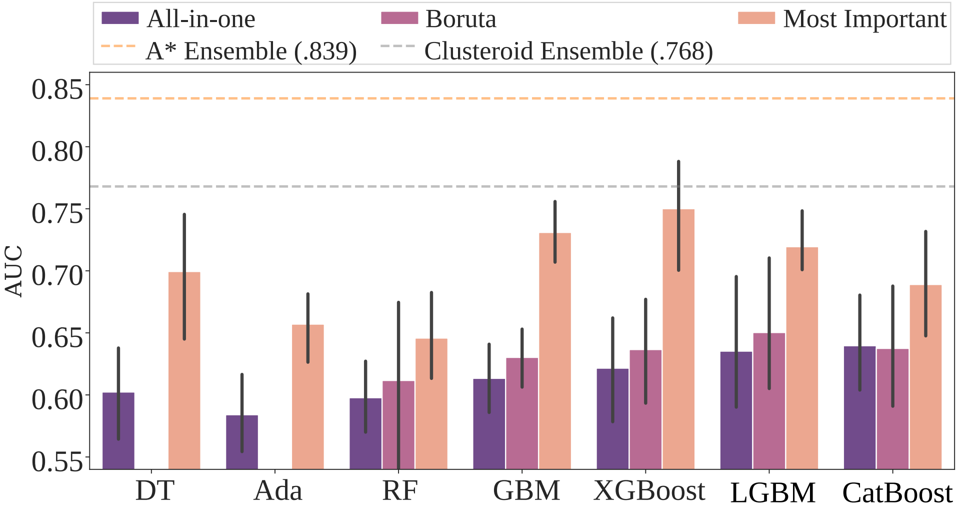


Table 5.1: Experimental results. We can observe that the default A* outperforms the remaining methods in most scenarios, thus justifying the A* (explorative) approach which refers to the optimizations discussed. N = 499 steel plates.

	Cluster															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	*Avg.*
Clusteroid	.60	.60	.61	.60	.60	.62	.60	.61	.60	.61	.60	.60	.62	.65	.61	*.61*
All-in-one	.64	.62	.65	.62	.64	.64	.62	.64	.61	.64	.62	.58	.62	.67	.65	*.63*
Greedy	.70	.70	.71	.70	.72	.71	.74	.71	.70	.73	.72	.70	.71	.70	.70	*.71*
A* (default)	.78	**.76**	.76	.76	**.83**	**.77**	**.80**	**.78**	.75	**.80**	**.79**	-	**.76**	**.79**	**.80**	***.78***
A* (backbone)	.70	**.76**	.77	.76	.81	.76	.78	.74	.76	.75	.77	-	.75	.77	.77	*.76*
MCTS (default)	.75	.74	**.77**	.75	.78	.75	.75	.76	.76	.76	.74	-	.73	.75	.73	*.75*
MCTS (backbone)	**.79**	.74	.75	.75	.78	.76	.76	.75	.76	.77	.75	-	.73	.77	.75	*.76*
A* + MCTS (default)	.78	**.76**	.77	.77	.79	.76	.77	.75	**.77**	.77	.77	-	.76	**.79**	.75	*.77*
A* + MCTS (backbone)	.77	.76	.76	**.77**	.78	.76	.76	.74	.76	.77	.76	-	.74	.77	.75	*.76*
A* (explorative)	**.80**	**.79**	**.80**	**.79**	.81	**.82**	**.81**	**.80**	**.79**	.78	**.80**	-	**.77**	**.80**	**.81**	***.80***

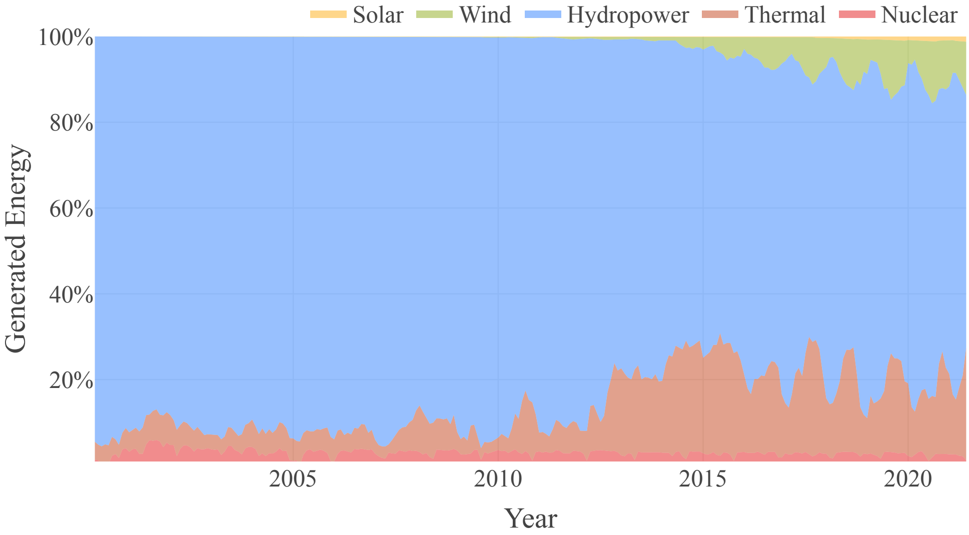
Chapter 6

Case Study: Identifying the drivers of energetic consumption

Energy systems are increasingly coupled with economic, social, and climate systems. For example, the deployment of renewable energy sources is critical for mitigating future climate change [IEA, 2021], and hydropower is the leading renewable energy resource in the world [Moran et al., 2018]. However, both energy and climate experts are raising concerns regarding hydropower's reliability under future climate change uncertainty and human development pressures. Expected growth in electricity consumption due to population and economic growth, the severity and frequency of extreme weather events, long terms effects of climate change on drought conditions, and expanding agricultural production all present immediate and urgent challenges to energy systems, particularly so for those dependent on renewables [Moazami et al., 2019] A crucial component of energy planning, therefore, requires an understanding of how these challenges are impacting electricity consumption and supply.

Such challenges are typified by Brazil, a country whose electricity system is heavily dependent on hydropower and is undergoing substantial economic and social transitions with populations living in geographically and climatologically diverse regions, all contributing to changing patterns in consumption. As of November 2021, Brazil's total generation capacity is 180.6GW, the third-largest electricity sector in the Americas and only second to China in total hydroelectricity capacity. Currently, 56.98% of Brazil's electricity sector is compromised of hydropower and 25.42% of thermal [ANEEL, 2021] with expected growth in the grid capacity of 2.7% yearly [EPE et al., 2021]. There are also indications that Brazil is on the brink of a multi-year energy crisis, spurred by low reservoir levels and drought, exacerbated by increases in electricity consumption [Ferraz, 2021]. Historically, Brazil relied almost completely on clean energy, but recently (see in Figure 6.1) there has been an increased reliance on thermal energy. In 2010 over 90% of the energetic demand was satisfied by hydro. In less than 10 years this changed to $50\% - 60\%$ of the generation coming from hydro-power, with thermal constituting about 20% of the total generated energy. The adoption of wind-based alternatives was able to somewhat halt the increase in thermal reliance, however, the main concern still stands:

Figure 6.1: Energy generation profile of Brazil. Although hydro-power constitutes over 50% of the generated energy, Brazil is experiencing a growth in the thermal share driven mostly by low reservoir levels, scarcity of rain, and fast load growth.



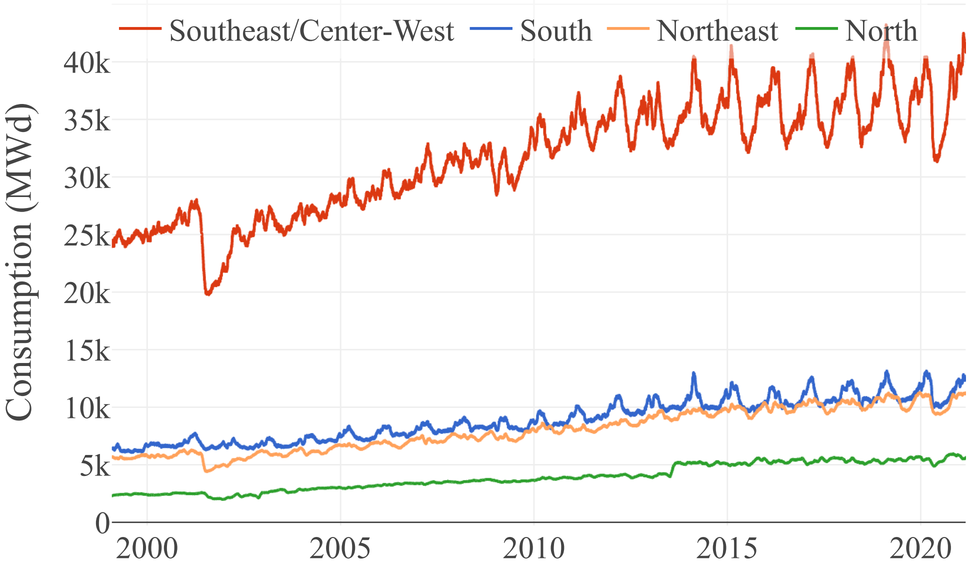
hydro-power might not be enough to supply the entirety of Brazil in the future and hydro-related crises are sure to occur more frequently. For these reasons, developing a deeper understanding of the Brazilian system is both timely and of critical societal importance [Hunt et al., 2022].

Climate and weather have been known to bear an influence on energy demand, one of the most meaningful variables being temperature [Davies, 1959, Hor et al., 2005]. A plethora of temperature-related metrics exist and many accurately approximate energetic consumption [Huang et al., 1986]. Although degree days remain one of the most common temperature-related metrics for load forecast, recent work has shown the significant impact of other weather variables. For instance, both Maia-Silva et al. [2020] and Woods et al. [2022] highlight the relationship between air humidity and the search for cooling during hotter days, and its direct consequence on summer electricity demand. Thus, many possible weather predictors that explain drivers of energetic consumption exist, suggesting the Rashomon effect and presenting a suitable scenario for our approach. Given the tight coupling of weather to the Brazilian electrical system at a national and regional level, the impact of disruptive events, and the need for future planning under climate change, we narrow our study's focus to one aspect of the energy system: weather determinants of consumption, as presented in our work [Zuin et al., 2022b].

6.1 Weather and Data

We applied two primary datasets in our study, Brazilian's National Energy System Operator (ONS) historical reports [ONS, 2018] and the ERA5 reanalysis [Hersbach et al., 2020]. Energy data was sourced directly from the Brazilian ONS website, containing daily measures dating back to 1999. The available information includes load, maximum consumption, mean and total daily megawatts (MWd), and hourly megawatts (MWh). The Brazilian electric grid is split into four subsystems associated with its main regions. Although there is some exchange of energy between adjacent subsystems, they are mostly independent in supplying local consumption. Figure 6.2 illustrates the load profiles for each of Brazil's macro-regions, all characterized by an increase in load over the last 20 years (1999-2021). We observe a rapid load growth, with an increase of nearly 60% in consumption in these 20 years. We also observe a close relationship between urban development and load, with Brazil's technological hubs having a higher energy consumption than less developed areas. Finally, there is a sharp drop in consumption near the year 2002 related to a curtailment intervention initiated by the Brazilian government.

Figure 6.2: Consumption for each subsystem in Brazil from 1999 to 2021 (30-day moving averages).



The ERA5 dataset consists of global hourly estimates from 1950-2021 for atmospheric variables, with a spatial resolution of 0.25 degrees (approximately a 30 x 30 km grid cell). To associate this weather data to Brazilian subsystems we extracted the mean values from grid cells using coordinates from each city with at least 100 000 inhabitants, weighting the values by the population so that larger cities will have a larger impact on mean values compared to smaller ones. Figure 6.3 depicts all considered cities and towns, encompassing over 70% of the total Brazilian population. Our final weather-related feature subset consists of daily temperature minimums, means and maximums, humidity, wind speed, and precipitation, heating degree days (HDD), cooling degree days (CDD), heat index, wind chill index, apparent temperature, HDD-derived from wind chill index and CDD derived from heat index.

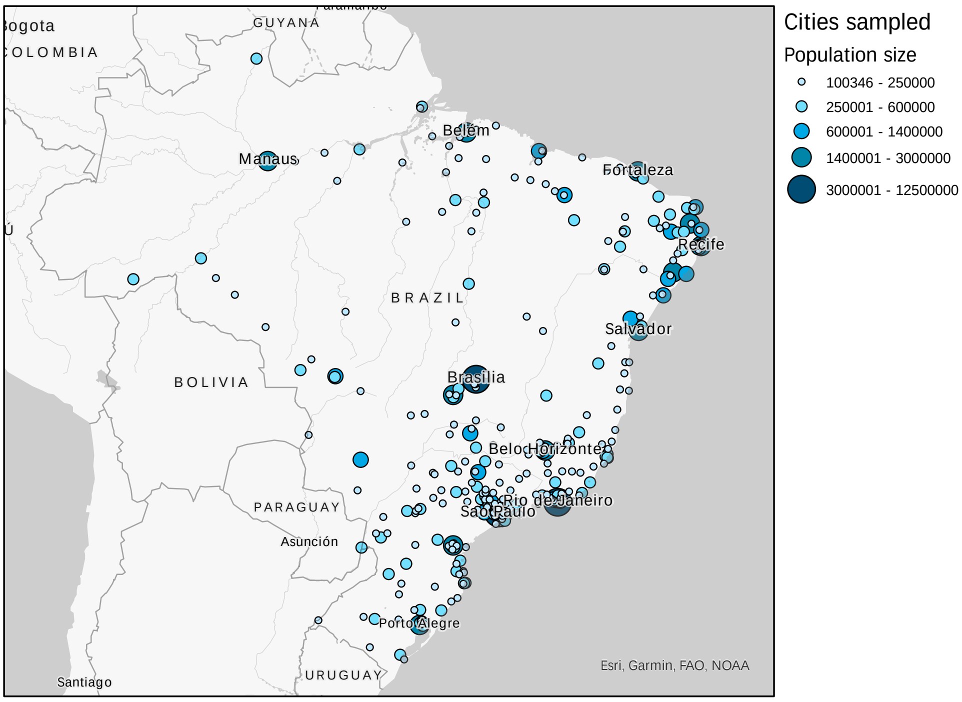
Degree days are a measure of heating or cooling, which has extensively been used for estimating energy consumption required for a household to reach a comfortable temperature and can be computed as an integral of a function of time over temperature. However, since temperature measurements are not continuous, but rather taken at discrete intervals, we can approximate heating degree days and cooling degree days for a single day as:

$$CDD = \frac{\sum_{i=1}^{T}(\theta_i - \theta_b)_{((\theta_i - \theta_b) > 0)}}{T} \tag{6.1}$$

$$HDD = \frac{\sum_{i=1}^{T}(\theta_b - \theta_i)_{((\theta_i - \theta_b) < 0)}}{T} \tag{6.2}$$

in which θ_i represents the temperature at instant i, θ_b is some constant representing the base comfort temperature and T is the number of equally spaced temperature measurement intervals. Thus, HDD is related to the amount of time that temperature remained below some established threshold and the size of this difference, implying that household heating was needed to reach the comfort temperature. Analogously, CDD is related to temperatures rising above this same threshold, leading to the need for cooling. A further

Figure 6.3: Brazilian cities with at least 100 000 inhabitants.



Esri, Garmin, FAO, NOAA

approximation employed in the literature considers only the difference between the base temperature and the mean daily temperature [Yang et al., 1995] or the maximum and minimum daily temperature [Ring et al., 1983]. We consider the base comfort temperature in Brazil as 65^oF.

However, as noted bySteadman [1984], the perceived temperature often differs from ambient dry bulb temperature. For example, on high dry bulb temperature days when the humidity is high, perception of heat is often greater than on days with an equivalent dry bulb temperature and lower humidity. The effect of wind has a similar perceptive impact of making cold days feel colder. Thus, Steadman defines the apparent temperature as the temperature equivalent perceived by humans caused by the combined effects of air temperature, relative humidity, and wind speed. The NOAA National Digital Forecast Database (NDFD) states that when the temperature falls below 50^oF, wind chill is a suitable measure of apparent temperature, and when the temperature rises above 80^oF, heat index should be employed. Between these two thresholds, humans experience the combined effect of both wind speed and humidity, and the ambient dry bulb air temperature is a reasonable measurement of apparent temperature. Heat index (HI) and wind chill index (WCI) can be computed as:

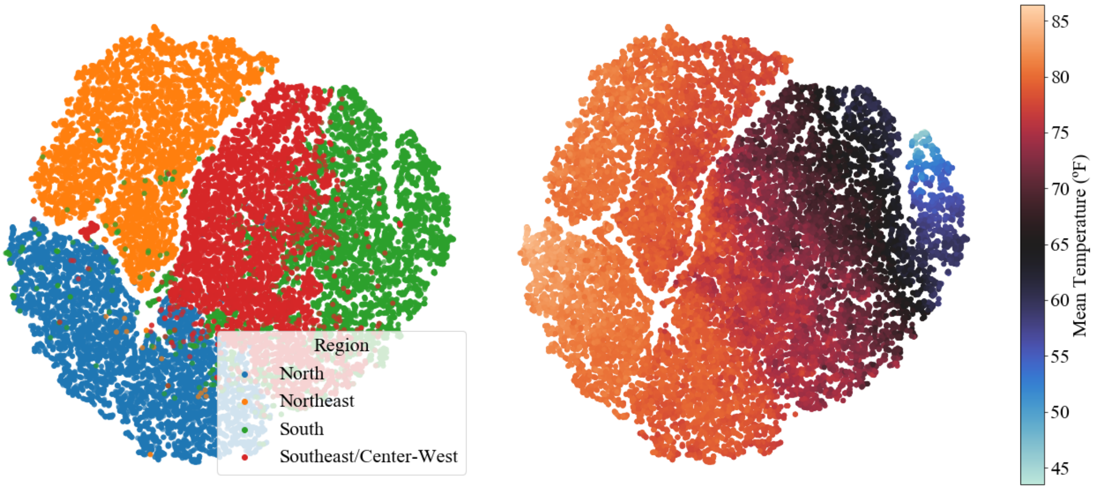
$$HI = c_1 + c_2\theta + c_3\phi + c_4\theta\phi + c_5\theta^2 + c_6\phi^2 + c_7\theta^2\phi + c_8\theta\phi^2 \tag{6.3}$$

$$WCI = 35.74 + 0.6215\theta - 35.75\nu^{0.16} + 0.4275\theta\nu^{-0.16} \tag{6.4}$$

in which θ is the dry bulb air temperature, ϕ is the relative humidity, ν is the air speed and c_i are constants used to approximate Steadman (1979) original heat index table [Steadman, 1979, Stull et al., 2000], valued as $c_1 = 0.363445, c_2 = 0.988622, c_3 = 4.777114, c_4 = -0.114037, c5 = -8.50208e^{-4}, c6 = -2.071619e^{-2}, c7 = 6.87678e^{-4}, c8 = 2.74954e^{-4}$. Since both HI and WCI are measured in ^oF, we can replace the temperature from Equations 6.1 and 6.2 to obtain both CDD and HDD derived from either heat index, wind chill or apparent temperature.

Brazil has relatively high climate diversity. This is mostly due to the size of the territory, the extent of its coastal regions, the variation in altitude, and the presence of different air masses that influence the temperature and humidity of each region. Most of Brazil is located between the Tropics of Cancer and Capricorn and for this reason it is called intertropical, but other types of climate are also present. According to Strahler and Strahler [2007], Brazil possesses six climate zones: equatorial, tropical, semi-arid, coastal, subtropical, and tropical high-altitude. This climate diversity, coupled with demographic heterogeneity, would make building a unified model for the entirety of Brazil a challenging task. However, by taking advantage of the political macro-regions and examining each of them individually, an analysis could be made more manageable and interpretable. This

Figure 6.4: TSNE representation of Brazilian weather. Each point represents a day in one of Brazilian's sub-regions. There seems to be a direct relationship between temperature and geographical location, that produces well-dived partitions on the plot on the left without explicitly performing clustering.



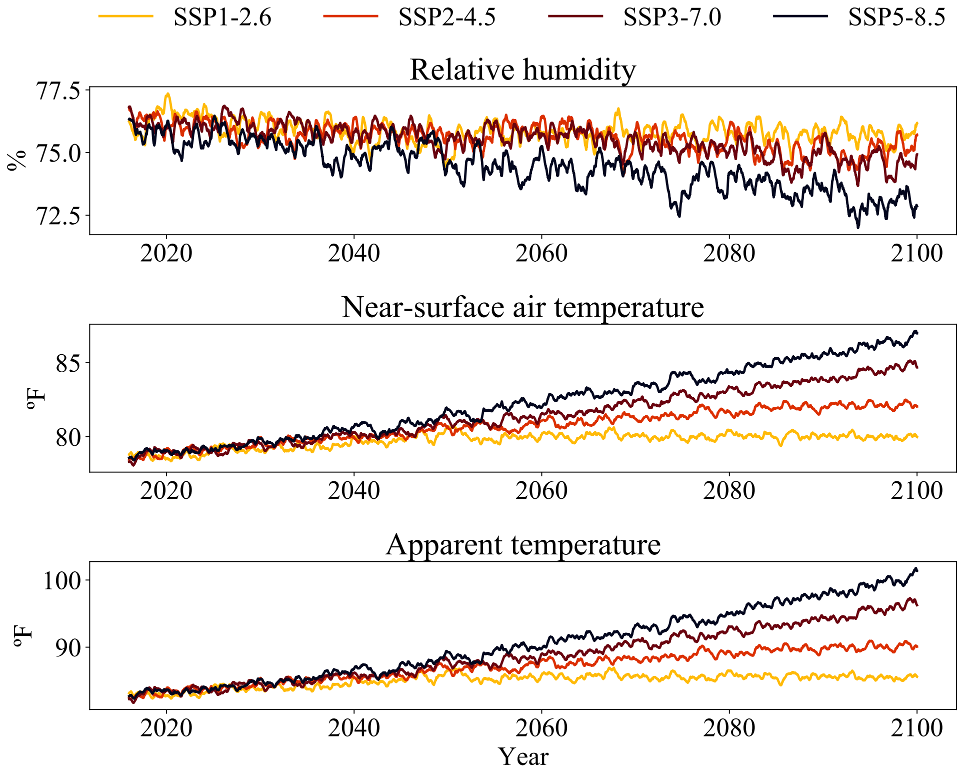
approach results in each region of interest encompassing at most two climate domains that are similar. However, our weighting approach relies on the weather of cities in the same region being similar, so climates from different regions do not offset. Figure 6.4 illustrates a two-dimensional reduction of the evaluated daily weather variables employing the TSNE algorithm, in which each point represents a day in our dataset that is colored to correspond to a region and its mean temperature.

We can observe that each region's weather seems to be well-defined, with a clear separation between most regions and small intersections. These clear boundaries validate our analysis of distinct regions and the proposed city-to-region aggregation approach for climate variables. Given that the main focus of this work is on the impact of temperature on energy consumption, we also wish to understand how temperature impacts this separation, and, after visualizing these measures, the relationship between geography and temperature is evident. Further, we can also draw some conclusions regarding the intersection between regions. The are only a handful of days where the weather in the South region is similar to the Southeast. From the temperature plot, we also observe that the colder Southeast days are similar to the hotter days in the South. An analogous pattern can be identified between the Southeast and the Northern region. The Northeast days appear to be isolated from the remainder of the plot. In our experiments, we verify that predicting consumption for this specific region is more difficult, and this might be one of the reasons for this observed phenomenon.

We also include data from various Coupled Model Intercomparison Project Phase 6 (CMIP6) models to forecast demand under future climate change, ranging from SSP1-2.6 to SSP5-8.5 [Eyring et al., 2016]. SSP-RCP (shared socioeconomic pathway-representative concentration path-way) narratives were designed to capture possible future scenarios for energy demand due to temperature and the shared socioeconomic pathways represent varying assumptions regarding future global development. Each of the five SSPs corresponds to a projection of greenhouse gas emissions, namely: SSP5 (fossil-fueled development), SSP4 (inequality), SSP3 (regional rivalry), SSP2 (middle-of-the-road development), and SSP1 (sustainable development). By design, SSPs were devised to work in conjunction with the representative concentration pathways. The various pathways relate to the levels of radiative forcing by the year 2100 and range from 1.9 to $8.5W/m^2$, in which higher values depict more global warming. Although various combinations of SSP and RCP are possible, many do not lead to feasible narratives of interest. For instance, the SSP5, characterized by fossil-fuel development and high gas emissions, is not a realistic pairing with low-emission RCPs such as $1.9W/m^2$. Here, we employ four Tier 1 SSP-RCP narratives from the Scenario Model Intercomparison Project (ScenarioMIP) within Phase 6 of the Coupled Model Intercomparison Project (CMIP6), SSP1-2.5, SSP2-4.5, SSP3-7.0 and SSP5-8.5 [Abram et al., 2019, O'Neill et al., 2016].

Individual models can introduce distinct biases driven by their methodologies and the physical phenomena that each attempts to represent. As such, forecasting and evaluation from a single model might lead to results that contradict another model. Considering the means from a group of models therefore can allow for a better evaluation of the overall trend for the forecasted RCP-SSP future pathways [Cheng and Zhu, 2016, Wang et al., 2018]. Specifically, we consider an ensemble for each specific SSP-RCP pathway compromised by the models ACCESS-CM2 [Bi et al., 2020], AWI-CM-1-1-MR [Semmler et al., 2020], BCC-CSM2-MR [Wu et al., 2019], CESM2-WACCM [Liu et al., 2019], CMCC-CM2-SR5 [Cherchi et al., 2019], CMCC-ESM2 [Lovato et al., 2022], CanESM5 [Swart et al., 2019], EC-Earth3 [Döscher et al., 2021], FGOALS-g3 [Li et al., 2020], GFDL-ESM4 [Dunne et al., 2020], IITM-ESM [Krishnan et al., 2019], INM-CM4-8 [Volodin, 2021], INM-CM5-0 [Volodin, 2020], IPSL-CM6A-LR [Boucher et al., 2020], MIROC6 [Tatebe et al., 2019], MPI-ESM1-2-HR [Müller et al., 2018], MRI-ESM2-0 [Yukimoto et al., 2019], and NorESM2-MM [Tjiputra et al., 2020], sampling maximum and minimum near-surface air temperature (tasmax and tasmin), precipitation flux (pr), near-surface eastward and northward component of wind (uas and vas), and near-surface relative humidity (hurs). Across all models and pathways, we can expect an increase in temperature coupled with a decrease in humidity by the end of the century. Largely due to warming, we can expect an increase in perceived temperature even in a future Brazil climate with less humidity, as illustrated in Figure 6.5. No clear pattern was identified concerning wind speed and precipitation, with most models providing conflicting predictions regarding the latter.

Figure 6.5: CMIP6 projections for temperature and humidity, and the resultant increase in apparent temperature.

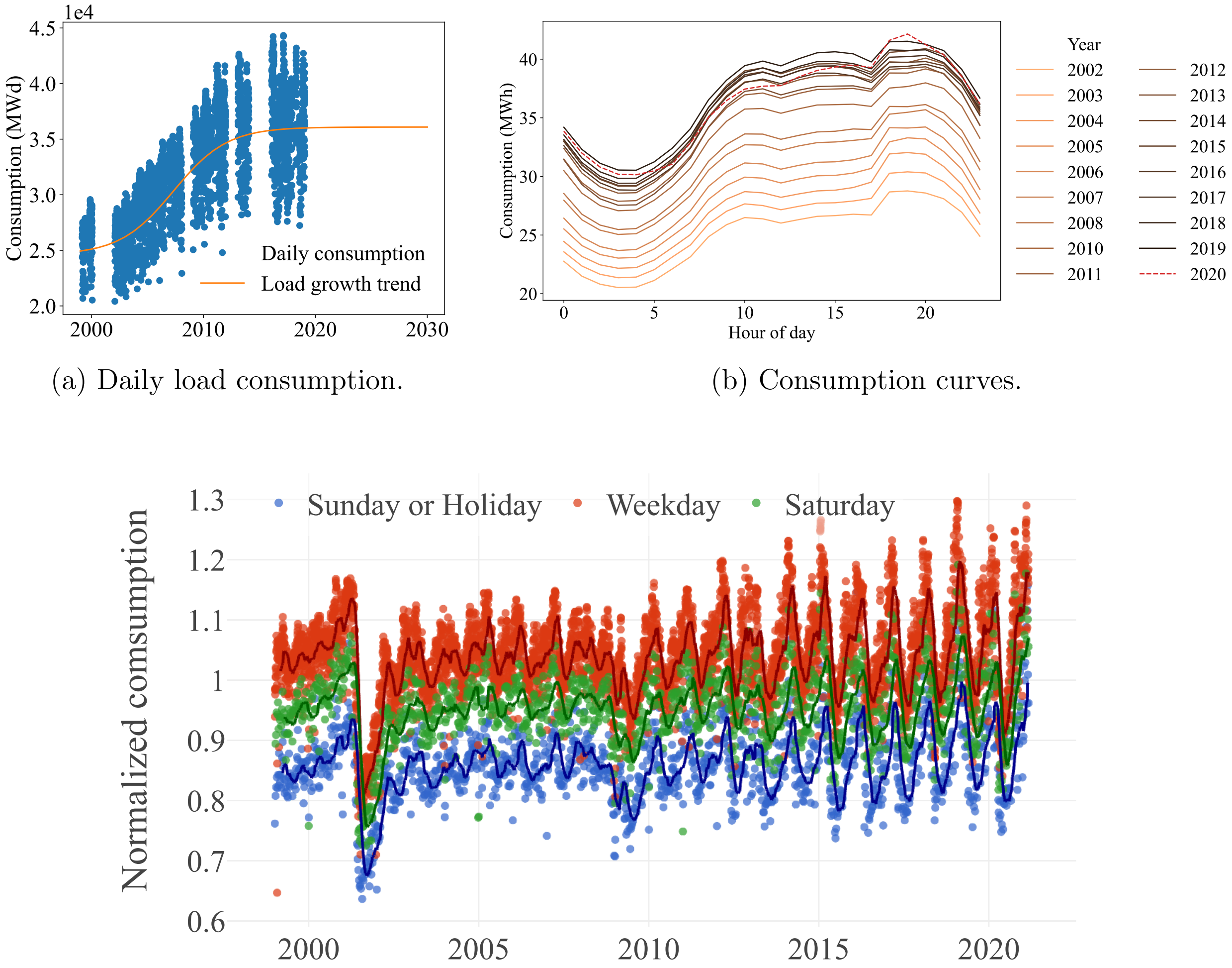


6.2 Building a Rashomon Ensemble

Our main objective in this case study is to predict consumption in the absence of any abnormal events, as this allows for a direct comparison of predicted and actual consumption to estimate the event's contribution to the consumption change. We formulate the problem as a regression. Given a set $w \in W$ weather descriptors and a set of $t \in T$ time descriptors, we apply a function $f(w; t; \sigma)$ parameterized by σ that maps a period to a consumption. To exclude disrupting factors, we search for optimal $W' \subset W$ and $T' \subset T$. Guided by existing literature [Giannakopoulos and Psiloglou, 2006], we propose the existence of three major groups of factors that drive electricity consumption:

- Load growth: increase in population, increase in consumer purchasing power, GDP, industrialization, etc.
- Historical events: pandemics, large-scale government policies, large celebrations, events, etc.
- Weather: heatwaves, winter storms, droughts, etc.

Figure 6.6: Consumption and load for the Southeast/Center-west subsystem.



(a) Daily load consumption.

(b) Consumption curves.

(c) Daily consumption normalized by load growth.

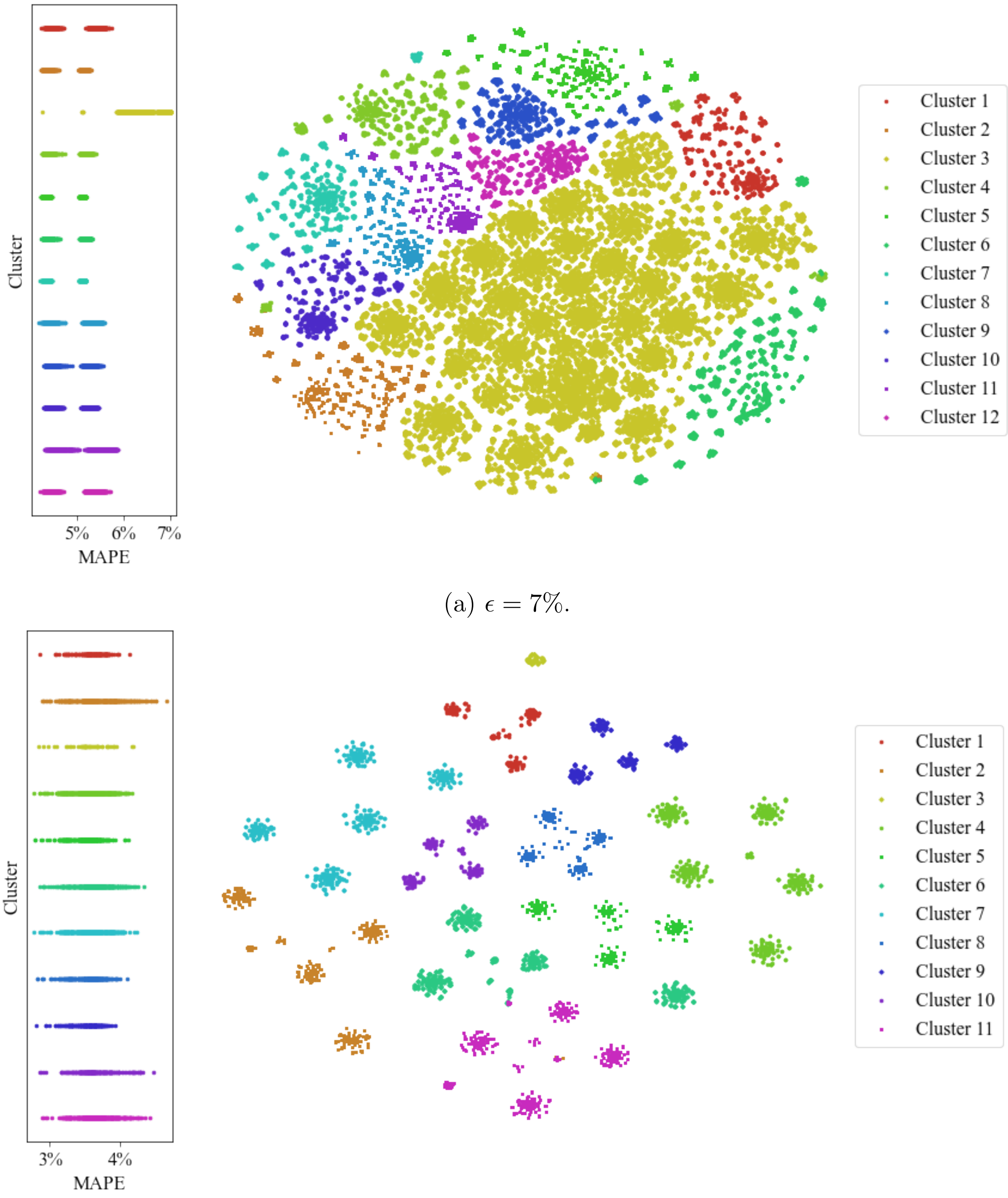
We observed that yearly energetic consumption follows a logistic growth trend as exemplified by Figure 6.6a and as typified by emerging countries [Tursun et al., 2016, Tuunanen et al., 2015]. When plotting the daily energy curves in Figure 6.6b, a similar pattern can be seen as the shape of the curves themselves remain relatively similar but there exists a step between one year's curve and the next, with the year 2020, highlighted in red, posing an abnormality. Normalizing daily consumption by the load growth function, which can be found from yearly load interpolation filtering out important atypical events, allowing us to build a counterfactual model focused on weather and temporal (e.g., day of the week, month, etc.) factors. We can then learn many models $f'(w; t; \sigma')$ employing different sets of features, and then build an ensemble that encompasses the many possible explanation biases per the Rashomon effect. Figure 6.6 illustrates the consumption for the Southeast/Center-West region after normalization removing the load growth trend.

Regarding the choice of a learning algorithm, we evaluated different possibilities as base constituents and found lightGBM to be performant. However, Tree-based algorithms are known to struggle with data pertaining outside the training domain, even if it follows the same distribution. While extrapolation beyond the data range can lead to biased results [Hahn, 1977], we use it as a way to gain scenario-based insights into future periods, such as forecasting consumption for the next century under climate change in Section 6.3.3. To address the challenge of out-of-distribution observations, Hooker [2004] proposes augmenting datasets to improve extrapolation using the Data-Augmented Regression for Extrapolation (DARE) algorithm, consisting of repeat sampling on the desired domain and the employment of a background model to provide weighted responses. To allow tree-based algorithms to cope with temperatures outside the training domain we augmented training data by 10% from 1872 Monte Carlo simulations uniformly sampling temperatures between 50°F and 120°F and employed a linear regression as the background model.

The first step to building our ensemble is sampling models to estimate the Rashomon space. However, in our approach, we often need to compare models with a different number of features. Although an standard literature metric, an unattractive property of the R^2 is its non-decreasing property. The addition of explanatory variables which bear no relevance to the target variable does not decrease the value of R^2, thus requiring some sort of penalization for models with too many features [Dufour, 2011]. Due to this fact, we elected the usage of the Mean Average Percentile Error (MAPE) to induce Rashomon sets. This approach led to the minimization of error, rather than the maximization of AU-ROC as in previous experiments. Figure 6.7 illustrates the found Rashomon space after sampling 100 000 models and also depicts the impact of different choices for the MAPE ϵ threshold to induce the Rashom set. For MAPE of 7%, we found a Rashomon ratio of .82 (82 114 of the original 100,000 models presented $MAPE \leq 0.07$) while decreasing this threshold to 5% reduced the ratio to .04 (4 064 of the original 100,000 models presented $MAPE \leq 0.05$). In Figure 6.7a, we observed a correlation between performance and cluster assignment, as nearly all underperforming models are assigned to the same sparse grouping. Including models from this cluster in the ensemble might not be productive as we are unlikely to find a good representative due to its sparsity. Properly tunning *epsilon* severely reduced the space, which allowed us to extract meaningful explanation groupings as seen in Figure 6.7b.

Following our proposed algorithm, we searched for optimal representatives under each explanation cluster. However, it is important to highlight that not all variables are relevant for forecasting. We represented the model space as a directed acyclic graph (DAG) and each node depicted a model built from a distinct feature subset. Let A and B be two nodes representing two distinct models. Vertex $A \rightarrow B$ exists if B can be reached by simple feature addition from A, thus constituting a transitive reduction of the complete model space. This approach presents two desirable properties: (i) any vertex is reachable from the $[\emptyset]$ model, and (ii) there exists a topological hierarchy, an ordering of all vertices into a sequence such that for every edge, the start vertex occurs earlier in the

Figure 6.7: Induced Rashomon spaces when setting ϵ threshold to 7% and 5% MAPE respectively. Overestimating the choice of ϵ leads to a larger Rashomon space bearing direct correlations between cluster assignment and explanatory factors.



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

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

sequence than the ending vertex of the edge for any feature set path. We proposed the existance of a set of optimal feature expansions that lead to performant models. These properties imply a partial ordering starting from the root node, enabling the search of the $N!$ combinatorial space. In our experiments, we applied the greedy beam search algorithm [Reddy et al., 1977], employing as a heuristic the R^2 of the model represented by a vertex and a beamwidth $\beta = 5$. Unlike when establishing the Rashomon set, we opted for choosing R^2 due to its informative properties in comparison to other regression metrics when evaluating models with a similar number of features [Chicco et al., 2021], as we sought the features that maximized the performance of each target model.

The beam search algorithm is an optimization of the best-first search parameterized by a beamwidth β, thus combining characteristics from breadth-first and depth-first searches. At each iteration, the best β nodes are considered expansion candidates, and all of their children are evaluated. If $\beta = 1$, then beam search behaves as best-first, in which only the best child in each iteration is expanded. If $\beta = \infty$, then beam search becomes a breadth-first search in which all possible expansion paths are considered. This allowed us to search the $N!$ combinatorial space of feature subsets to select the best-performing specialized models.

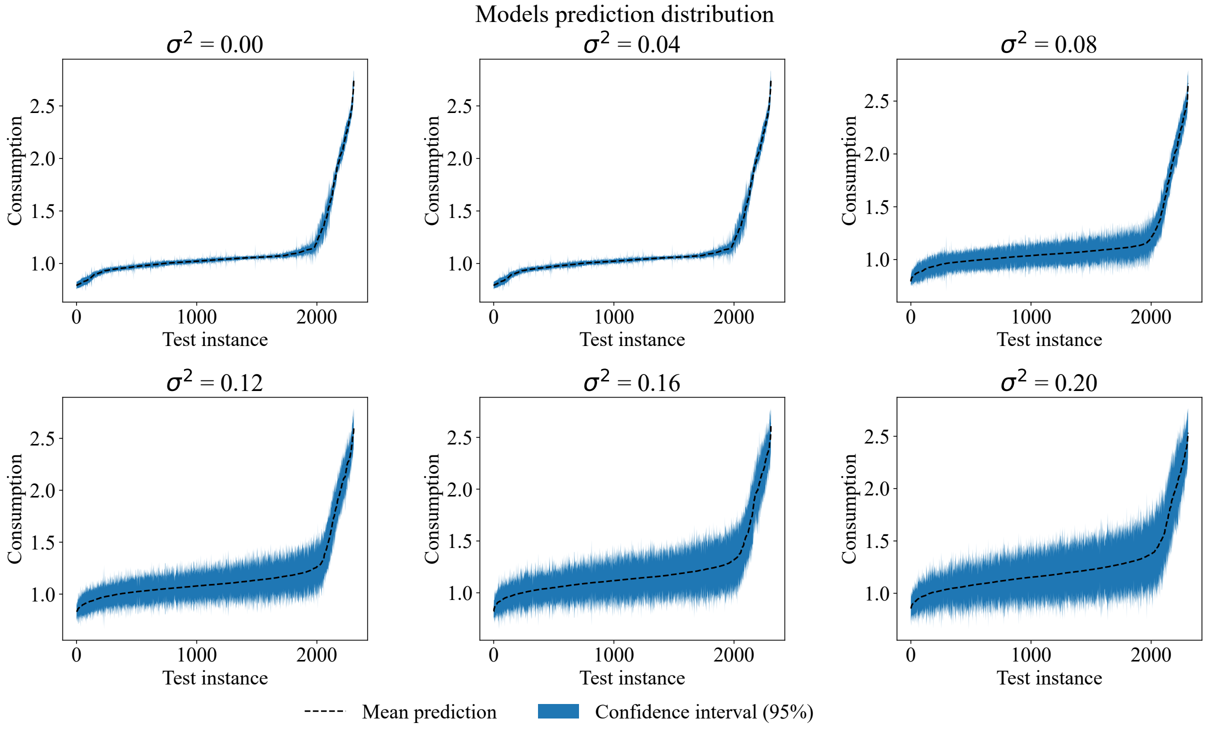
Once models are selected, their suitability as Rashomon constituents was verified. In Figure 6.8, we introduced increasing amounts of gaussian noise to the normalized features and observed the normalized consumption estimated by each individual constituent. As expected, in the absence of noise all models behaved in a statistically similar manner and the confidence interval was narrow. As we introduced noise, models quickly started disagreeing. Like in the previous experiments, we verified a direct relationship between noise and the confidence interval width, thus signifying that models become more divergent under drift and ensemble reliability decreases. In fact, even if the mean prediction of the ensemble stayed relatively the same, small amounts of noise already introduced drastic effects on predictions.

To apply our approach in production, our hypothesis is that measuring model disagreement allows for the estimation of ensemble reliability. However, the previous definition of agreement is ill-defined for regression as a visual inspection of confidence intervals is undesirable. When considering a classifier, we can state that two models agree if they predict the same class. It is unlikely that two regressors output the same value. We would argue that they 'agree' if the predictions are 'close', a subjective measure that depends on context. We opted for the coefficient of variance (C_V) between regressors as our measure of agreement, defined as:

$$C_V = \frac{\sigma}{\mu} \tag{6.5}$$

The standard deviation σ provides a measure of spread. Dividing σ by μ provides a dimensionless metric representing the extent of variability concerning the population

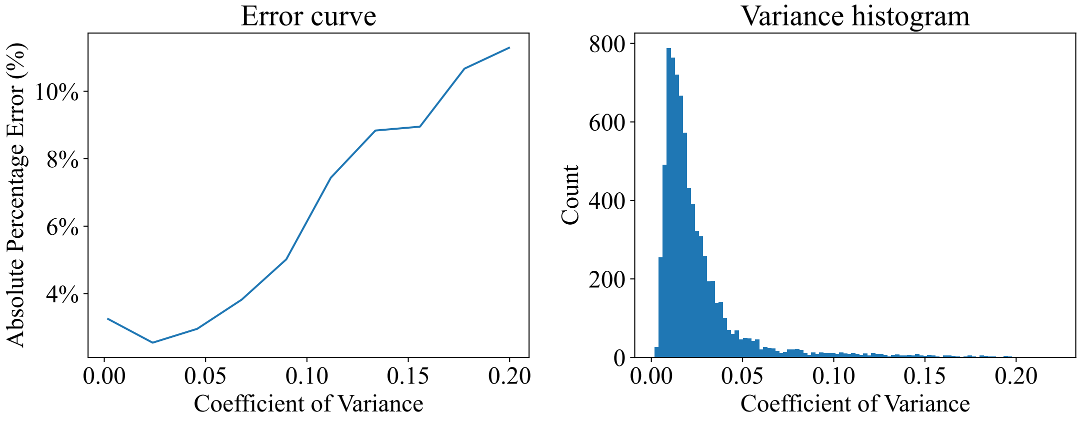
Figure 6.8: Energetic consumption: effect of introducing noise to ensemble constituents' input features



means. Therefore higher the C_V, the greater the dispersion and disagreement between constituents. Under a voting scheme, the ensemble prediction is μ, C_V states how far apart are the individual constituents to the final ensemble. Section 6.3.1 provides an example of how this metric can be used during production to estimate ensemble reliability.

Finally, we validated our ensemble by a comparison between performance and constituent predictions dispersion C_V, as illustrated in Figure 6.9. We inferred a direct relationship between these two metrics. We also considered the distribution of constituent dispersion and observe that most instances are located below $C_V = 0.05$. That is, the majority of instances presented less than 5% dispersion between constituents of the ensemble. In this scenario, we can expect a MAPE below 4%, which is desirable. As dispersion rose, error quickly increased. Thus, we can conclude that disagreement between constituents turns predictions unreliable following our hypothesis. This approach enabled us to extract relevant scenario insights by applying our model to different scenarios, described in the following Section.

Figure 6.9: Relantionship between constituent agreement and ensemble performance. The coefficient of variance between constituent predictions was used as an agreement metric.

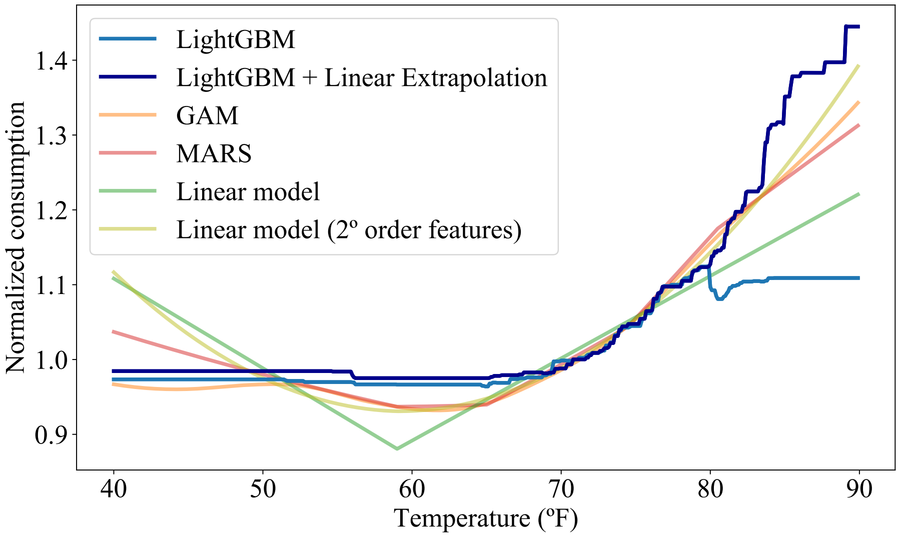


6.3 Experiments and Results

We assessed the statistical significance of our measurements through a pairwise t-test with p-value ≤ 0.05 and a one-year walk-forward validation. An exhaustive grid search was employed for hyperparameter tuning. We considered many algorithms for base constituents, but the performant model found was a lightGBM optimizing the L2 loss function. The quantile regression used to obtain the prediction intervals employs the same hyperparameters but optimizes the pinball loss function. The learning rate was set to $5e^{-}2$, with 64 bins and training 100 trees with 30 leaves and a maximum depth of 50.

Figure 6.10 presents a simulation in which we only altered temperature and observed the deviation in energy consumption. The lightGBM model trained on augmented data from a linear regression was able to extrapolate consumption past the 80°F threshold, at which point the standard lightGBM produced an unrealistic flat consumption curve. However, we can expect a trade-off between extrapolating and predictive power. Table 6.1 presents the performance in terms of R^2 and MAPE for all evaluated methods in each Brazilian subregion and on the whole dataset. Gradient boosting outperforms all other methods in the overall scenario. Thus, introducing a dependence on a linear model which presents a larger error reduces performance in comparison to the standalone lightGBM. Nevertheless, from the Table, we verify that this loss in predictive power is minimal and for some regions, there is no statistical difference between both lightGBM methods. Thus, DARE augmentation was able to retain the robustness of gradient-boosting machines and the extrapolative ability of the linear models.

Figure 6.10: Consumption simulation while varying temperature. Southeast/Center-west 2020-11-16, weekday.



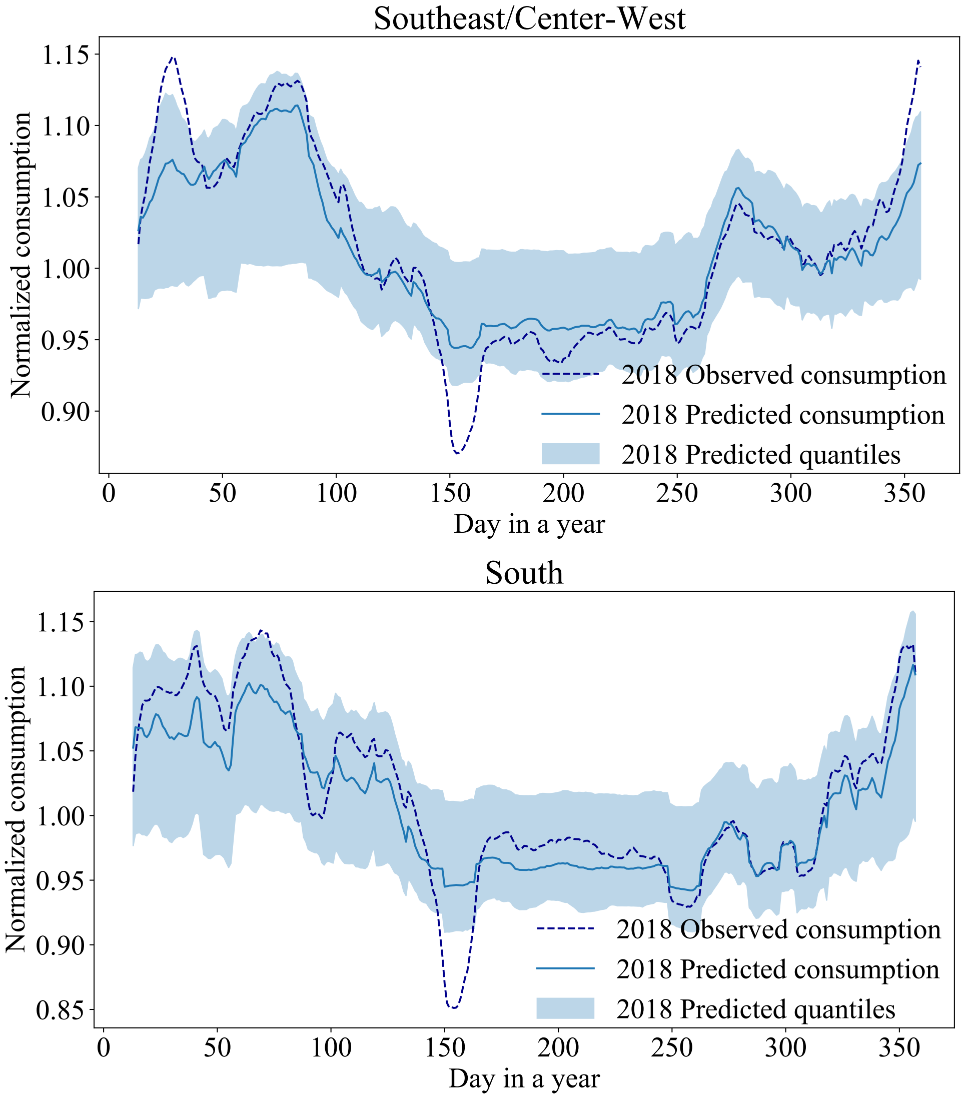
6.3.1 Atypical Historical Events Impact

To understand the impact of events on energy consumption, we compared the predicted consumption from our counterfactual model with the observed consumption. To get sensible ranges, we performed quantile regression through the pinball loss function to obtain the 5% and 95% quantile predictions. If the observed consumption lies outside this range, we assumed that this event generates an atypical consumption profile, reflecting a low (< 10%) probability of this occurring.

We show three relevant case study periods for analysis regarding RQ2. The first case study period is 2018 in which the Brazilian economy recovered after a long recession during a period with no severe heatwaves, droughts, or other extreme weather events of note, serving as a suitable year for baseline comparisons. Figure 6.11 illustrates the consumption for the most developed region in Brazil for these two scenarios. The only anomaly observed during 2018 was the truck drivers' strike that lasted from May 21st to May 30th, resulting in the disruption of entire supply chains across Brazil, empty gas stations, and people forced to work from home due to limited transportation options. The dramatic impacts of this strike on consumption are easily identifiable in Figure 6.11.

The second case study period is the year 2020, in particular during the early stages of the COVID-19 pandemic which resulted in substantial restrictions on mobility and a reduction in GDPs that were also experienced on a global scale. In Figure 6.12 we compared consumption with the Oxford Stringency Index, which measures the strictness of COVID-19-related policies that restrict population behavior [Hale et al., 2020]. During the early COVID pandemic, we observe a close relationship between the drop in consumption

Figure 6.11: Consumption for 2018 in the Southeast/Center-west and South subsystems.

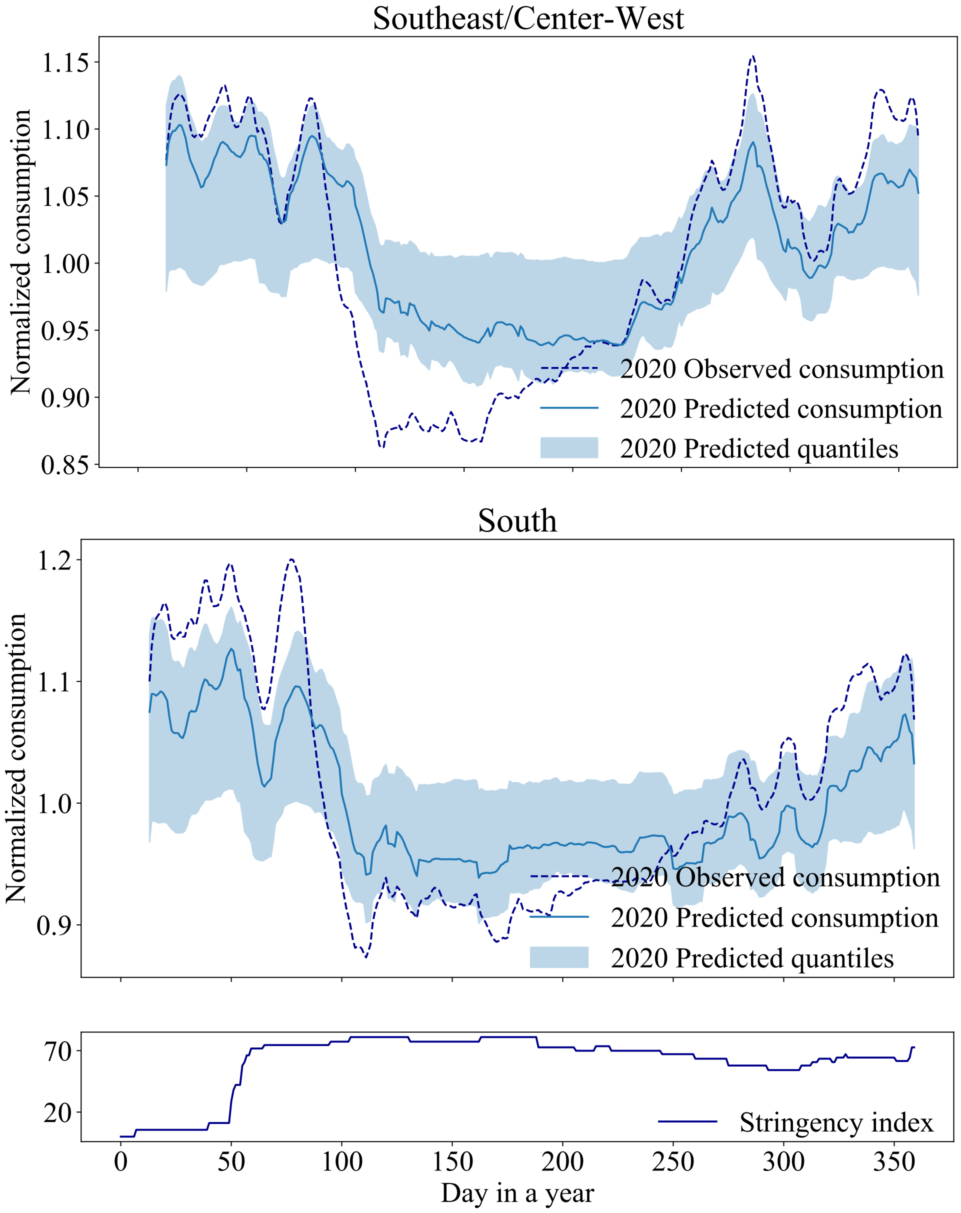


and the Stringency index for Brazil. The exception is the South region, which overall was less impacted by the pandemic compared to other regions in Brazil. One explanation for this lower observed impact is due to the region's economic dependency on tourism and lower levels of population concern about the COVID-19 virus in March. As such, by April most restrictions had already been lifted [GULLO, 2020].

The absence of variables concerning demographics and population behavior makes our model unable to predict consumption during the COVID-19 pandemic, which enables us to measure its impact in a counterfactual approach. Therefore, we expect all constituent errors to be equally high across 2020. Figure 6.13 presents ensemble and constituent performance when trained on data from 2014 to 2018 and employed for prediction in 2019 and 2020. As expected, errors for 2019 are similar to those seen in training, meaning similar weather variables distributions.

However, in 2020 we witness erratic behavior from the Rashomon ensemble. While in 2019 the coefficient of variance was 0.04, we observed a variance of 0.14 in May 2020,

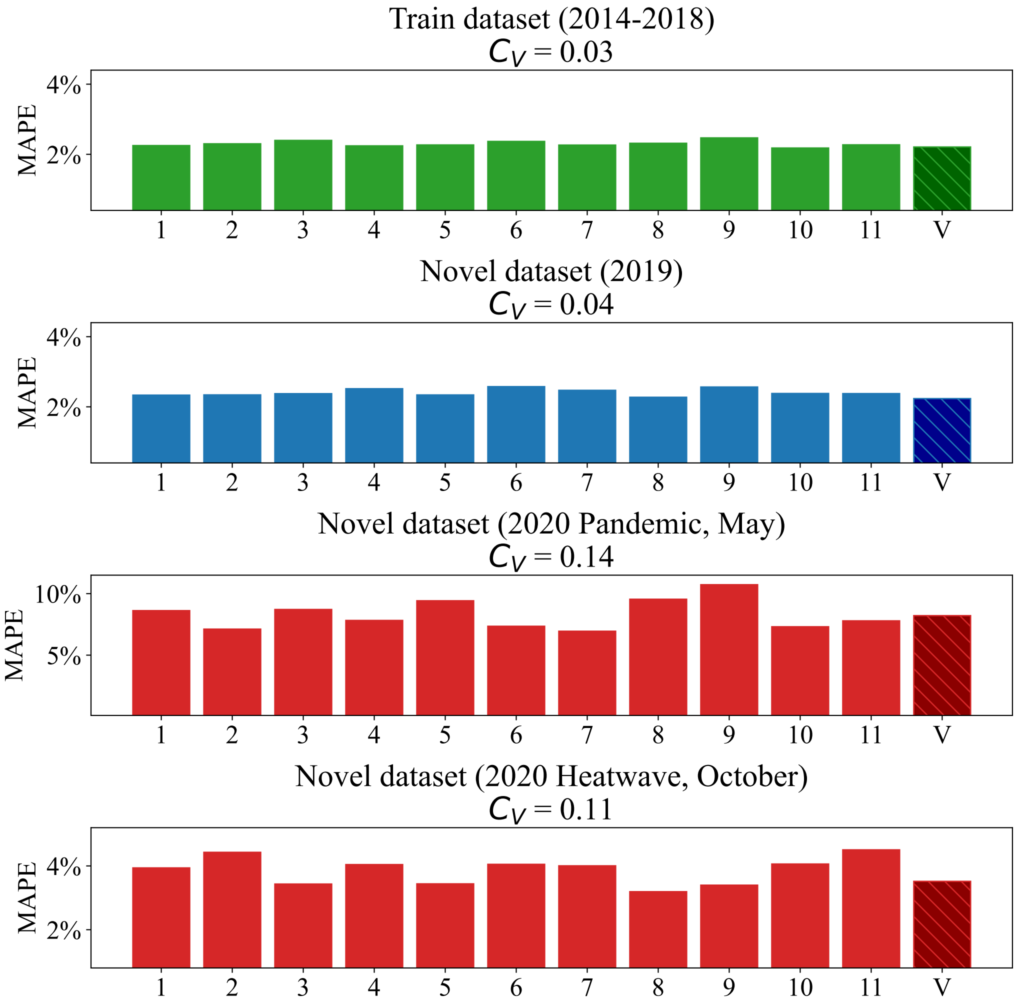
Figure 6.12: 2020 COVID-19 impact on Brazil. Stringency data from Ritchie et al. [2020].



implying that the weather for 2020 did not conform to earlier years. Since this pertains to the high point of restrictions in Brazil, we can conclude that the pandemic effect shadowed this change in weather. Brazil registered record-breaking heatwaves in October, surpassing temperatures from the past 100 years. Under such extremes diverging from 2014-2018 distributions, predictions became unreliable due to error rising alongside ensemble dispersion. If prediction errors in May were exclusively due to the pandemic, we would expect similar performance across constituents. However, their erratic behavior implies that the weather for the entirety of the year was atypical. Months before the heatwave, we could already forecast drift.

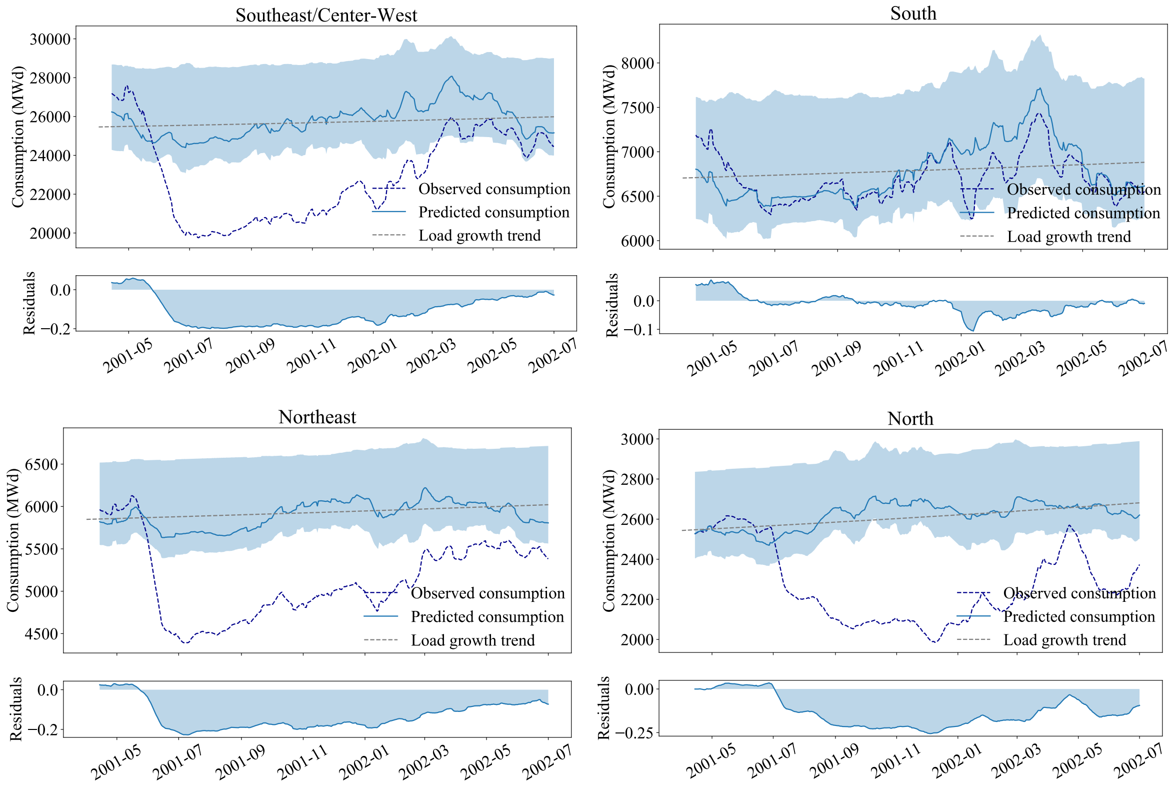
Our third case study period spans from 2001 to 2002. After a long drought with low reservoir levels in early 2001, system operators were concerned that the electric grid

Figure 6.13: Comparison of model performances across periods. Each constituent model is represented by the Cluster from which it hailed. For 2019 data, all models behave similarly to training and we observe a low MAPE. During the year 2020 individual constituent performance becomes erratic implying data coming from distributions different from the ones seen in training. This is to be expected, as October's heatwave registered century-long record-breaking temperatures.



might collapse. To avoid a worst-case scenario, the Federal Government enacted a series of policies aimed at reducing energy consumption by 20% [Bardelin, 2004]. These ranged from awareness campaigns and propaganda to increasing energy prices at peak hours. It was also heavily incentivized to turn off all lights during certain periods of the day, with the intent of keeping all non-essential energy consumption to a minimum to minimize grid strain. An appeal of examining this period is the presence of an exact measure of expected impact, which is uncommon in counterfactual literature and allows for a direct evaluation of our approach. Figure 6.14 highlights the impacts of these policies on all regions, with the South excluded from this rationing plan.

Figure 6.14: Consumption on 2001's Brazilian *Apagão* (Blackout).

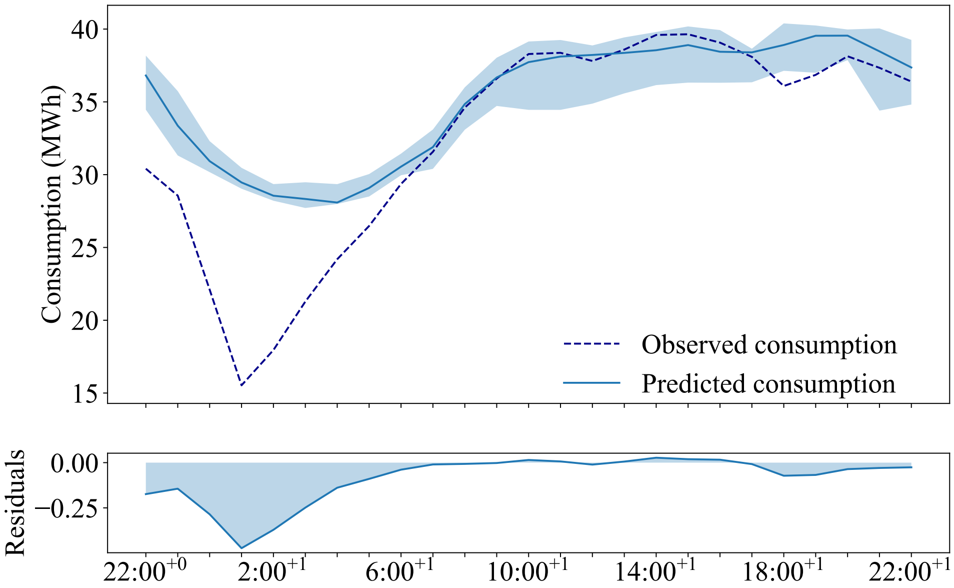


6.3.2 Outages Impact

Extreme events such as the pandemic and heat waves usually leave a lasting impact which, as demonstrated above, can be evaluated from daily energy consumption measures. However, atomic or short-duration events might not provide the same consumption signature. This is true for most outages, given that the recovery time can happen within a few hours. In this respect, using the same daily consumption model might lead to underpredicting their impact. For events such as these, we applied the hourly counterfactual model. Using the same methodology as previously described, we can measure an outage impact by computing the residuals between the counterfactual model output and the observed consumption. Additionally, we can accurately measure grid recovery time. Even if the utility indicates that electricity is completely restored, it might take some time for the grid to stabilize, and this approach accounts for this potential delay.

One of the largest outages in Brazil's history was the 2009 Brazil and Paraguay blackout. According to Brazil's Ministry of Mines and Energy, adverse "atmospheric factors" caused the failure of three transmission lines from the Itaipu Hydroelectric Power Plant, which resulted in its complete shutdown on November 10th, 2009 [Brasil, 2009]. The massive blackout affected not only 18 of Brazil's 27 states, but also the entire population

Figure 6.15: 2009 Brazil and Paraguai blackout on 10th November 2009. The energy was restored between 1:00 and 6:00, and affected an estimated 90 million people in Brazil

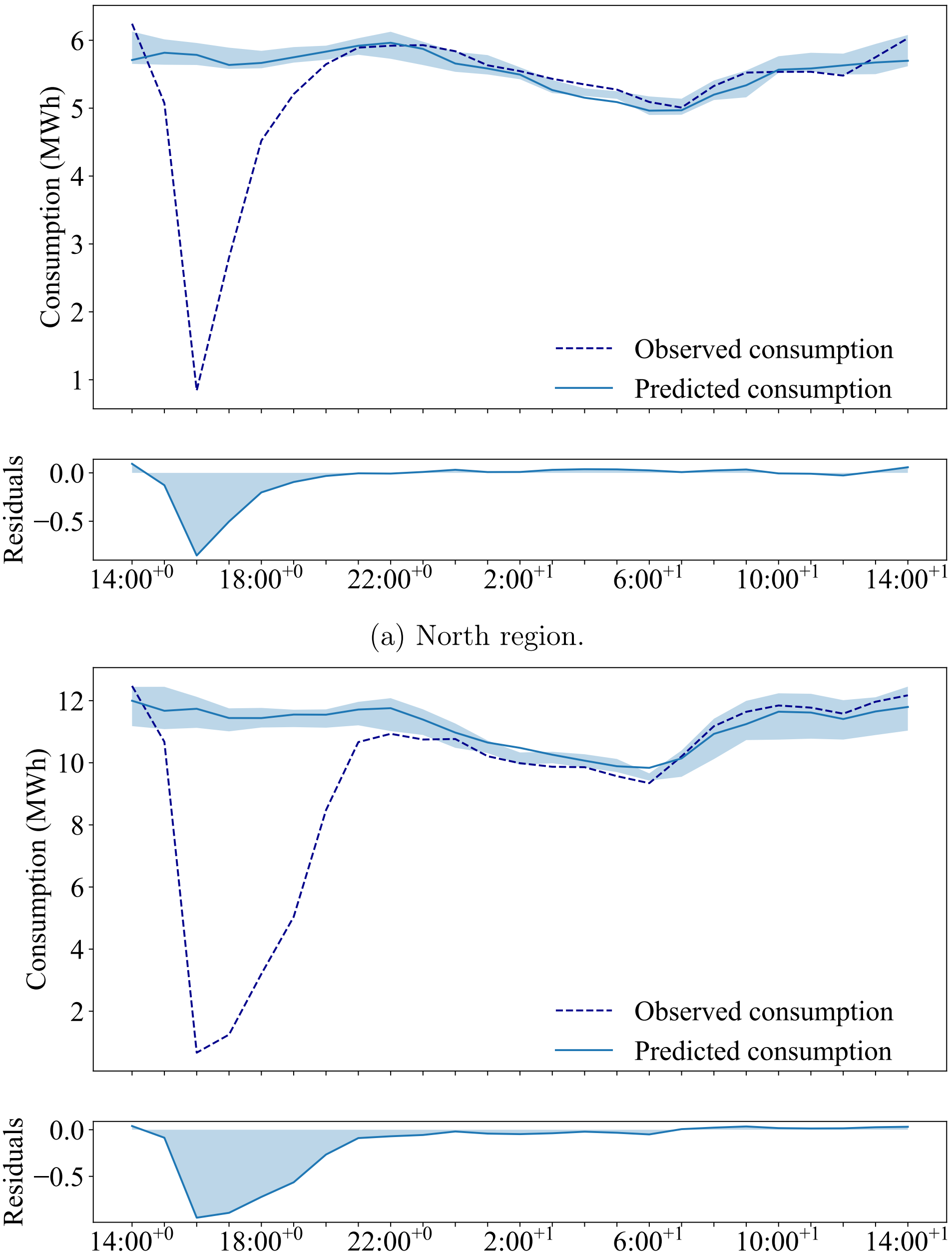


of Paraguay [Conti, 2010]. As a result, four Brazilian states and 90% of Paraguay were left in complete darkness. It was reported that the outage started at 22:13 GMT-3 and that the Itaipu power plant returned to normal operations at 6:00 AM GMT-3 the following day.

There is some debate however regarding recovery time from this outage event. For example, Hudedmani (2019) states that the energy was restored at 2:45 AM GMT-3 [Hudedmani et al., 2019] as reported by national news media at the time. However, Globo, one of the largest news portals in Brazil, states that the southeast region began seeing recovery after 3 hours of blackout Globo [2022], therefore at 1:00 AM GMT-3. Figure 6.15 illustrates the energy consumption for the South-east and Center-west regions which contains all four states that experienced outages during this event. Indeed, at 1 AM the start of the recovery can be seen with consumption only returning to pre-outage levels between 6 AM and 7 AM.

Another wide-scale outage occurred on March 21st of 2018, reaching all regions of the country, with greater intensity and duration in the North and Northeast. According to the Brazilian Electric System Operator, human failure was responsible for the outage, which resulted in an overload of the electric grid and eventually, its collapse [Silveira, 2018]. This also raised concerns regarding the security of the power grid [Liu, 2019]. The outage began at 15:48 GMT-3 when the transmission line connected to the Belo Monte Power Plant failed after not being able to support an increase in load. Belo Monte lies in the Northern state of Pará close to the border of the Northeastern region. However, by 16:15 GMT-3 electricity was already restored in the Southeast and Center-west region, thus these regions were not severely affected. The same cannot be said of the regions that

Figure 6.16: 2018 North and Northeast outage in Brazil, on 21st March 2018. The energy was restored between 16:00 and 21:00 and affected at least 80 million people in Brazil.

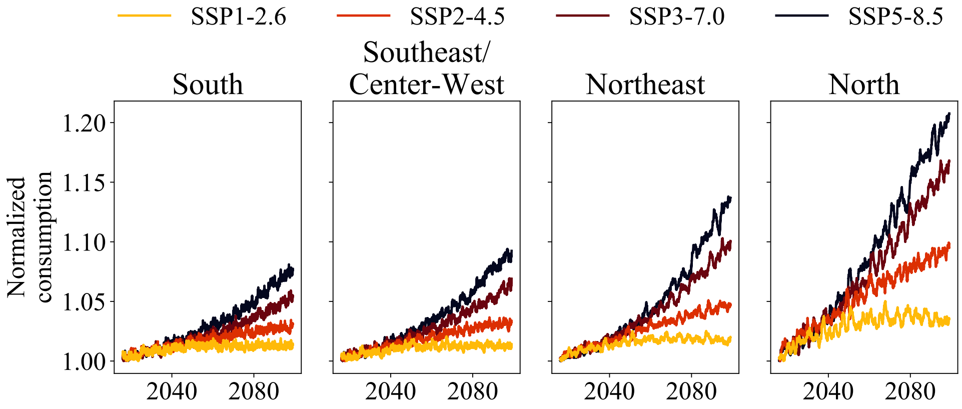


(a) North region.

(b) Northeast region.

depend on Belo Monte more heavily. Figure 6.16 shows the consumption for March 21st, 2018, in the North and Northeast region. Indeed we can see the drop in consumption at the 16:00 hour mark, coinciding with the beginning of the outage. Recovery followed immediately after, with an estimated total grid recovery occurring shortly after 21:00, when consumption returned to pre-outage levels. Computing the area between observed and counterfactual consumption, we obtained approximately a loss of 200% (12MW) and 300% (36MW) of energy for the North and Northeast regions respectively.

Figure 6.17: Normalized consumption forecast for each Brazilian subsystem under all evaluated RCP-SSP scenarios, considering the means of the respective CMIP6 ensembles.



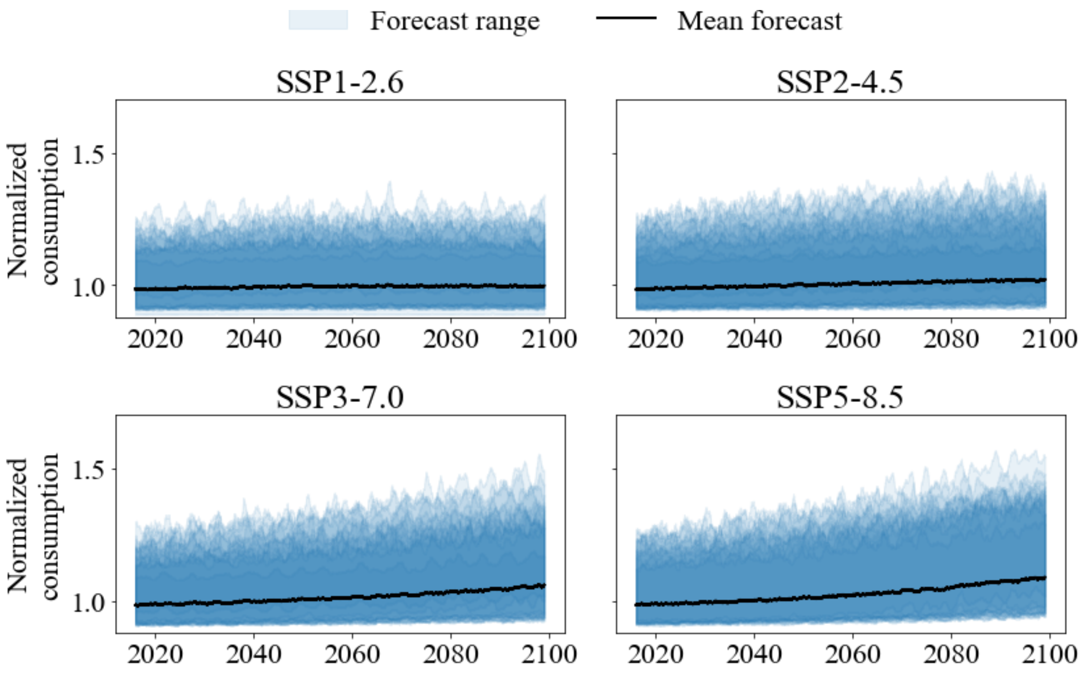
6.3.3 Forecasting Under a Changing Climate

Earth's mean temperature is expected to increase by the end of the century compared to pre-industrial levels, and how much of an increase will depend on future actions taken by human populations to mitigate carbon emissions. Due to climate change, other weather transformations are also probable such as a decrease in the humidity of dry regions and shorter rainy seasons. All these climate changes are likely to impact energy usage, as there is a close relationship between weather and electricity consumption. However, not all regions of the globe will be subject to the same climatic changes. Furthermore, there exist many possible scenarios highlighting different climate change projections. To account for this variability, we propose forecasting energy consumption under 106 distinct CMIP6 models encompassing SSP1-2.6, SSP2-4.5, SSP3-7.0, and SSP5-8.5 scenarios, building ensembles for each pathway taking into account the future projections for precipitation, temperature, humidity, and wind speed.

We explored how each of the Brazilian regions' energy consumption is expected to change under the different warming scenarios in Figure 6.17. North and Northeast, the hotter regions, present a larger increase in energy usage than the remaining ones. This conforms with the expected temperature increases for the continent, as the northern portion is expected to become significantly hotter than the southern portion. When comparing one subregion to another, we can associate the disparity in energetic usage due to heat index. The Northern region encompasses the Amazon rainforest and presents a significantly higher humidity than the primarily semi-arid Northeast. As such, the effects of high temperatures are exacerbated in the form of a larger increase in the perceived temperature.

We also considered the overall consumption increase for the whole country. Each pathway provides minimum and maximum temperature ranges, which we employed to

Figure 6.18: Normalized consumption forecast for each evaluated SSP-RCP pathway on the entirety of Brazil. Each area illustrates the consumption forecast range from an individual CMIP6 model. While the lower bound for both temperature and energetic consumption is similar across all scenarios, there is an increasing range and variance regarding the upper bounds as the scenarios become more pessimistic.

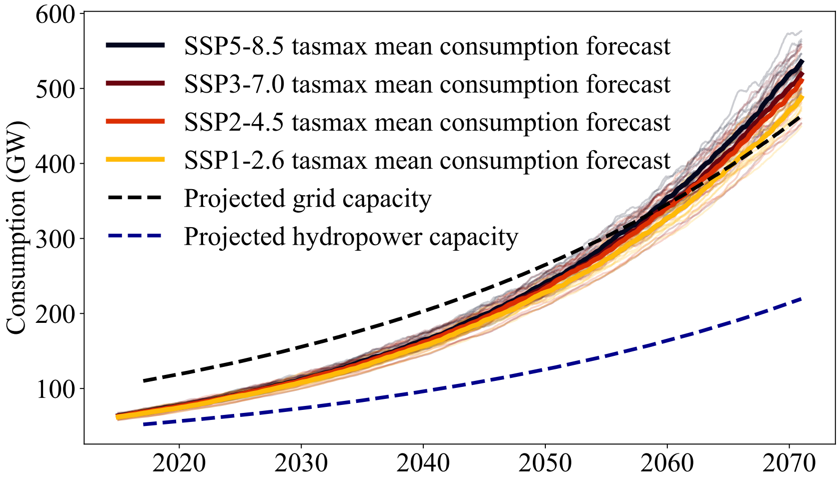


obtain the consumption band in Figure 6.18. Nearly all CMIP6 models provide similar minimum near-surface air temperatures (tasmin), regardless of the pathway. The same cannot be said regarding maximum near-surface air temperatures (tasmax). As the scenarios become more severe in terms of projected climate change, the mean maximum temperature increases and we observe a larger variance near the extremes of the distributions.

For the next five years (2021 - 2026), Brazilian's Energy Research Company (EPE), National Electric System Operator (ONS) and Electric Energy Trading Chamber (CCEE) expect a 3.6% and 2.7% yearly load and power grid expansion respectively [EPE et al., 2021]. Figure 6.19 presents a simulation of daily consumption through the year 2070 assuming that Brazil retains its 60% hydro-based energy supply. We presented 360 rolling day averages to suppress seasonality and evaluate consumption tendencies. Although Brazil currently holds 100GW+ of installed hydro capacity, it is unrealistic to assume that any hydropower will be able to constantly operate at 100% efficiency. Itaipu, the largest hydropower plant in Brazil and responsible for supplying over 10% of country-wide consumption, has historically operated at an average of 61.15%[1] power over the past 36 years. Thus, we considered a mean 60% hydropower operative power to estimate Brazil's total grid capacity.

[1] Available in Portuguese at `https://www.itaipu.gov.br/energia/geracao`

Figure 6.19: Electricity consumption forecast for Brazil through 2070. Light-colored lines represent forecasts from individual CMIP6 models, while bold lines illustrate the mean of the ensemble.



6.4 Discussion

While many statistical methods for demand prediction have been proposed and are demonstrated to be performant [Wang et al., 2021], one primary limitation has been an overreliance on linear or quasilinear relationships [Jiang et al., 2020]. For example, Wang et al. [2021] compares classical statistical methods utilizing linear relationships against more complex statistical approaches and found that gradient-boosting methods have higher performance when predicting city-wide energy consumption. However, this is expected as tree-based approaches are known to struggle with out-of-distribution data [Meyer and Pebesma, 2021]. If applied to demand forecasting this could be problematic as impacts from extreme events and climate change can exceed historical records. In our work, we proposed applying the DARE algorithm to a tree-based consumption model so that it can accommodate out-of-distribution data, which we hypothesized to enable state-of-the-art performance in face of extreme and disruptive events.

Disruptive events can impact electricity consumption, with the most notable recent event being the COVID-19 pandemic, which has contributed to uncertainty in consumption forecasting at the country scale. While the pandemic is an example of one type of disruptive event [Ruan et al., 2020], other disruptive events such as economic recessions can have dramatic impacts on the power sector [Bardelin, 2004]. In addition to such disruptions, long-term social and economic trends, such as population growth and urban development, will likely contribute to load growth. There is also the potential that future climate change – which will result in warmer yearly temperatures – will in turn result in increasing consumption. Warmer temperatures could also change the timing and intensity of daily electricity consumption, with peak electricity demand occurring at different hours of the day and higher levels due to new deployment and utilization of building cooling technologies [Burillo et al., 2019]. Moreover, these challenges and impacts may not be ubiquitous, and in the Brazilian context may differ substantially due to climate, geographic, and economic variations across regions [Swart and Brinkmann, 2020].

While some disruptive events may have impacts that persist across longer time horizons, such as the COVID-19 pandemic, economic recessions, and climatological changes, there are also shorter-term extreme weather events such as heat/cold waves, hurricanes, and flooding that generate dramatic impacts on energy supply and demand (e.g., the 2021 Texas electricity crisis [Kemabonta, 2021]). Research has identified the importance of generating better seasonal, subseasonal, and even weeks ahead forecasts to account for extreme events and their potential impacts on everything from renewable energy generation to electricity demand [Orlov et al., 2020]. However, forecasting these extreme events, and even generating a better understanding of their direct impacts on the electricity system, remains an ongoing challenge [Kumar et al., 2021, Lerch et al., 2017, Vitart and Robertson, 2018]. In acknowledgment of these challenges, industry and regulatory groups are now calling for the introduction of reliability and resource adequacy metrics that better account for extreme weather events [EPRI, 2022].

Counterfactual modeling has emerged as a technique to understand the impacts of disruptive events on the electricity system in a more generalized manner, allowing for comparisons across disparate areas of the world, including North America, South America, Europe, and Asia. For example, Buechler et al. [2022] used a counterfactual modeling approach to understand the impact of COVID-19-related restrictions on electricity consumption across 58 different countries/regions. In this work, counterfactual models of electricity consumption were developed for each country/region to simulate what consumption would have been in the absence of the pandemic and then compared against actual demand during this same period.

Concerning the Brazilian energy system, Belançon [2021] evaluated 20 years of electricity load for the Brazilian grid and investigated scenarios for the balance of electricity supply and demand in 2030 considering projected renewable penetration. In related work, Trotter et al. [2016] generated a large number of realistic weather paths and presented a probabilistic electricity demand forecast for Brazil for 2016-2100. In addition to these research efforts, a spatial econometrics approach was applied in de Assis Cabral et al. [2017] to forecast regional electricity consumption in Brazil. Focusing on the residential sector in Brazil, de Assis Cabral et al. [2020] proposed spatiotemporal models to estimate electricity demand.

In our first set of experiments, we focused on identifying the best algorithm to build a counterfactual model of consumption prediction. lightGBM, although performant, fails to extrapolate beyond training data. This makes it unable to cope with previously unrecorded extreme weather events, such as unprecedented heat waves made more severe due to global warming. We propose augmenting training data with the predictions from a linear regression model in temperatures outside the training scope. This enables the lightGBM model to handle these previously unseen scenarios albeit at a small loss of performance.

In all experiments, the North subsystem remained the most challenging. This region was not part of the integrated energy system until 2014, which induced two distinct consumption distributions. To counteract this, we filtered data on the North before 2014 leading to a scarcity of data. Furthermore, the northern region had the smallest nominal consumption, and as a consequence, small fluctuations led to large relative effects. When exploring the drivers of consumption from SHAP, the most important feature identified is whether a given day is a weekend/holiday, with weekdays being associated with higher consumption. We observe that further stratifying by day of the week is helpful, as Saturdays tend to have higher consumption than Sundays, and Mondays lower than other weekdays, a characteristic also found in Europe [Ziel, 2018]. Since SHAP assumes feature independence, co-related features can share the credit for their importance. This is particularly true for the temperature-derived variables, as they share a causal relationship. Nevertheless, we verify that CDD remains one of the most impactful drivers, suggesting a near-directly proportional relationship – potentially due to Brazil's tropical climate.

Some insights into disruptive event impacts emerge when comparing the counterfactual output with the observed consumption in 2001 when the Federal Government enacted a series of policies to reduce consumption by 20%. From our counterfactual model, the mean and median relative residuals for the Center-West region between June 2001 and January 2002 are $-18.1\%(\pm 1.6\%)$ and -18.6%. This matches the expected -20% reduction on the period in which the restriction policies were in place, from July 1st, 2001 to February 19th, 2002. A similar pattern was also observed in the North and Northeast regions with -19.1% and -18.9% relative residuals.

When evaluating the beginning of 2020 in Figure 6.12, we observe that our counterfactual predictions closely match the observed consumption. In March, state Governors started issuing recommendations about social distancing during the initial stages of the COVID-19 pandemic. During this initial period, electricity consumption decreased with a pattern that remained mostly constant from April to June, with a mean $-8.87\% \pm 1.2\%$ residuals. However, by July we observe a recovery to pre-pandemic levels that coincides with the relaxation of COVID-19-related restrictions. Although more severe restrictions were again imposed in the latter part of the year, consumption was not as impacted compared to the initial stages of the pandemic.

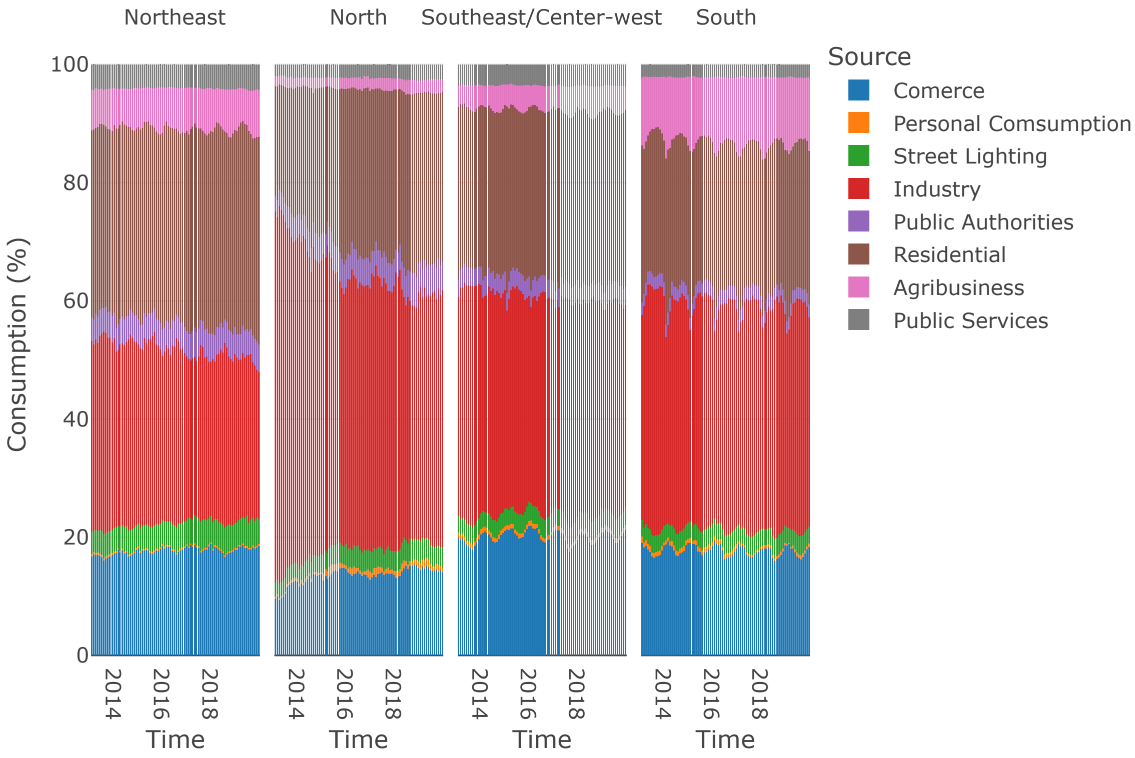
We identified a clear distinction in behavior during weekdays and weekends and proceeded to evaluate them separately. Indeed, during the COVID-19 pandemic, we observe that there is a larger error margin between observed consumption and counterfactual output during weekdays. This finding corroborates with other works addressing mobility and consumption during the early pandemic, in which many commercial and industrial sectors either closed or adopted a home-office strategy. We proceeded to examine the consumption curve shapes and we can observe that curves during the weekdays match the shapes seen during the weekends. That is, due to the higher levels of restrictions to daily life displayed by the Stringency Index, as well as the adoption of the home office, we observe weekday consumption patterns to appear more similar to weekends compared to before the pandemic. For most regions, we see a change from a consumption peak near mid and late afternoon to a steady ramping consumption till late night. This phenomenon is particularly significant for the Southeastern-center western subsystem, the most industrialized region.

Regarding heatwaves, we observe a compound phenomenon of the later-stage pandemic and higher temperature conditions. Many commercial establishments had reopened at this time but a portion of the population remained in the home office. This contributed to commercial and industrial consumption returning to pre-pandemic levels, but residential consumption increasing. One potential explanation is an extreme heat event coupled with the record increase in energy-intensive cooling systems adoption (e.g., air conditioning), as well as less adherence of the population to restrictions [Freitas, 2020]. During the heatwave, we observe erratic behavior from the constituents, which is to be expected when we consider this period to have broken temperature records from the past century and be a clear example of data drift.

We also examined the monthly consumption profiles for each region in Brazil. For all regions, the major sectors are commercial, industrial, and residential loads. We observe that the consumption profile of the Southeast/Center-West region closely matches one of the Southern regions. This might be a reason why the model performed well on both of these subsystems. For the Southern region, we see a significant part of the load tied to Agribusiness which presents a strong seasonality pattern. The Northern and Southeast regions likewise show this same pattern, which cannot be seen in the Northern region. This is to be expected given the extractive focus and the large proportion of the Amazon rainforest in the region. Figure 6.20 illustrates each region's consumption profiles.

We also found there to be a drastic decrease in the industrial prevalence in the Northern region after 2014, with an increase in residential consumption. Further analysis is needed, but one explanation is that in the second half of 2014 this region became part of Brazil's National Interconnected System (SIN), which in turn is tied to a large increase in consumption. It might be the case that many small cities and towns, which are quite abundant in the region as seen by the absence of points in the Northern region of Figure 6.3, were not properly accounted for in the data before this period. Nevertheless, we can conclude that overall, the consumption profiles of all regions did not change much between early 2013 and late 2019, regardless of the financial crisis that happened during

Figure 6.20: Consumption profiles for each Subsystem.



this period. This is a strong indicator that we can expect the same to be true for future projections. The same cannot be said of the 2020 pandemic, which was unlike previous events that impact consumption that Brazil had previously experienced.

Our final experiments explored future periods by applying forecasting from CMIP6 projections. We observe that given Brazil's current sources of generation, it will not be able to sustain its consumption, and it will face shortages sooner than initially thought even under optimistic assumptions. Under this scenario, we can expect a considerable reliance on non-hydro energy sources by 2030 to meet demand. This suggests a need to promote renewable energy alternatives, as currently, the next most available energy resource in Brazil is carbon-intensive thermal power stations. Additionally, under the SSP5-8.5 forcing scenario, we project that Brazil will be unable to meet its consumption by 2070.

Table 6.1: Baseline comparison of algorithms. Highlighted cells represent statistically superior results.

	Model	N		NE		S		SE/CW		Overall	
		R^2	MAPE	R^2	MAPE	R^2	MAPE	R^2	MAPE	R^2	MAPE
First order features	Linear Regression	.610	2.57%	.738	2.40%	.875	3.23%	.826	2.59%	.788	2.76%
	GAM	**.621**	**2.52%**	.712	2.55%	.875	3.21%	.831	2.57%	.783	2.78%
	MARS	.614	2.58%	.723	2.52%	.875	3.16%	.824	2.53%	.787	2.78%
	SVM	.261	3.53%	.529	3.44%	.786	4.49%	.679	3.91%	.616	3.95%
Second order and logarithm transformations	Linear Regression	**.621**	**2.52%**	.721	2.50%	.878	3.19%	.832	2.54%	.790	2.74%
	GAM	.618	**2.54%**	.713	2.55%	.877	3.17%	.831	2.58%	.784	2.79%
	MARS	**.624**	**2.51%**	.713	2.55%	.877	3.16%	.834	2.53%	.788	2.73%
	SVM	.313	3.39%	.566	3.33%	.821	4.11%	.710	3.61%	.649	3.72%
	LightGBM	.610	2.61%	**.773**	**2.33%**	**.887**	**3.05%**	**.850**	**2.50%**	**.854**	**2.64%**
	LightGBM + Linear Extrapolation	.599	2.78%	**.773**	**2.33%**	**.885**	**3.07%**	.845	2.55%	.848	2.69%

Chapter 7

Case Study: Diagnosing COVID-19 from Complete Blood Counts

At the end of 2019, the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-Cov-2) appeared in the city of Wuhan, China, [Wu et al., 2020] which led to a global outbreak weeks later [Hui et al., 2020]. This highly transmissible novel Coronavirus disease was named Coronavirus disease 2019 (COVID-19) [Wu and McGoogan, 2020]. At the time this work is being written, over 630 million cases of COVID-19 infections and over 6.67 million deaths have already been reported worldwide. One of the main challenges for its diagnosis is the list of initial symptoms: fever, dry cough, and/or tiredness [Dias et al., 2020] which are all common in many other respiratory diseases.

Currently, the golden-standard tests for SARS-Cov-2 direct detection include the Reverse Transcription Polymerase Chain Reaction exam (or simply, RT-PCR) and the serology count analysis. The first action of the RT-PCR exam is the use of the enzyme reverse transcriptase to transform the RNA of the virus into complementary DNA. RNA is produced from a DNA molecule and presents information with which it is possible to coordinate the production of proteins. With a complementary probe to a particular virus, it is possible to verify whether the molecular content corresponds to that of the suspected infectious agent. However, in particular, for the case of SARS-Cov-2, the RT-PCR is more efficient at the peak of the infectious cycle [Singanayagam et al., 2020]. This leads to high false-negative occurrences with a sensitivity rate of between 50% and 62% according to Guan et al. [2020], Wang et al. [2020]. Xiao et al. [2020] verified instances of over 20% infected individuals with a positive RT-PCR result only after two consecutive false-negative results. Serology exams have been found to reach a sensitivity and specificity rate of .95+ but only after 15 to 28 days of symptom onset [Iyer et al., 2020]. Furthermore, both exams are relatively expensive and results take longer to process when compared with other kinds of laboratory tests, such as the complete blood count.

Complete blood counts (or simply, CBC) are extensively used for general individual diagnosis [Walters and Abelson, 1996]. As a low-cost test that measures analyte levels of the white and red series in the blood, it is a useful tool to support medical decisions, as intrinsic variations of analytes can bring relevant insights regarding potential diseases.

Patients with most kinds of infectious diseases have noticeable changes in their CBC tests. However, proving that these results can be interpreted as sufficient to support a particular diagnosis is a considerably more difficult task, as changes in analyte values could be easily confounded for different diseases' patterns.

In analyzing complete blood counts of individuals with COVID-19 infection in isolation, we find some changes to be quite characteristic of the disease [Foldes et al., 2020, Formica et al., 2020, Hu et al., 2020]. This implies that machines, which can detect patterns not easily noticeable by humans, could be employed for automatic detection and preliminary screening of the disease. 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. Indeed, many models have been proposed for automated COVID-19 diagnosis through CBCs and omics data. Further, we argue that the detection performance of these models is possibly biased - or overestimated - as many patterns are not unique to SARS-Cov-2. With this, the performance of these models will likely drop significantly as the prevalence of other respiratory viruses increases. This work employs a dataset collected between 2016 and 2021 in partnership with *Grupo Fleury* containing exams of individuals who underwent blood tests in conjunction with RT-PCR exams throughout Brazil, both for COVID-19 and for other pathologies like Influenza-A or H1N1. More specifically, our dataset includes individuals who underwent a CBC at an interval of 60 days before or after a RT-PCR test.

For 2020 and 2021 we collected laboratory data for 900 220 unique individuals, 809 254 CBCs, and 1 088 385 RT-PCR tests, of which 21% (234 466) were positive and less than 0.2% (1 679) were inconclusive. This work does not investigate demographic, prognostic, or clinical data, such as ethnicity, hospitalization, or symptomatology, as these fall out of laboratory scope. We propose modeling the task as a binary classification problem and analyzing two distinct timeframes: one considering the early pandemic stage, namely the first wave of COVID-19 cases in Brazil; and a second stage after November 2020, when the second wave of COVID-19 started, and when we saw the emergence of a new variant of concern, *P1*, which eventually led to the health system collapse in the capital state of Amazonas in late December [He et al., 2021, Naveca et al., 2021].

One of the key highlights of this case study is the analysis of other RNA respiratory viruses. We also collected 120 807 CBCs from 2016 to 2019 of 16 940 individuals who tested positive for Influenza-A, Influenza-B or H1N1, as well as other respiratory viruses, and additionally 307 978 unlabeled CBCs. In particular, these additional CBCs included exams from the 2016 H1N1 surge in São Paulo [Santos et al., 2017], during which the population developed similar hygienic habits to the ones recommended in 2020, like social distancing and the use of masks, although at a minor scale. To the best of the authors' knowledge, this is the most extensive and comprehensive COVID-19-related dataset to date. Therefore, we hypothesize that employing Rashomon Ensembles is not enough if we cannot somewhat reduce biases coming from the data collection stage.

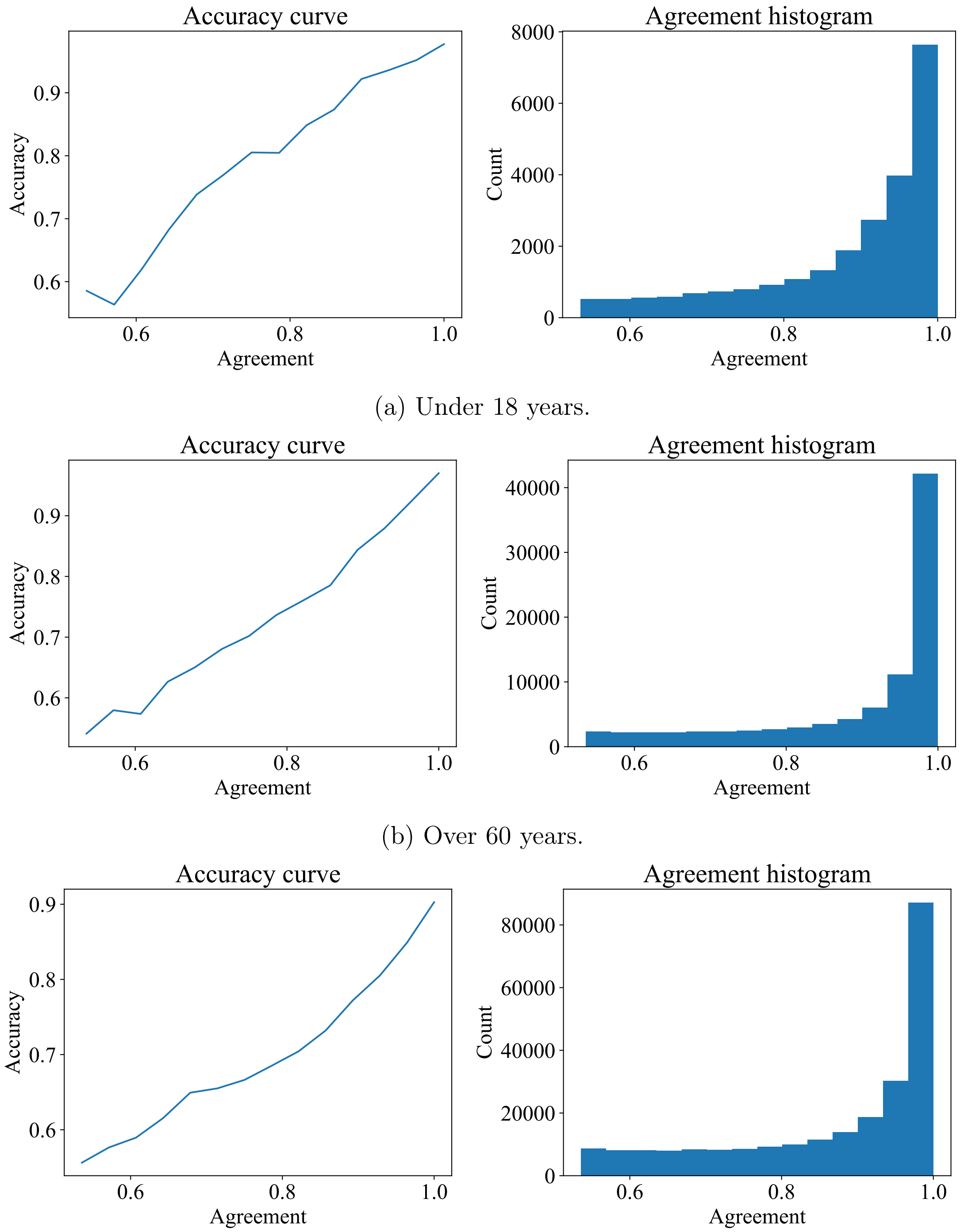
We follow the guidelines provided by the IJMEDI[1] checklist [Cabitza and Campagner, 2021] regarding the application of machine learning to medical data, allowing for both higher quality work and an easier reproducibility and understanding of results. Our analysis focused on patients older than 18 years. We employed our Rashomon ensemble technique to predict COVID-19 and, as one of its advantaghes, used it as means to gain insights of predictions between different age groups. We verified the Accuracy x Agreement curves for both young and old patients and found no significant difference in ensemble performance thus implying that our model is well-tuned, exemplified in Figure 7.1. We believe more experiments are necessary to guarantee performance for children and teens, but data regarding these age groups are present in all training and test sets.

Throughout our experiments, we train an ensemble of machine learning models on this million-scale dataset to predict Sars-CoV-2 positivity. To guarantee the correct labeling of training instances, we focus on the CBC results as close to the first positive result as possible. Our analysis shows that the additional data from other RNA respiratory viruses is a fundamental aspect of properly screening for COVID-19. In the absence of such information, models are prone to confound SARS-Cov-2 with other respiratory viruses or infections. This finding corroborates with many studies that raised concerns regarding bias in COVID-19 research such as Bastos et al. [2020], Palayew et al. [2020], Wynants et al. [2020]. We also demonstrate the necessity of maintaining a model as up-to-date as possible to allow any machine learning model to keep up with the different stages of a pandemic surge. Our model retains high-performance values across multiple evaluation scenarios and on simulations with varying prevalences of COVID-19, properly differentiating Sars-CoV-2 from other confounding viruses, thus demonstrating the robustness of our approach presented in our work [Zuin et al., 2022a].

Our stacking definition extends all previously related COVID-19 learning approaches by building specialized models targeted at confounding viruses. When building the final model, we can expect to learn prediction relationships between COVID-19 and other respiratory infections. For example, in a scenario of a moderately high chance of Influenza, we would need an exceedingly high COVID-19 probability to confirm a positive infection hypothesis.

[1] International Journal of Medical Informatics

Figure 7.1: Relationship between Rashomon Ensemble accuracy and intra-constituent agreement. As we hypothesized, there exists a direct correlation between ensemble performance and agreement. Similar patterns can be found across all age strata. Further, the concentration of data points in high agreement values implies that the returned predictions are mostly trustable and empirical risk found in training can be extrapolated to production.



(a) Under 18 years.

(b) Over 60 years.

(c) All ages.

7.1 Complete Blood Count and Data

The Fleury database structure was created on 10/1997 using an InterSystems Caché and Ensemble, version 1.4[2], a high-performance architecture that is commonly used to develop software applications for healthcare management (Cambridge MA). The database was built using standard healthcare industry practices to ensure the accuracy, completeness, and security of the data collected. The results of the laboratory tests are automatically inserted in a Microsoft SQL database after verification of the RT-PCR output. Within a few seconds, data is replicated to the Cache Database – Intersytems – for permanent storage. Once stored in the database, the result is made available to patients. All users have a username and password, maintained by AD Windows (Active Directory). All registry changes to the database are tracked through a log and are restricted to users with high-level administrative permissions. Information is kept secure through a separate network firewall, accessed only by authorized persons within the Fleury Group's domains. Data stored in this database has been used previously in several clinical studies before the COVID-19 outbreak [Baldo et al., 2019, Brandão et al., 2021, Candido et al., 2020, Chauffaille et al., 2021, Idrees et al., 2021]

This project was submitted, evaluated, and approved by the Research Ethics Committee (CEP) of Grupo Fleury (CAAE: 33790820.3.0000.5474), duly qualified by the National Research Ethics Committee (CONEP) of the National Health Council of Brazil. The Research Ethics Council (CEP) is an interdisciplinary and independent collegiate of public relevance, consultative, deliberative, and of educational character, created to defend the interests of research participants in their integrity and dignity as well as to contribute to research development within the highest ethical standards. By decision of the CEP, since this project uses retrospective and anonymized data, there is no need to apply an e-Free and Informed Consent Term (TCLE) to participating patients.

Quality control is performed daily using 3 control levels for each parameter. Measurements are analyzed using the InsightTM Interlaboratory Quality Assessment Program for Sysmex hematology analyzers, where data from users worldwide are compared. To guarantee equivalence and reproducibility of our analysis and enable the use of common reference intervals for different measurement procedures [Miller and Greenberg, 2021], harmonization of equipment is performed per the Clinical and Laboratory Standards Institute's (CLSI) guidelines [Hayward et al., 2015]. Results are accepted if the percentage difference is less than 50% of the total error for each parameter, which allows us to devise reference values for each measurement [Medicare, 1992, Ricós et al., 2015].

CBC measurements were obtained from EDTA-K3 collected peripheral blood sam-

[2] Caché, InterSystem, 2018; `https://docs.intersystems.com/`; November 2020

ples analyzed by the Automated Hematology Analyzer XT or XN series from Sysmex (Sysmex Corporation, Kobe, Japan). In total, 72 pieces of equipment are distributed in 36 laboratories over the country. Red blood cells (RBC) and platelets were counted and sized by direct current impedance with hydrodynamic focusing and heath flow direct current (DC) detection was used. The hematocrit was determined from the RBC pulse height. The hemoglobin was measured using sodium lauryl sulfate spectrophotometry. CBCs also include the physical features of t RBC: Mean corpuscular volume (MCV) is a measurement of the average size of red blood cells; Mean corpuscular hemoglobin (MCH) is a calculated measurement of the average amount of hemoglobin; Mean corpuscular hemoglobin concentration (MCHC) is a calculated measurement of the average concentration of hemoglobin; Red cell distribution width (RDW) is a measurement of the variation in RBC size. The white blood cells (WBC) and six-part differential were determined by fluorescence flow cytometry. Specifically, the WBC subpopulations were separated based on cell complexity (side-scattered fluorescent intensity), cell size (forward-scattered light), and fluorescence signal (side fluorescent light).

Abnormal increases or decreases in cell counts may indicate an underlying biological process taking place, like inflammation or immune response. Also, values such as the Neutrophil-Lymphocyte ratio, Platelet-Monocyte ratio, or Platelet-Lymphocyte ratio are recognized as inflammatory markers [Nanava et al., 2020]. Table 7.2 shows analyte means and standard deviations, as well as the employed units of measure in each of our cohorts. We can easily identify some patterns that might help us in sorting COVID-19-infected patients from the remaining ones. We can also clearly perceive that the distributions for each gender are slightly different. This is to be expected, as it is known that CBC values vary with age and gender [Bain et al., 2016]. However, introducing an explicit gender variable into our model could entail bias. To avoid this, we instead normalize each analyte by the corresponding gender and age reference values devised by Grupo Fleury, thus building a unified model that considers CBC analyte values regardless of gender. Specifically, we perform normalization by employing the reference ranges as a pivot. Let R be the reference values of an analyte, the general formula scaling features is given as:

$$x' = \frac{x - \Omega(R(x|\text{sex} = s, \text{age} = a))}{O(R(x|\text{sex} = s, \text{age} = a)) - \Omega(R(x|\text{sex} = s, \text{age} = a))} \tag{7.1}$$

where x is an original value, x' is the normalized value, $R(x|sex = s, age = a)$ describes the reference values for x given the sex s and age a of a patient, and Ω and O represent the lower and upper bounds respectively. For example, suppose a male adult presents a 5.0 millions/mm^3 RBC and knowing that the reference values lie in the range $[4.30 - 5.70]$, we first subtract 4.30 from 5.0 and divide the result by 1.4 (the difference between the maximum and minimum reference values), thus obtaining the normalized 0.5 RBC count. Consequently, normalized values above 1 represent abnormally high cell counts. Likewise, normalized values below 0 represent abnormally low counts. Our model analyzes normalized cell counts and their corresponding pair-wise ratios as potential features for building our models.

However, proper feature selection is not the only step needed to ensure performance. It is worth mentioning that CBCs and RT-PCRs are part of different exam batteries, and are therefore often collected on different dates for the same individual. Thus, an important decision is the ideal time frame between the collection of a CBC and that of the RT-PCR test used to validate its label. It is challenging to validate the precise moment the infection has initiated considering the lack of information concerning the onset of symptoms. We also observed abnormalities in the CBCs associated with recovered individuals. These differences could be related to drug usage and/or other therapies, or be due to symptoms that persist even after the virus has been eliminated. In this context, we have the hypothesis that CBCs, even when associated with a positive RT-PCR, may be affected by treatment-related effects. Figure 7.2 shows the concentration distribution of some analytes along with the disease progression time frame. The lower the ratio between white blood cells (WBC) and red blood cells (RBC), the higher the probability of the individual being positive for COVID-19. Additionally, we observed that the lowest value for this ratio lies on day 0. Since our working dataset consists of patients who went to one of Grupo Fleury's laboratories to undertake an exam, we hypothesize that the search for an RT-PCR, in particular for patients who obtained a confirmatory diagnosis of COVID-19, might be associated with the start of symptoms onset, explaining this particular pattern. We did not observe similar behavior for other evaluated viruses, perhaps due to the relative difference in public awareness/concern regarding SARS-Cov-2 and Influenza infections.

Furthermore, we also observed that most analytes tend to present abnormal values for up to 30 days. This might be related to the natural evolution of COVID-19 onto the inflammatory stage, the effects of treatments, or even long-lasting effects on patients' immunological systems. We concluded that the safest and most effective gap to use for labeling CBCs with RT-PCR outcomes' is the 24-hour window centered on the first positive RT-PCR result of an individual, with the remaining frames being highly uncertain about a positive diagnosis, and thus discarded.

To ensure safe labeling, one should also address any gender or age biases. Figure 7.3 presents the age distribution of each pathology subset. We verify a small prevalence in male-positive COVID-19 cases and Female positive Influenza. To address this, we subsampled the training sets to remove possible biases that could jeopardize learning and validated unsampled data to properly verify model behavior in real-world scenarios.

Another point of attention is the possible existence of false-negative results for RT-PCR exams. In particular, we often see cases of the same individual having negative results interspersed with two or more positive results. Therefore, it is also necessary to carry out a pre-processing step to guarantee the authenticity of negative labels and to

Figure 7.2: Average values of the most impactful analytes along with the disease time frame, from 30 days before the first positive RT-PCR result up to 30 days after. N=120 726 patients.

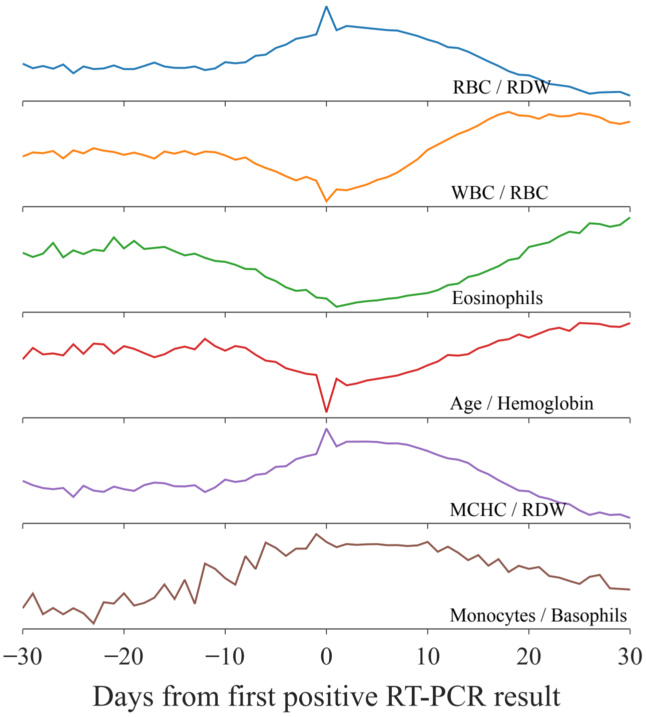
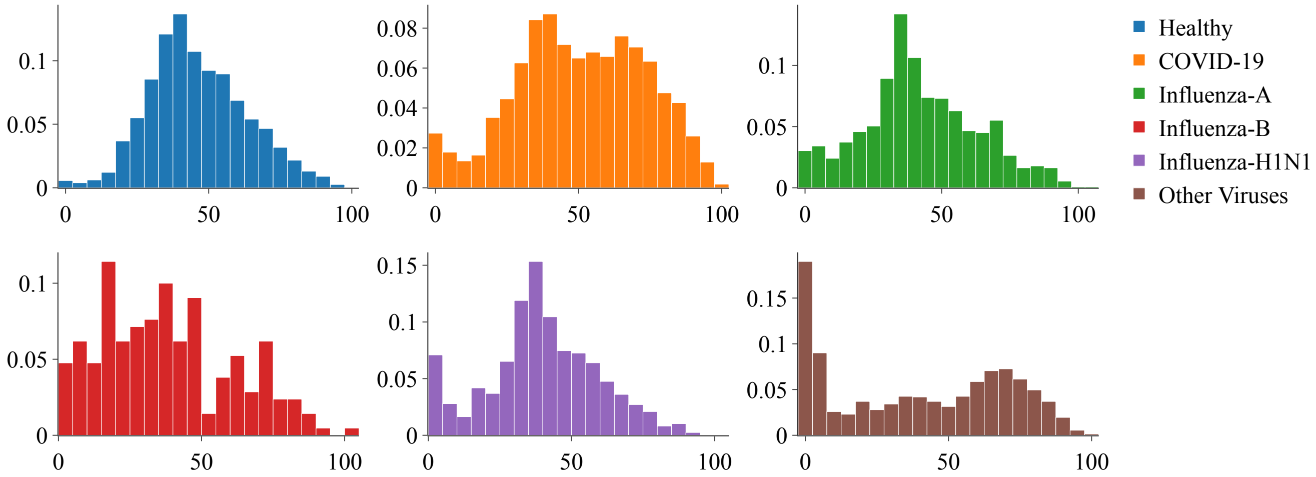


Figure 7.3: Age distribution of the patients across all the evaluated diseases. Only patients with positive RT-PCR were considered. The number of patients available on Table 7.1.



ensure that the model is as faithful as possible to the real scenario of COVID-19, and not to the limitations of the RT-PCR exam. We filter out any negative RT-PCR results issued after the first positive RT-PCR result, thus focusing our analysis on pre-covid individuals and those in the preliminary stages of the infection. We also consider individuals that never had any contact with SARS-Cov-2 in our negative cohort, namely individuals with exams dating before 2020.

Finally, the scale of our dataset allows us to produce high-quality training sets and massive validation sets. Table 7.1 provides the gender and RT-PCR results distribution employed for training and evaluating our models. In addition to SARS-Cov-2, Influenza-A, Influenza-B, and Influenza-H1N1, our dataset also comprehends a variety of other viruses, including Coronavirus OC43, Human Metapneumovirus A, Adenovirus, Parainfluenza 1, Coronavirus HKU1, Enterovirus B, Parainfluenza 2, Coronavirus NL63, Respiratory Syncytial Virus A, Mycoplasma pneumoniae, Respiratory Syncytial Virus B, Rhinovirus, Human Metapneumovirus B, Coronavirus 229E, Chlamydophila pneumoniae, Bordetella pertussis, Parainfluenza 3, Bocavirus, and Parainfluenza 4. We argue that taking this variety of confounding viruses into consideration is of utmost importance to learning models that are specific to COVID-19.

7.2 Building a Rashomon Ensemble

Our main objective in this case study is to predict COVID-19 disgnosis on blood related data. As a secondary objective, we also sought to demonstrate that training a model directly on COVID-19 data is not enough to guarantee robustness if multiple respiratory infections are present, as might be expected to occur in a possible COVID-19 endemic scenario. This was true even in the case of a massive dataset such as the one we employed for our study. We built resilient models for a pre-selected case of core confounding viruses and showed that we can retain similar COVID-19 detection performance in a scenario containing the prevalence of COVID-19 as well as achieving high discriminatory figures in low-prevalence scenarios with an abundance of other respiratory infections. Furthermore, we also demonstrated that the model indeed learns useful relationships between CBC patterns of other respiratory infections. To ensure the relevance of the results, we assess the statistical significance of our measurements through a pairwise t-test [Sakai, 2014] with $p-\text{value} \leq 0.05$ and 5-fold cross-validation.

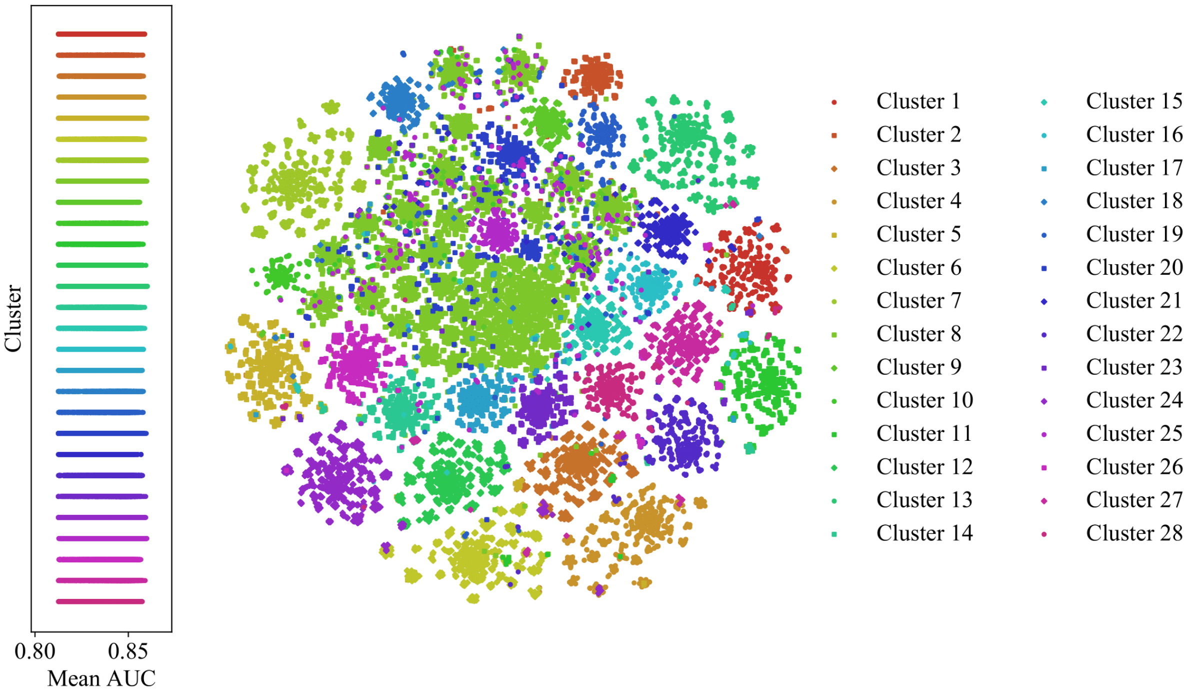
Our models were trained with the objective of distinguishing CBCs (+) from CBCs (−) (refer to Table 7.1). We followed a stacking procedure, that is, the training stage consists of creating multiple specialized models for each of the viruses considered (i.e., COVID-19, Influenza-A, Influenza-B, Influenza-H1N1, and other viruses) and then combining their outputs to obtain a final prediction about the target disease. For the COVID-19 predictor, we employed our Rashomon Ensemble technique with gradient-boosting machines as base models. We divided the training samples into two equally sized batches. The first one was used to train the specialized models and the second one to train the final stacked meta-model. Each specialized model only had access to label information regarding the corresponding virus, and the stacking model employs CBC (+) and CBC (−) labels. Table 7.1 shows training and validation sets for the two waves that occurred during the COVID-19 outbreak in Brazil. The training set for the first wave comprises labeled CBCs acquired until 26 June 2020, whilst its validation set comprises labeled CBCs acquired between 27 June 2020 and 05-September-2020. The training set for the second wave comprises labeled CBCs acquired until 30 September 2020, whilst its validation set comprises labeled CBCs acquired between 01 October 2020 and 28 February 2021. Both training and validation sets contain data corresponding to viruses other than SARS-Cov-2: the training sets contain instances from 2016 to 2018, while the validation sets contain instances from 2019.

Both specialized models as well as the final stacking model were trained with lightGBM [Ke et al., 2017a], a fast implementation of a tree-based gradient boosting technique. We employed the SHAP algorithm [Lundberg and Lee, 2016, 2017, Lundberg et al., 2018] to obtain an interpretation of the model's prediction, allowing us not only to have a probability that a specific CBC is associated with a positive RT-PCR for COVID-19 but also an explanation consisting of the feature importance leading to the model decision. We assessed performance by calculating AUROC, sensitivity, and specificity in the validation sets as well as running 5-fold cross-validation in the training sets. We performed extensive grid-search for hyperparameter tuning for all the aforementioned models. Our final models employ 100 Gradient-Boosted Decision Trees estimators with a maximum tree depth of 50 and a maximum number of leaves of 50. The learning rate was set to $2e^{-1}$ optimizing the binary cross-entropy function.

As the first step of our algorithm, we sampled 100 000 models from the complete model space, considering both raw analytes as analyte ratios as possible features. Section 7.4 presents related literature which helps in finding a suitable baseline for predicting COVID-19 from blood tests and infer a reference model. We considered the AUROC value of .81 as a performance threshold to consider a model minimally performant and establish the Rashomon set, the lowest AUROC value found in the literature. This resulted in a sampled model space $\mathcal{H}'$ containing 47 708 models out of the original 100 000 (47.71% of the models perform better than the minimally performant model from literature). Such a large Rashomon set implies that blood-related features are useful for preliminary disease diagnosis. Figure 7.4 illustrates the induced Rashomon space and the found divisions after clustering models according to their explanatory vectors.

Similarly to our previous case studies, not all CBC analytes are relevant features for differentiating the base targets (i.e., each virus), and some features may be detrimental to the task. To properly perform the subsequent step of our algorithm and find a set of relevant features for each representative Rashomon constituent, we represented the model space as a directed acyclic graph (DAG) in which each node represents a distinct feature

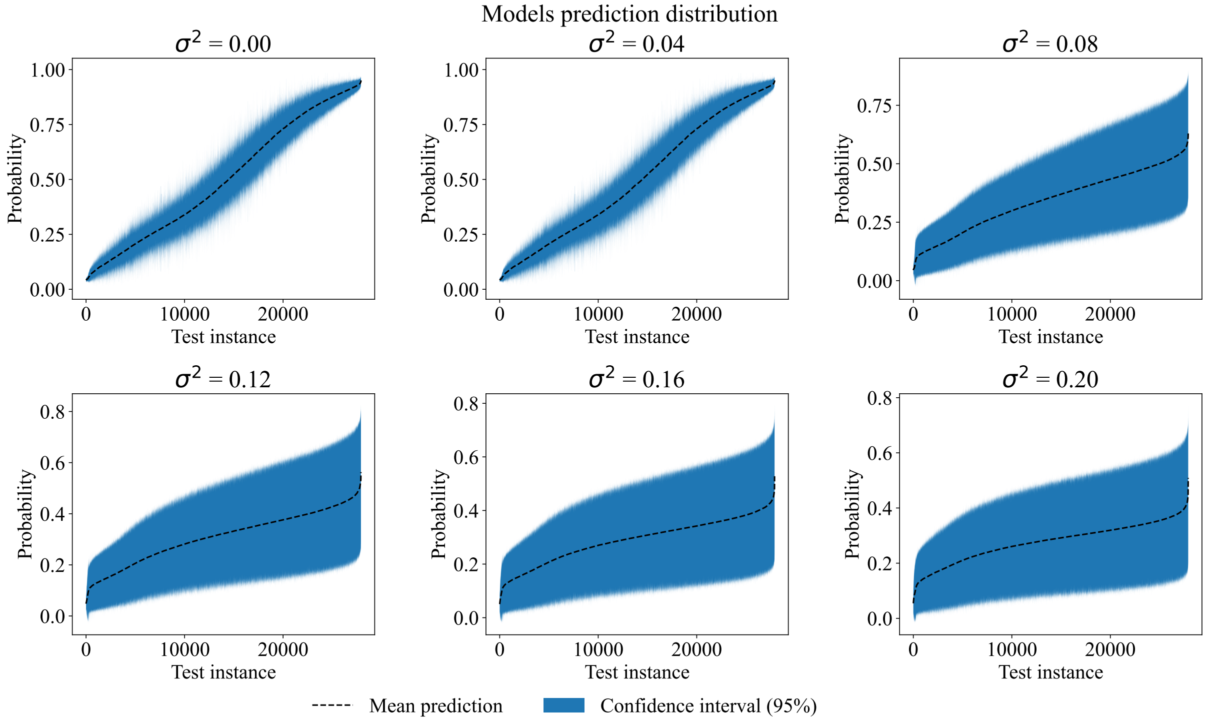
Figure 7.4: TSNE visualization of the Rashomon space. No clear relationship exists between cluster assignment and predictive power. Cluster 11 appears to be more spread over the space overlapping with other clusters, while the remaining ones are mostly concise, reminiscing the steel-plate defects of Rashomon space. N=47 708 models.



subset, and vertex $A \to B$ is connected if B can be reached by simple feature addition from A, thus representing a transitive reduction of the more complex combinatorial complete model space. This modeling approach presents two desirable properties: the first being that any vertex is reachable from the $[\emptyset]$ model, the second being that, for any feature set path, there exists a topological ordering, an ordering of all vertices into a sequence such that for every edge, the start vertex occurs earlier in the sequence than the ending vertex of the edge. These properties imply a partial ordering of the graph starting from the root node, which allows us to search it in an orderly manner. We apply the A* algorithm [Hart et al., 1968], employing as heuristic the AUROC of the model represented by the feature set of a given vertex. We hypothesize that there exists a set of optimal feature expansions that lead to the best-performing models for each specific base task. This allowed us to search the $N!$ combinatorial space of feature subsets to select the best-performing specialized models.

Once models are selected, it is important to verify their suitability as Rashomon constituents. Our main hypothesis is that by exploiting models that disagree under data drift and that encompass diverse explanations, we can build a more robust ensemble. In Figure 7.5, we introduced increasing amounts of gaussian noise to the normalized

Figure 7.5: COVID-19 diagnosis from blood analytes: effect of introducing noise to ensemble constituents' input features.

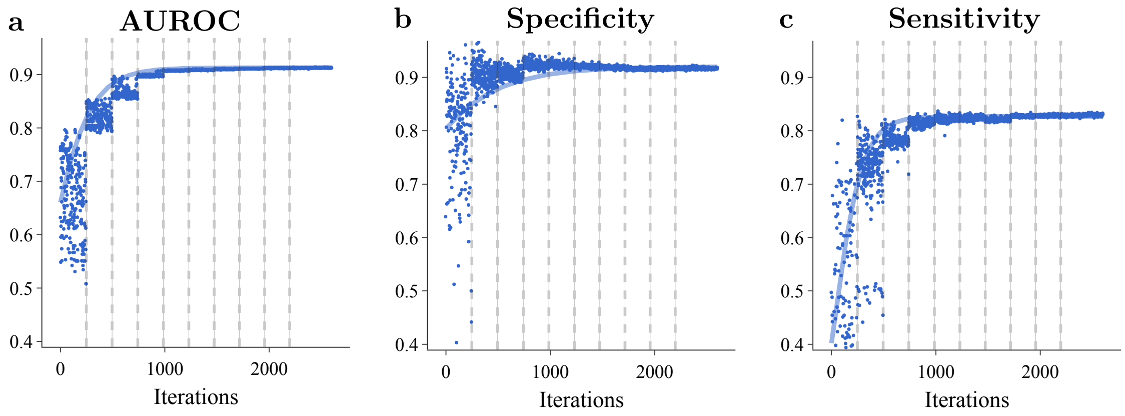


features and observed each constituent's returned probability distributions under each scenario. We verified a direct relationship between noise and the confidence interval, thus signifying that models become more divergent under drift and ensemble reliability decreases, as intended. This suggested that our constituent selection was appropriate and that we could deploy them. The following Section presents multiple experiments in which we applied our ensemble, enabling us to extract relevant scenario insights and deepen the literature behind COVID-19 AI-guided diagnostics.

7.3 Experiments and Results

Our first set of experiments is dedicated to validating that CBCs are useful sources of information for identifying SARS-Cov-2 virus infection. It is worth mentioning that in this initial experiment we did not employ information about infections other than COVID-19 while training the model, that is, CBC (−) is composed only by the sub-population in COVID-19 (−). We trained a COVID-19 model with the labeled CBCs within the first two quarters of 2020 and evaluated it with the labeled CBCs within the third quarter of 2020. Figure 7.6 shows the AUROC improvement as we proceed to include more features in the COVID-19 model. We can verify that employing only three features is already enough to surpass the .85 AUROC mark. Our final COVID-19 model achieves an AUROC of .922, specificity of .918, and sensitivity .824, thus clearly indicating the potential of employing large volumes of CBCs to identify SARS-Cov-2 virus infection. Since electing more than 15 features doesn't seem to significantly improve performance, our sampling step for the Rashomon approach did not explore larger models. Figure 7.7 presents the 15 most important features identified by our algorithm as well as their contribution to the final specialized COVID-19 model prediction.

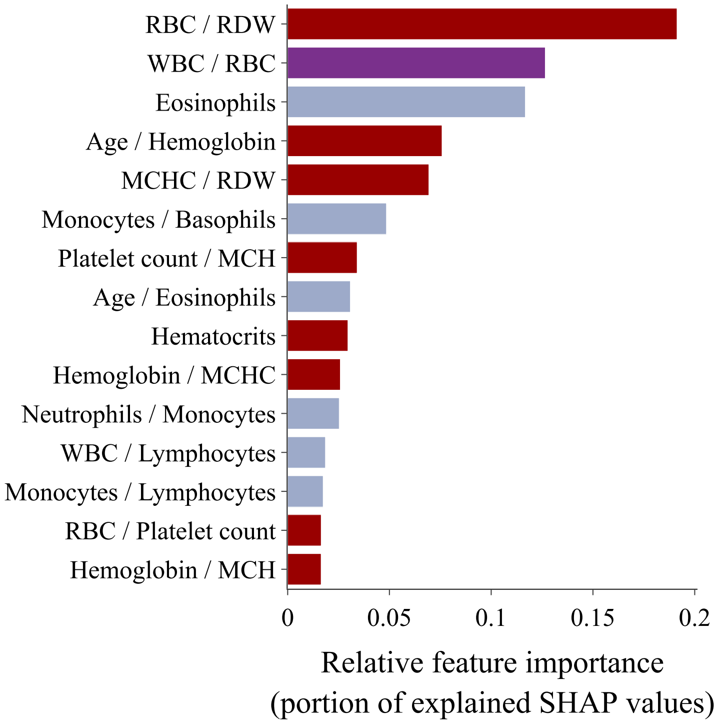
Figure 7.6: Increase in performance as we allow more features to enter the model while performing a greedy search (i.e., each iteration increases the feature size). Each point in the figure represents a COVID-19 model, and the number of features within a model is given according to the corresponding dashed lines. **a** Performance increase of the area under the Receiver operating characteristic curve. **b** Performance increase of specificity. **c** Performance increase of sensitivity. N=75,923 patients.



7.3.1 SARS-Cov-2 Mutations and Variants

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 [Emmerich, 2021]. 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 [He et al., 2021]. To evaluate the performance of our COVID-19 model as the SARS-Cov-2 virus mutates, we trained it on two distinct 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 (as shown in Table 7.1). 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 (as

Figure 7.7: Features in red are exclusively from the red series (roughly 60% of total importance). Features in gray are exclusively from the white series (roughly 26% of total importance). Features in purple involve analytes from both red and white series (roughly 14% of total importance). N=103 822 CBCs.

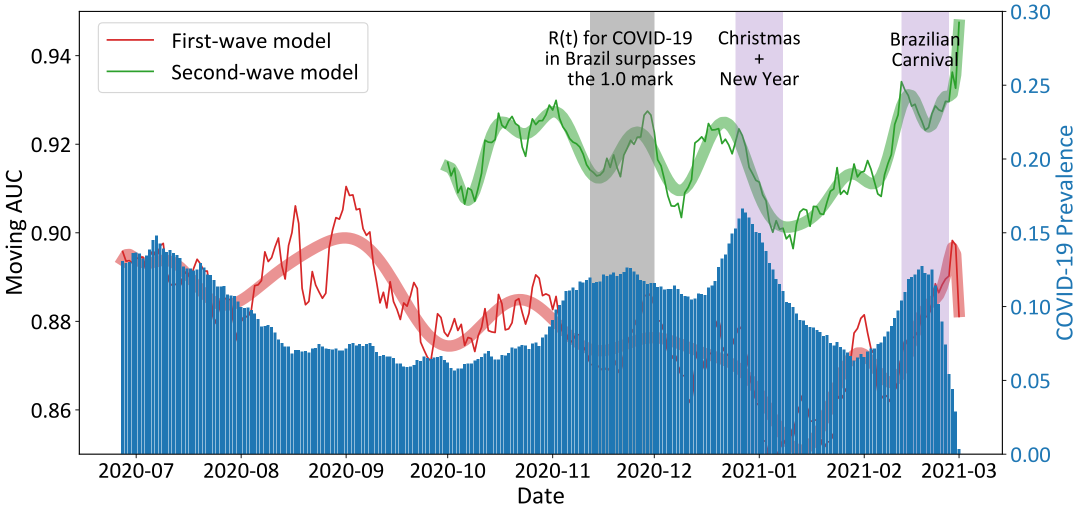


shown in Table 7.1).

Figure 7.8 presents the AUROC obtained after the application of each of these two models during the pandemic, up to March 2021, considering a 7-days sliding window, as well as the respective COVID-19 prevalence (i.e., the proportion of positive cases over all RT-PCR exams in a given period). We investigate three periods of interest: $R(t) > 1.00$, a period in which the SARS-Cov-2 reproduction number was above 1.00 uninterruptedly for several days. During this period the virus spread quickly through the entire country; Christmas + New Year's day, a period in which families reunite, spreading the virus and resulting in a clear increase in COVID-19 cases. This was observed in the entire country; and during Carnival, a period in which large crowds fraternize. Carnival events were canceled for 2021, but many gatherings were reported in some regions of the country, such as Rio de Janeiro, Natal, and Recife.

We evaluate the COVID-19 Rashomon ensemble on both periods using the model trained with data up to October, illustrated in Figure 7.9. Performance on both periods appears to be comparable, thus implying that the constituents were able to properly generalize onto the second wave. We also verify that the Rashomon ensemble remains a suitable approach, outperforming all constituents in either scenario. Further, the empirical

Figure 7.8: AUROC fluctuation over time considering a 7-day sliding window. 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. N=357 956 CBCs.



risk found during training can be used to estimate the empirical risk on production as no significant divergences were observed.

The performance of the First-wave model, trained on data up to May, seems to deteriorate with time mostly as a result of periods of high COVID-19 prevalence due to SARS-Cov-2 variants. On the other hand, the Second-wave model reaches AUROC values as high as .952. The periods analyzed affected the two models in different ways, and the experiment highlights the importance of retraining the models so that they can account for eventual virus variants. Thus, a key concern is an ability to distinguish between different respiratory viruses. After a careful study, we further trained specialized models in an attempt to predict the RT-PCR result for various types of Influenza and other respiratory viruses. Our approach employs stacking to combine the outputs of each specialized model (i.e., COVID-19, Influenza-A, Influenza-B, H1N1, etc.) to perform a final prediction for COVID-19. Specifically, we used half of the training data to learn specialized models, and the other half to train the final stacked model. As illustrated in Figure 7.10, our stacked COVID-19 model achieves performance as high as .913 (cross-validation on the stacking training sets shown in Table 7.1) and .917 (using stacked training and validation sets shown in Table 7.1) while retaining .80 sensitivity and .91 specificity.

Figure 7.9: 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 constituents model behave similarly. We should expect data distributions from late 2020 and early 2021 to be properly represented in data before October 2020.

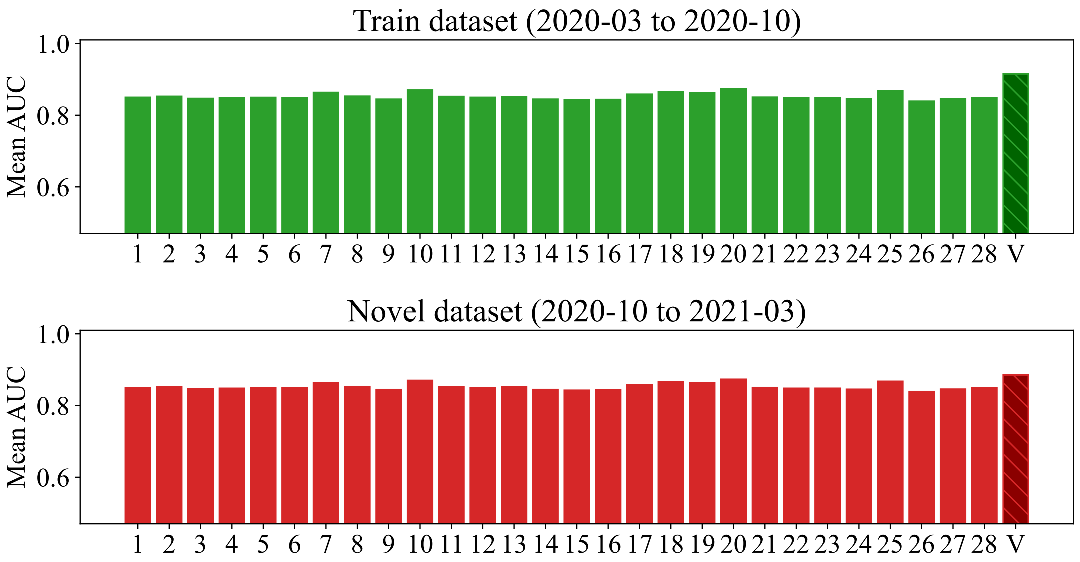
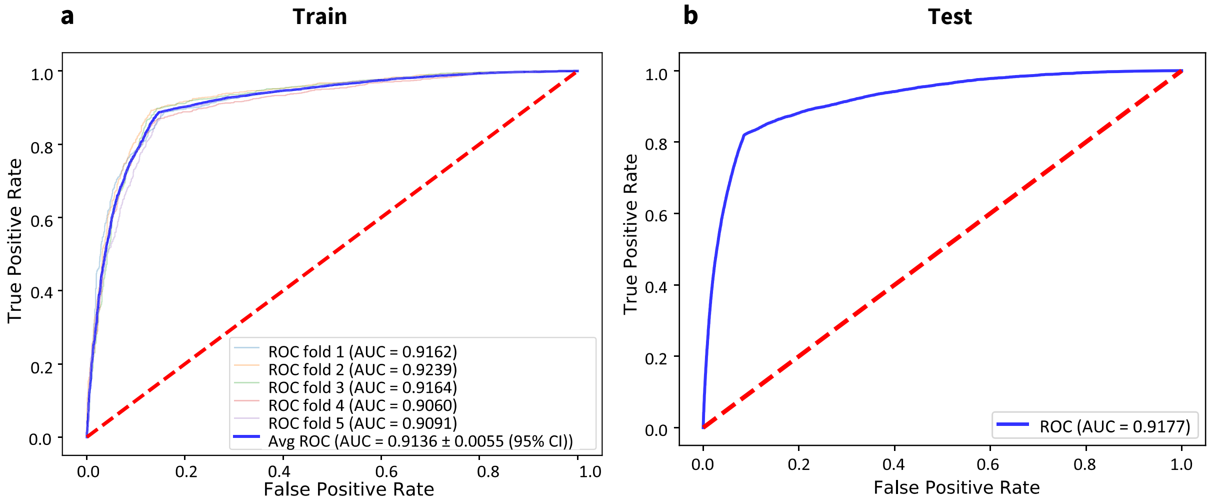


Figure 7.10: AUROC values for the proposed stacking model. **a** Cross-validation performance. **b** Test set performance. N= 91 014 train CBCs and 261 630 test CBCs.

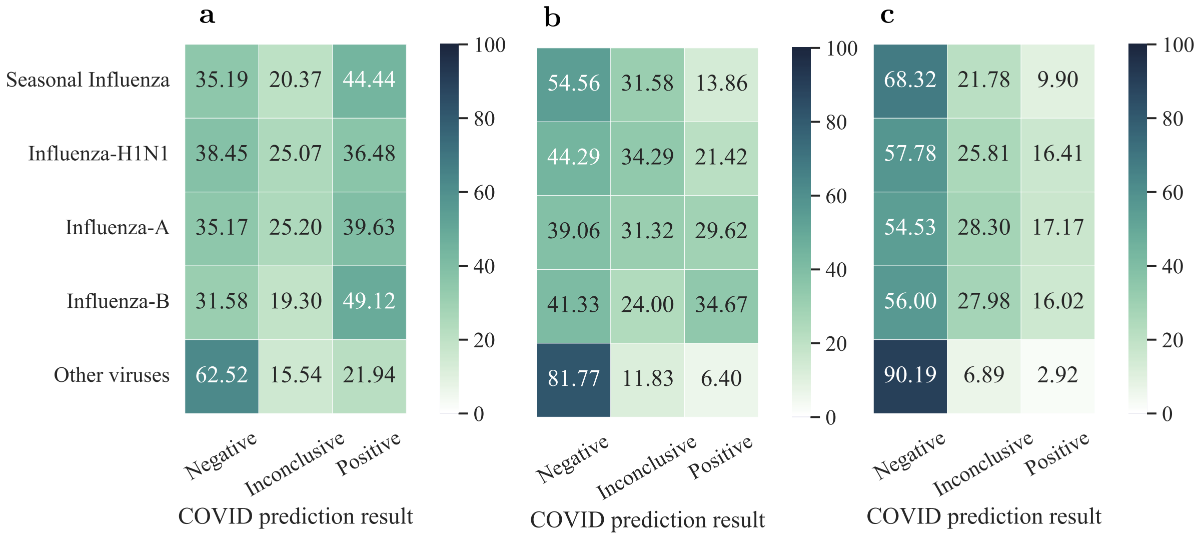


7.3.2 Presence of Confounding Diseases

While the stacked model achieves high performance in predicting COVID-19, it is also important to verify its specificity by analyzing the predictions performed for individuals infected with viruses other than SARS-Cov-2. Figure 7.11 shows how different models perform specifically on individuals that were infected by some viruses in 2019. The ideal result would be all predictions being negative for COVID-19. As discussed before, models trained solely on SARS-Cov-2 data are very effective in identifying COVID-19 cases, but the result on 2019 data indicates that these models performed poorly on other viruses (Figure 7.11a). Including viruses other than SARS-Cov-2 during training increases the performance of 2019 data (Figure 7.11b). The stacked model proves to be much more specific for COVID-19 than both previous models (Figure 7.11c).

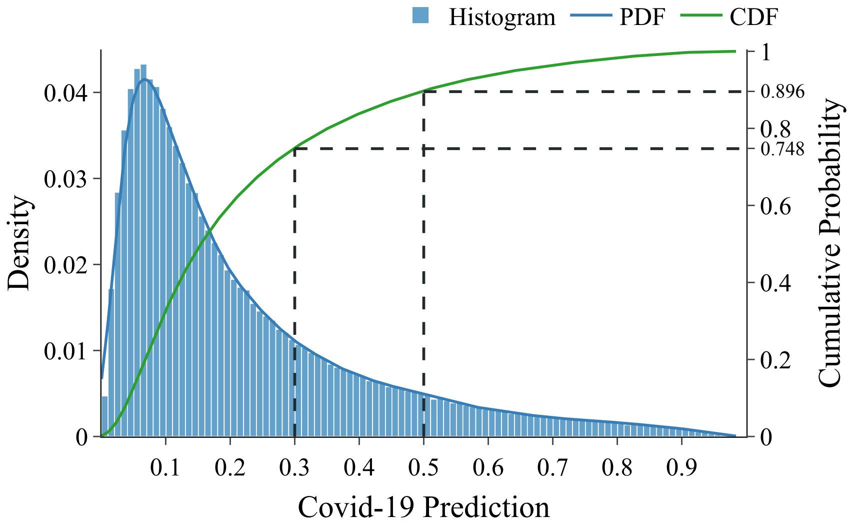
Figure 7.12 also investigates the specificity of the stacked model by showing the prediction distribution on the 2019 data (i.e., individuals infected by a virus other than SARS-Cov-2). The stacked model associates 0–10% COVID-19 probability to roughly 44% of the predictions on 2019 data. Furthermore, the stacked model correctly places almost 80% of the evaluated individuals below the 30% COVID-19 prediction mark, with over 40% being placed below the 7% probability mark.

Figure 7.11: Results of different models evaluated on 2019 individuals with confirmed RT-PCR results for diverse viruses, including Influenza-A, Influenza-B, Influenza-H1N1, and Seasonal Influenza. **a** Model trained only on SARS-Cov-2 data. CBC (−) includes only COVID-19 (−). **b** Model trained using data of diverse viruses, including SARS-Cov-2. CBC (−) also includes viruses other than SARS-Cov-2. **c** The stacked model. CBC (−) also includes viruses other than SARS-Cov-2. Specialized models are trained using half of the training sets, and then these specialized models are combined using the other half of the training sets. N=11 116 CBCs.



We also consider how the model would perform in an endemic scenario in which individuals infected with SARS-Cov-2 could be scarce, and where other types of confounding viruses might be present. To simulate different scenarios, we evaluate the stacked model on data with different COVID-19 prevalences. Specifically, we sample exams from the second wave validation and 2019 data to control the COVID-19 prevalence. The main goal is to stress the stacked model by presenting cases before any safety and/or social

Figure 7.12: Stacked model probability of predicting COVID-19 on 2019 data. Nearly 90% of the cases lie below the 50% covid probability threshold, with roughly 75% being concentrated below the 30% probability threshold. N=307 978 CBCs.



distancing policies could take place, in an attempt to mimic what could happen in an endemic future. These results are summarized by the AUROC, sensitivity, and specificity numbers for each evaluated COVID-19 prevalence presented in Table 7.3. To guarantee statistical significance, we perform 30 repetitions of each simulation and present the respective 95% confidence intervals. The stacked model proved to be robust on varying levels of COVID-19 prevalence.

7.4 Discussion

The CBC is a simple and inexpensive exam. It is part of most laboratory routines, so *"astute practitioners may use nuances and clues from the CBC in many clinical situations"* [Walters and Abelson, 1996]. Liu et al. [2021] devised a high-accuracy risk assessment tool that can predict mortality for COVID-19 through CBCs. Tan et al. [2020] verified that the low count of white blood cells is related to COVID-19 severity by analyzing 12 death cases of COVID-19 and 18 individuals with moderate to severe symptoms, verifying low lymphocyte percentage in most of the cases. Although our dataset had no indicator of severity, we did find a drop in lymphocyte count the closer individuals were to their first positive RT-PCR results, corroborating this finding. Furthermore, we also verified many other analytes that shared a similar pattern. Although more research is needed, we believe that the key analytes indicated by our model might provide possibilities for future research. Literature suggests that there might be existing intrinsic relationships between analytes that might be characteristic of COVID-19. For instance, Nalbant et al. [2020] found that the neutrophil/lymphocyte ratio (NLR) might be particularly typical of COVID-19 infection. However, there is a profusion of other possible promising ratios and patterns currently being under-analyzed for the sake of COVID-19 diagnosis. One of the secondary goals of this case study was to investigate this hypothesis, and we confirmed that our search algorithm tends to favor ratios over analyte count values.

We identified several works attempting to exploit blood counts to detect COVID-19 with the help of machine-learning algorithms. Avila et al. [2020] trained a naive Bayes classifier with data from 510 individuals admitted to hospitals presenting COVID-19-like symptoms with a reported AUROC of .84. Silveira [2020] devised a solution based on gradient boosting machines that focuses primarily on white series analytes. They achieved an AUROC of .81 in a dataset composed of anonymous data from 1 157 individuals. Banerjee et al. [2020] trained both a shallow neural network as well as a random forest model to distinguish COVID-19 cases on data from 954 individuals, reaching an AUROC of .94 for those who were admitted to the hospital with severe symptoms and an AUROC of .80 for individuals with mild symptoms. Cabitza et al. [2021] evaluate different machine learning algorithms on both a COVID-19-specific dataset as well as another dataset including individuals who exhibited pneumonia symptoms in 2018, consisting of data from $1,624$ cases. By exploring a variety of biomarkers, including the analytes from CBCs, they were able to achieve an AUROC of .90. However, a point of concern for such studies is the data scale. We know from the literature that complex machine learning models are prone to overfitting and, with small sample sizes containing only a few hundred individuals, all these works are at risk of presenting unreliable results and overestimated performance.

Wynants et al. [2020] provided a study of 37 421 research titles, with 169 studies describing 232 prediction models, of which 208 contained unique, newly developed models. These models contained both a diagnostic solution to identify suspected infection cases as well as a prognostic evaluation. One of the key findings was that all models were at high ($n = 226, 97\%$) or unclear ($n = 6, 3\%$) risk of bias according to an assessment with PROBAST, suggesting a risk for unreliable predictions when employed in the real world. A similar finding was also reported by Bastos et al. [2020], which verified that, out of the 49 risk assessments performed over 5 016 references, and 40 studies, 98% reported a high risk of individual selection bias. Only 4 studies included outpatients and only two performed some sort of validation at the point of care. This kind of problem is not specific to COVID-19-related research and has been present in many previous medical studies. As mentioned by DeCamp and Lindvall [2020],

> *"... failure to proactively and comprehensively mitigate all biases − including latent ones that only emerge over time − risks exacerbating health disparities,*

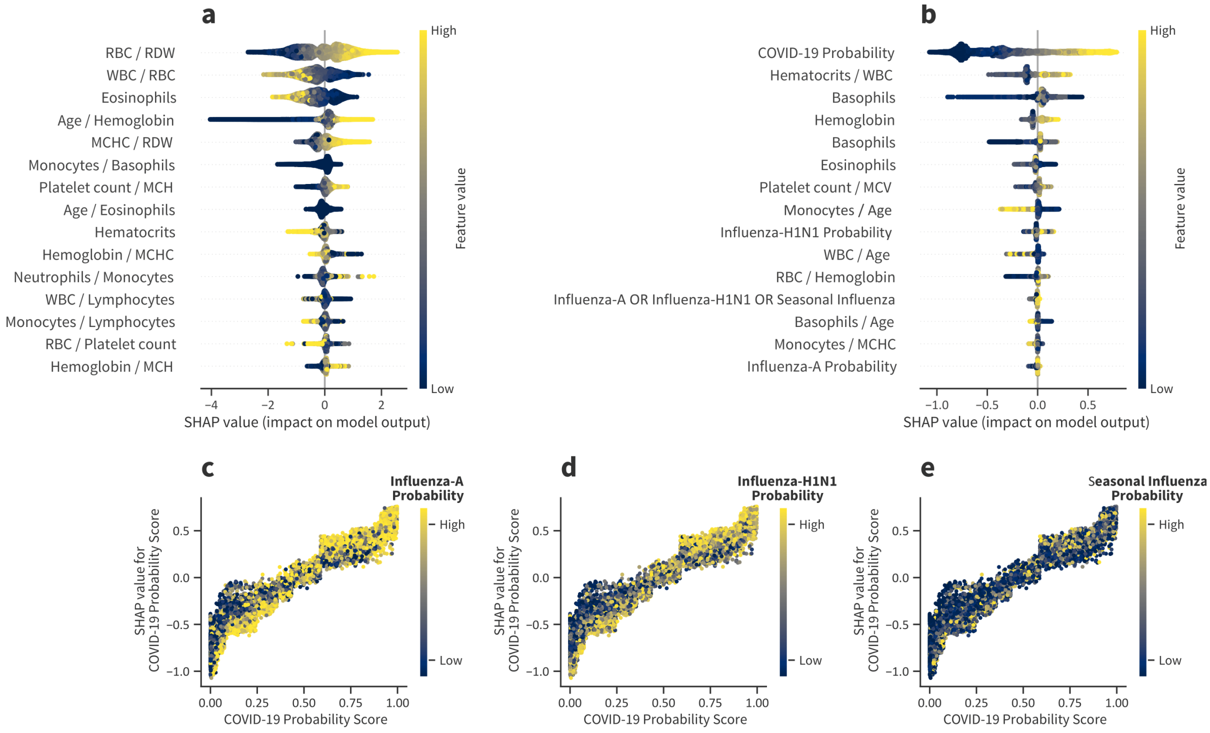
> *eroding public trust in healthcare and health systems, and somewhat ironically, hindering the adoption of AI-based systems that could otherwise help individuals live better lives."*

With that in mind, it is important to highlight the work of Soltan et al. [2020] which, with the help of the Oxford University Hospital, included 114 957 individuals in a COVID-negative cohort and 437 in a COVID-positive cohort, thus establishing a dataset of 115 394 individuals for a full study. Before our work, this was the most extensive COVID-19 study to date. While exploring a variety of scenarios regarding COVID-19 prevalence, they reported AUROC values ranging from .88 up to .94 if their model employs additional data from CBCs, blood gas, and other vital signs collected in routine clinical exams. However, one key concern in this study is the low prevalence of Influenza-like infections ($< 0.1\%$), which drew our attention to a different kind of selection bias in COVID-19 research. Due to the hygiene habits acquired by the population worldwide after the pandemic outbreak, we believe that many other confounding diseases might be underrepresented in most performed datasets. As such, models might be learning patterns that are associated with a general infectious condition rather than specifically with COVID-19.

Our concern regarding data bias in the latest COVID-19 research appears to be valid, as was verified during our experiments assessing performance on data before 2020. Several instances of individuals with different variants of the Influenza virus were initially labeled as potential COVID-19 infected, which we knew not to be true. As such, we devised an approach to insert information regarding other diseases into our model without harming accuracy. In particular, we explored two approaches: the first one being simply retraining our specialized model with the added data of negative COVID, whilst keeping positive results for other diseases. The second approach had the objective of creating an ensemble of models with constituents specialized in other virus infections. We observed similar AUROC results between both, with the first one having a slightly higher AUROC result at the cost of lower differentiation capabilities.

We plot the importance of each feature for every individual, and these results are shown in Figure 7.13. Yellow points are associated with individuals for whom the corresponding feature shows a relatively high value. Blue points, on the other hand, are associated with individuals for whom the corresponding feature shows a relatively low value. Furthermore, there is a vertical line separating individuals for whom the feature is contributing either to decrease (left side) or increase (right side) the probability of active SARS-Cov-2 infection. Figure 7.13a shows the COVID-19 specialized Rashomon ensemble, and the CBC patterns shown in the figure are not specific to COVID-19, as discussed in previous experiments. Figure 7.13b shows the stacking model, where the COVID-19 specialized model is included as one of the features (i.e., COVID-19 probabil-

Figure 7.13: Learned analyte patterns and disease prediction relationships. **a** SHAP Summary plot for the COVID-19 specialized model. **b** SHAP Summary plot for the stacking model, which combines different specialized models and CBC patterns. **c** Partial dependence plots with the relationship between COVID-19 specialized model's predictions and Influenza-H1N1. **d** Partial dependence plots with the relationship between COVID-19 specialized model's predictions and Influenza-A. **e** Partial dependence plots with the relationship between COVID-19 specialized model's predictions and Seasonal Influenza. N=75 923 CBCs.

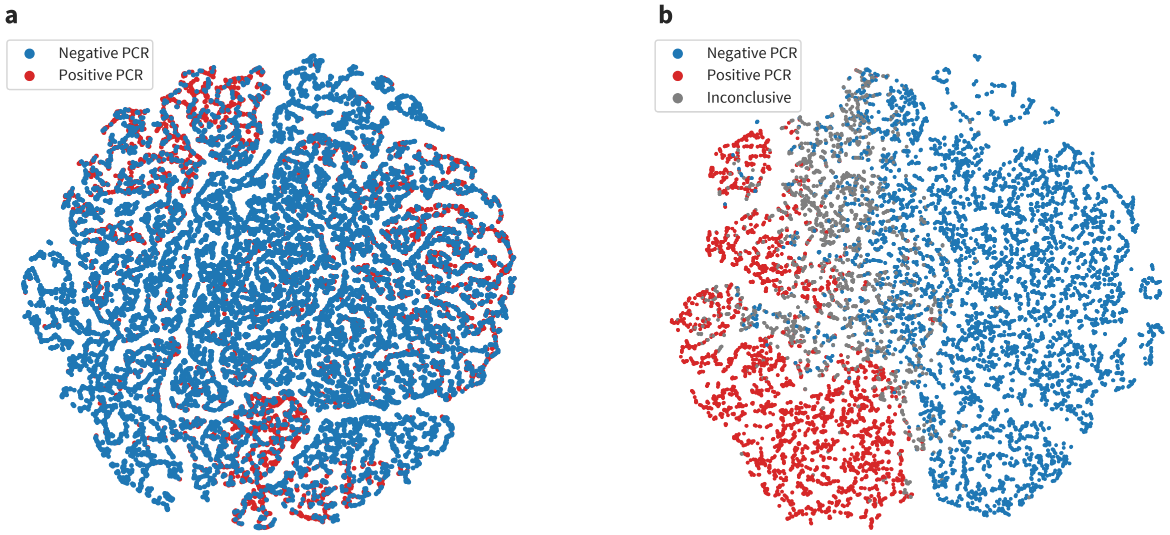


ity). As the stacking model takes into consideration the probability of diverse infections, COVID-19-specific CBC patterns are found.

The stacking approach allows us to study how the physiological patterns found in CBCs of different diseases co-relate. Figure 7.13c to Figure 7.13e illustrate dependence plots of our COVID-19 specialized prediction concerning remaining diseases, which present relevant patterns that enhance the credibility of our approach. For instance, looking at the right portion of Figure 7.13c, we observe a concentration of high Influenza H1N1 predictions (yellow points) on the upper side of the plot, with a similar pattern on the left side of the plot and a concentration on the lower portion. This behavior shows us that in cases of suspicion of H1N1, the overall prediction of COVID is significant, be it to confirm an H1N1 hypothesis (left side) or rule it out (right side). However, when there is a lower probability of H1N1, we likewise see a lower scoring attributed to the COVID-19 model. The stacking ensemble learns to use the information regarding all diseases for these hard-to-predict individuals. We observe similar patterns in Influenza-A (Figure 7.13d), and Seasonal Influenza (Figure 7.13e).

Employing Shapley values as an explanation technique not only allows us to understand the model's final prediction but also to understand the testing time frame. Figure 7.14 shows a 2D representation of the tests of several individuals contained in the dataset and their respective RT-PCR results for COVID-19. In Figure7.14 we observe no clear distinction between exams of infected or healthy individuals and represent what might be observed in an attempt to draw linear correlations between analytes. Figure Figure7.14b shows a visualization of the decision process of the model in the shape of a 2D representation of the returned Shapley values. This scenario reflects all the non-linear relationships present in a CBC that might be challenging for humans to extrapolate on their own. Not only can we draw clear divisions between both individual populations but we are also able to infer a measure of confidence. The closer to the decision boundary, the higher the uncertainty of the prediction and, thus, the more important the discerning capabilities when combining these results with other relevant factors for diagnosis, such as reported symptoms and possible disease onset period.

Figure 7.14: Representation of the CBC space. **a** TSNE representation of the CBC space using the analyte's raw data. **b** TSNE representation of the CBC space using analyte's Shapley values and model predictions. N=16,958 patients.



Predicting data from the second wave proved to be particularly hard, as we observed a deterioration in the performance of our first wave model as time went on, which might be associated with concept drift. In particular, we observed that the peak in performance on the second half of the chart is associated with a lower COVID-19 prevalence, which implied that the model was losing its ability to predict COVID-19 infections. We hypothesize some explanations for this behavior, including the effect of the distribution of COVID-19 prevalence in 2020 and across 2021, as well as the prevalence of other possible confounding diseases, which changed as restriction measures were lifted. Likewise, one of the main characteristics of the second wave is the emergence of a new COVID-19 strain, namely the P.1 variant that ran rampant in Brazil during the analyzed period. It might be the case that the physiological reaction of the body to the new strain was distinct from the earlier variants, resulting in degradation in performance. Finally, another possibility is that RT-PCR tests at the time of evaluation might not have been tuned to properly identify the new strain, thus inducing a divergence between model output and ground-truth data due to possible false negatives.

The proposed solution consisted of employing data close to the start of the second wave, simulating a scenario where we keep the model as up-to-date as possible before the start of a new pandemic stage. Although we could not test for each of these hypotheses, the proposed approach should solve all of the three possible explanations described. With this approach, not only did we verify a higher performance from the start, but the model was able to largely mitigate the concept drift phenomena, retaining an AUROC above the .90 threshold throughout most of the evaluated period. When observing the Accuracy x Agreement, we were able to observe that the curves for the training and deployment stage were similar, as well as individual constituent performance, thus presenting indications that our technique was successful in handling this issue.

A point of attention that should be addressed by any health professional when employing our approach is the presence of co-infections. For instance, multiple cases of COVID-19 hospital cross-infections have been identified [Chen et al., 2020]. As we do not have data explicitly concerning co-infections, we cannot provide insights regarding the blood profiles that emerge in such situations which might confuse the model. It is also important to highlight the impact of ethnicity on CBC results [Lim et al., 2015]. Although the large data sample and the demographic plurality of Brazil serve as indicators of robustness, further testing is needed to understand if the Brazilian model can be directly applied to other contexts. Nevertheless, our method is generalist to an extent that the achieved results could be potentially replicated anywhere on Earth if data concerning a specific region/scenario is collected.

Table 7.1: Entire dataset, training sets, and validation sets for the two waves that occurred during the Brazilian COVID-19 outbreak. Training sets were obtained after applying the inclusion-exclusion criteria to the entire data and downsampling the COVID-19(-) class in the training sets to account for class unbalance. We considered October 1st as the split point between the first and second wave data to eliminate possible incubation periods before the start of the second wave in early November. As such, validation for the first wave encompasses data from late June to late September, and validation for the second wave ranges from early October to late February. N=1 138 728 CBCs.

Entire data

	CBC (+)	CBC (−)				
Gender	COVID-19 (+)	COVID-19 (−)	Influenza-A (+)	Influenza-B (+)	H1N1 (+)	Other (+)
Male	11.3%	34.0%	46.7%	46.5%	48.4%	59.5%
	(122,793)	(369,787)	(3,160)	(1,384)	(4,108)	(20,107)
Female	10.3%	44.4%	53.3%	53.5%	51.6%	40.5%
	(111,673)	(482,453)	(3,604)	(1,588)	(4,380)	(13,691)

Training set: first wave data

Male	12.9%	9.8%	4.2%	2.1%	6.0%	12.8%
	(5,859)	(4,469)	(1,895)	(975)	(2,742)	(5,825)
Female	12.1%	15.2%	4.9%	2.8%	6.9%	10.3%
	(5,527)	(6,918)	(2,223)	(1,214)	(3,118)	(4,656)

Validation set: first wave data

Male	4.9%	37.6%	1.0%	<0.1%	1.4%	2.3%
	(5,808)	(44,637)	(1,113)	(188)	(1,660)	(2,710)
Female	4.7%	43.4%	1.1%	<0.1%	1.6%	1.7%
	(5,647)	(51,550)	(1,343)	(134)	(1,842)	(2,028)

Training set: second wave data

Male	25.9%	10.5%	2.0%	1.0%	3.1%	6.3%
	(24,104)	(9,770)	(1,895)	(975)	(2,742)	(5,825)
Female	24.2%	15.1%	2.3%	1.3%	3.3%	5.0%
	(22,404)	(14,088)	(2,223)	(1,214)	(3,118)	(4,656)

Validation set: second wave data

Male	4.5%	38.9%	0.4%	<0.1%	0.6%	1.0%
	(11,860)	(101,655)	(1,113)	(188)	(1,660)	(2,710)
Female	4.3%	48.1%	0.5%	<0.1%	0.7%	0.8%
	(11,021)	(125,776)	(1,343)	(134)	(1,842)	(2,028)

Table 7.2: Mean and standard deviation for all considered cell counts in each cohort. N=1 138 728 CBCs.

Analyte	Covid-19 (+)	Covid-19 (-)	Influenza (+)	Other Viruses (+)	Entire Data
Male patients					
RBC (10^{12}/L)	5.06 ± 0.52	4.21 ± 0.98	4.73 ± 0.60	3.67 ± 0.87	4.28 ± 0.96
Hemoglobin (g/dl)	14.9 ± 1.4	12.4 ± 2.8	14.0 ± 1.7	10.8 ± 2.5	12.6 ± 2.7
Hematocrit (%)	43.8 ± 4.0	36.8 ± 7.9	41.0 ± 4.9	31.7 ± 7.3	37.4 ± 7.7
MCV (fL)	86.8 ± 4.7	88.1 ± 6.4	87.0 ± 6.7	86.9 ± 8.0	88.0 ± 6.2
MCH (pg/cell)	29.5 ± 1.9	29.6 ± 2.3	29.6 ± 2.3	29.6 ± 2.6	29.5 ± 2.2
MCHC (g/dL)	34.1 ± 1.1	33.6 ± 1.4	34.0 ± 1.1	34.1 ± 1.4	33.6 ± 1.4
RDW (%)	13.0 ± 1.0	14.3 ± 2.2	13.6 ± 1.2	15.1 ± 2.1	14.1 ± 2.2
WBC (10^9/L)	6.07 ± 2.37	8.07 ± 3.81	6.96 ± 2.81	5.87 ± 4.69	8.02 ± 3.81
Monocytes (10^9L)	0.66 ± 0.29	0.68 ± 0.35	0.75 ± 0.37	0.66 ± 0.46	0.66 ± 0.34
Lymphocytes (10^9L)	1.40 ± 0.72	1.67 ± 1.05	1.23 ± 0.92	1.25 ± 1.40	1.54 ± 0.99
Eosinophils (10^9/L)	0.07 ± 0.09	0.18 ± 0.20	0.07 ± 0.10	0.10 ± 0.16	0.15 ± 0.20
Basophils (10^9/L)	0.02 ± 0.02	0.03 ± 0.02	0.02 ± 0.01	0.02 ± 0.02	0.03 ± 0.02
Neutrophils (10^9/L)	3.92 ± 2.22	5.53 ± 3.50	4.90 ± 2.57	4.08 ± 3.93	5.64 ± 3.57
Platelets (10^9/L)	195.7 ± 56.7	222.0 ± 102.3	182.9 ± 63.6	145.8 ± 115.6	222.7 ± 99.9
Female patients					
RBC (10^{12}/L)	4.57 ± 0.44	4.03 ± 0.75	4.62 ± 0.67	3.75 ± 0.78	4.05 ± 0.75
Hemoglobin (g/dl)	13.3 ± 1.2	11.8 ± 2.1	13.6 ± 1.9	11.0 ± 2.1	11.8 ± 2.1
Hematocrit (%)	39.8 ± 3.4	35.4 ± 6.1	40.3 ± 5.4	32.8 ± 6.4	35.6 ± 6.0
MCV (fL)	87.3 ± 5.0	88.3 ± 6.3	87.7 ± 6.7	87.9 ± 8.1	88.3 ± 6.2
MCH (pg/cell)	29.2 ± 2.0	29.3 ± 2.3	29.7 ± 2.3	29.4 ± 2.7	29.3 ± 2.2
MCHC (g/dL)	33.5 ± 1.0	33.2 ± 1.3	33.8 ± 1.2	33.5 ± 1.4	33.2 ± 1.3
RDW (%)	13.1 ± 1.1	14.2 ± 2.1	13.7 ± 1.3	14.9 ± 2.1	14.1 ± 2.1
WBC (10^9/L)	5.87 ± 2.40	8.03 ± 3.71	7.11 ± 3.15	6.62 ± 4.63	7.84 ± 3.66
Monocytes (10^9/L)	0.56 ± 0.24	0.62 ± 0.32	0.70 ± 0.35	0.61 ± 0.43	0.60 ± 0.31
Lymphocytes (10^9/L)	1.54 ± 0.80	1.85 ± 1.05	1.36 ± 0.95	1.54 ± 1.40	1.78 ± 1.02
Eosinophils (10^9/L)	0.06 ± 0.08	0.16 ± 0.18	0.075 ± 0.11	0.09 ± 0.18	0.15 ± 0.18
Basophils (10^9/L)	0.02 ± 0.01	0.03 ± 0.02	0.01 ± 0.01	0.02 ± 0.02	0.03 ± 0.02
Neutrophils (10^9/L)	3.68 ± 2151.51	5.39 ± 3.36	4.94 ± 2.97	4.56 ± 3.81	5.29 ± 3.34
Platelets (10^9/L)	222.6 ± 63.0	249.2 ± 101.4	185.0 ± 69.1	188.9 ± 123.8	248.4 ± 100.4

Table 7.3: COVID-19 endemic and pandemic simulations. AUROC, Specificity and Sensitivity, and the respective confidence intervals for different COVID-19 prevalence simulations under 95% confidence. N=30 simulations with 20 000 unique patients each.

COVID-19 Prevalence	AUROC	Specificity	Sensitivity
1%	0.928 ± 0.093	0.875 ± 0.018	0.913 ± 0.152
2%	0.881 ± 0.117	0.877 ± 0.024	0.812 ± 0.250
3%	0.917 ± 0.046	0.874 ± 0.016	0.873 ± 0.099
4%	0.922 ± 0.037	0.882 ± 0.033	0.896 ± 0.087
5%	0.918 ± 0.046	0.874 ± 0.012	0.879 ± 0.104
6%	0.909 ± 0.041	0.874 ± 0.032	0.857 ± 0.116
7%	0.910 ± 0.024	0.883 ± 0.018	0.840 ± 0.083
8%	0.904 ± 0.054	0.879 ± 0.036	0.849 ± 0.102
9%	0.907 ± 0.046	0.872 ± 0.025	0.871 ± 0.085
10%	0.896 ± 0.059	0.871 ± 0.025	0.848 ± 0.118
20%	0.916 ± 0.029	0.866 ± 0.025	0.878 ± 0.045
30%	0.906 ± 0.021	0.871 ± 0.018	0.862 ± 0.059
40%	0.911 ± 0.016	0.871 ± 0.024	0.873 ± 0.032
50%	0.913 ± 0.032	0.886 ± 0.030	0.863 ± 0.028
60%	0.901 ± 0.015	0.868 ± 0.031	0.852 ± 0.037
70%	0.906 ± 0.021	0.867 ± 0.033	0.858 ± 0.035
80%	0.902 ± 0.040	0.869 ± 0.074	0.854 ± 0.025
90%	0.911 ± 0.030	0.889 ± 0.081	0.864 ± 0.022

Chapter 8

Conclusions

In this thesis, we studied an underexplored link between explanatory modeling and Rashamon sets, leading to a novel approach for ensemble learning. In this chapter, we briefly summarize the results already obtained and also present some directions for future work.

8.1 Main Results

We proposed a novel approach to estomate the risk of employing a model in production and used its finding to propose a choice of constituents for ensemble learning based on explainability. We do show, however, that there are some constraints in the choice of this model. First, we must establish a Rashomon subset around the target model. That is, the set of models that satisfy the same predictive accuracy criteria equally, but process information in the data in substantially different ways. From this, we can induce perturbation on our held-out test set to simulate out-of-distribution data and obtain a model that diverges under this scenario, while also estimating the ensemble loss of predictive power as its constituents diverge. After selecting said models, we can estimate at deployment time the reliability of predictions by constituents output distance. The further appart are each individual's predictions, the higher our approximated risk.

A solution is therefore to collect several models and employ them in production, but this selection process is non-trivial. In many situations, the data is inherently composed of several local structures and sub-populations. This work aimed to show, based on evidence, that in these situations it is advantageous to exploit the concept of local structures for the induction of models that are more robust and consistent with the data. We argue that each local structure can be mapped to a context domain and, by harnessing model explanation techniques, we can single out different underlying explanations of the studied phenomenon.

Our proposed approach is grounded in three core concepts: (i) models that compose the ensemble should be diverse in terms of their explanatory factors, (ii) the higher the difference between data distributions, the more divergent constituent behavior tends to be, and (iii) candidate models should be organized by seeking stability in the sense that models that perform similar predictions should be also similar in terms of their explanatory factors, which enable clustering of models. We evaluate our ensemble learning approach in many tasks. In problems where one can expect the existence of multiple local structures, our approach presented consistent gains in AUROC in comparison to other tree-based ensembling techniques. In the absence of such structures, our approach proved to be robust enough to retain high performance. When applied to the problem of predicting in a scenario where the generator function may be different than the one seen in training, we observed a direct relationship between accuracy and model agreement. That is, if the constituents agree, then the accuracy is high and we should trust the predictions. On the other hand, if the agreement between models is low, the accuracy is likewise low and the predictions should not be trusted, which is exactly what we aimed to show and highlights the robustness of our approach.

We also verified a key limitation, that being the diversity scarcity of some hypothesis spaces. Out of the three datasets in which our approach was not able to beat the state-of-the-art in Chapter 3, two presented narrow Rashomon sets quantified by the low Rashomon ratio. In most scenarios, a fair share of the sampled performant models presents performance statistically close, which directly translates to explanation diversity. In the aforementioned datasets, less than 0.5% of the sampled models were comparable to the all-in-one model. The Rashomon ratio is a property of both the data and the hypothesis space, serving as a gauge for the simplicity of the learning problem. A small Rashomon ratio implies a harder learning problem in the sense of model and feature selection thus typifying a limitation on the set of decisions a model might make to belong to a Rashomon set. This harms our sub-space division leading to non-representative groups and poor predictive power. Fortunately, we can quickly estimate this ratio with high confidence and a small sample size during the preliminary sampling stage. For instance, less than 10 000 models drawn at random need to be evaluated to admit a maximum error of 0.01 on the estimative of Rashomon ratio with 95% confidence.

Our method presented satisfactory results when applied to the real-world problems analysed is this study, both when formulating the tasks as binary classifications or regressions. In the COVID-19 case study, to the author's knowledge, we employed the largest COVID-19 dataset to date. We found that training machine learning models solely on 2020 data are not enough to guarantee robustness while also reaching high performance in the wake of scenarios with both prevalence and absence of COVID-19 infections, attaining AUROCs of .90+. Similar patterns were found in the energetic consumption problem, with a R^2 of 0.848 and a mean average percentile error below 2.7%, outperforming all other state-of-the-art approaches while also giving relevant understandings of the present and future energy demand.

A special mention lies in the stainless steel case study, in which we asked for inputs from the metallurgical experts after performing selecting base constituents. The main lesson learned is that there are cases where some conclusions found do not fit with realistic scenarios. For instance, some models implied that an oversaturation of Carbon was preferable. However, high concentrations of this element might disqualify a steel plate from categorization as Duplex stainless steel. By design, many patterns are actionable, and experts can freely select explanations to include as ensemble constituents. This is true for all evaluated case studies and proved particularly effective concerning key chemical elements used in the stainless steel production process. After filtering those patterns that do not fit realistic scenarios, the most relevant ones were turned into production rules and employed in the 2019 and 2020 steelmaking process. A reduction of over 50% in the occurrence of heating slivers was reported, showing the potential of this strategy in a real-world problem, validating the proposed framework, and serving as a pivotal argument towards human-centered AI.

8.2 Future Work

We are currently collecting more accessible datasets to strengthen our benchmark evaluation. As of now, we evaluate ten datasets of different sizes involving binary classification problems, all acquired from the UCI machine learning repository [Asuncion and Newman, 2007] and the OpenML database [Bischl et al., 2017]. We believe our method to be robust enough to handle both multi-class and regression tasks. As such, our benchmark suite should encompass datasets that approach such problems. We also only provided preliminary results in the case studies of hemogram-based disease detection and energetic consumption. Both of these are relevant real-world problems that deserve an in-depth analysis concerning deployment.

For instance, it is imperative to assess the impact of our approach on a hospital's daily flow concerning the COVID-19 Rashomon Ensemble, as the adoption of new technology can potentially disrupt existing processes. This should enable us to collect data concerning other relevant analyses. We are currently studying the implementation of the developed algorithm in different Brazilian hospitals using an API framework connected directly to their databases. In these scenarios, we aim to understand how a tool can be introduced into a hospital's existing workflow in the least disruptive way, as well as find out how comfortable health professionals feel when using it. Imperative to the human-centered AI narrative, some validation thesis includes observing health professionals' interactions with an API, changes in procedures, protocols, and decision-making processes, and the benefits of the solution if applied in fast-paced and high-volume contexts. For example, the Rashomon ensembles for COVID-19 diagnostics could be employed to help in prioritizing patient queues and minimize the occurence of COVID-19 infected patients being assigned to the same waiting rooms of non-infected ones.

Further, our search inside each Rashomon set could also be improved. Applying Monte Carlo proved to be less effective than directly employing an informed search algorithm such as $A*$. However, due to the high degree in the action space of relationship graphs, $A*$ often gets stuck in a series of null expansions and barely outperforms a greedy expansion approach. The other evaluated search algorithms, them being greedy and beam search, suffer from similar problems. Further, since we do not know the 'goal' end model or even its desired performance, we need to make assumptions that harm $A*$ search speed. For instance, a reinforcement learning approach similar to Silver et al. [2016] should be able to encounter promising regions of the model space, being a further refinement of our improved $A*$ search. Another option is directly applying feature selection approaches to the feature sets $Xc \subseteq X$ for each explanation and modifying the equations to take into account cluster membership. Kissel and Mentch [2021] provides relevant theories regarding model path selection, which bear a close relationship to our graph-based model path search. Further experiments need to be performed but forward stability might be a suitable heuristic metric to use alongside the $A*$ algorithm in lieu of performance.

Finally, another direction for future work lies in exploring ensembles constituting different algorithms. In all our experiments, we constrained the Rashomon sets to include only models subject to the same learning algorithms, such as the hypothesis space of all decision trees with depth up to some value. However, employing different methods enables a final model to capture nuances that are particular to all base methods, such as combining a convolutional neural network and an LSTM to capture both temporal and context-aware patterns [Zuin et al., 2018]. However, our method relies on computing feature importance to compare models. When under the same constraining learning algorithm, a comparison of Shapley values is intuitive but it becomes non-trivial when different algorithms also need to be compared to one another. For example, we do not know if a feature importance of 20% in linear regression is comparable to the same 20% in a deep neural network. The problem arises partly due to the non-linearities and interactions that features suffer under more complex models. Heskes et al. [2020] explored the notion of causality in SHAP attributions to disentangle direct influences of a single feature from indirect influences the feature has in a coalition. This causal theory might enable us to compare different learning algorithms and is a study in progress.

Bibliography

N. Abram, J. Gattuso, A. Prakash, L. Cheng, M. Chidichimo, S. Crate, H. Enomoto, M. Garschagen, N. Gruber, S. Harper, E. Holland, R. Kudela, J. Rice, K. Steffen, and K. von Schuckmann. Framing and context of the report: Supplementary material. In *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*, pages 73–129. Intergovernmental Panel on Climate Change, Switzerland, 2019.

S Aeberhard, D Coomans, and O De Vel. Comparison of classifiers in high dimensional settings. dept math statist, james cook univ, north queensland. Technical report, Australia. Tech Rep, 1992.

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

Izuwa Ahanor, Hugh Medal, and Andrew C Trapp. Diversitree: Computing diverse sets of near-optimal solutions to mixed-integer optimization problems. *arXiv preprint arXiv:2204.03822*, 2022.

ANEEL. Expansão da matriz elétrica brasileira - Setembro/2021. *Agência Nacional de Energia Elétrica*, 2021.

Justice Peter Applegarth. *The Australian Institute for Progress Ltd v The Electoral Commission of Queensland & Ors.* Number No. 2. QSC 174, 2020.

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

Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

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

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

Barbara J Bain, Imelda Bates, and Mike A Laffan. *Dacie and lewis practical haematology e-book*. Elsevier Health Sciences, 2016.

Danielle Cristiane Baldo, Alessandra Dellavance, Maria Lucia Gomes Ferraz, and Luis Eduardo C Andrade. Evolving liver inflammation in biochemically normal individuals with anti-mitochondria antibodies. *Autoimmunity Highlights*, 10(1):1–14, 2019.

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

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

Cesar Endrigo Alves Bardelin. *Os efeitos do racionamento de energia elétrica ocorrido no Brasil em 2001 e 2002 com ênfase no consumo de energia elétrica.* PhD thesis, Universidade de São Paulo, 2004.

Mayara Lisboa Bastos, Gamuchirai Tavaziva, Syed Kunal Abidi, Jonathon R Campbell, Louis-Patrick Haraoui, James C Johnston, Zhiyi Lan, Stephanie Law, Emily MacLean, Anete Trajman, et al. Diagnostic accuracy of serological tests for covid-19: systematic review and meta-analysis. *bmj*, 370, 2020.

Marcos Paulo Belançon. Brazil electricity needs in 2030: trends and challenges. *Renewable Energy Focus*, 36:89–95, 2021.

Daohua Bi, Martin Dix, Simon Marsland, Siobhan O'farrell, Arnold Sullivan, Roger Bodman, Rachel Law, Ian Harman, Jhan Srbinovsky, Harun A Rashid, et al. Configuration and spin-up of access-cm2, the new generation australian community climate and earth system simulator coupled model. *Journal of Southern Hemisphere Earth Systems Science*, 70(1):225–251, 2020.

Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Frank Hutter, Michel Lang, Rafael G Mantovani, Jan N van Rijn, and Joaquin Vanschoren. Openml benchmarking suites and the openml100. *stat*, 1050:11, 2017.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Occam's razor. *Information processing letters*, 24(6):377–380, 1987.

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. 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):511–528, 2004.

Olivier Boucher, Jérôme Servonnat, Anna Lea Albright, Olivier Aumont, Yves Balkanski, Vladislav Bastrikov, Slimane Bekki, Rémy Bonnet, Sandrine Bony, Laurent Bopp, et al. Presentation and evaluation of the ipsl-cm6a-lr climate model. *Journal of Advances in Modeling Earth Systems*, 12(7):e2019MS002010, 2020.

Cynthia M Álvares Brandão, Maria Izabel Chiamolera, Rosa Paula Mello Biscolla, José Viana Lima Junior, Cláudia M Ferrer, Wesley Heleno Prieto, Pedro de Sá Tavares Russo, José de Sá, Carolina dos Santos Lazari, Celso Francisco H Granato, et al. No association between vitamin D status and COVID-19 infection in São Paulo, Brazil. *Archives of Endocrinology and Metabolism*, (AHEAD), 2021.

BBC Brasil. Apagão foi 'acidente' causado pelo mau tempo, diz ministro. [Online] `https://www.bbc.com/portuguese/lg/noticias/2009/11/091111_apagao_rc`, 2009.

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

Leo Breiman. Arcing the edge. Technical report, Technical Report 486, Statistics Department, University of California, 1997.

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

Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001b.

Elizabeth Buechler, Siobhan Powell, Tao Sun, Nicolas Astier, Chad Zanocco, Jose Bolorinos, June Flora, Hilary Boudet, and Ram Rajagopal. Global changes in electricity consumption during covid-19. *Iscience*, 25(1):103568, 2022.

Phuong Bui Thi Mai. Underspecification in deep learning. 2021.

Daniel Burillo, Mikhail V. Chester, Stephanie Pincetl, Eric D. Fournier, and Janet Reyna. Forecasting peak electricity demand for los angeles considering higher air temperatures due to climate change. *Applied Energy*, 236:1–9, 2019.

Federico Cabitza and Andrea Campagner. The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical ai studies. *International Journal of Medical Informatics*, 153:104510, 2021.

Federico Cabitza, Andrea Campagner, Davide Ferrari, Chiara Di Resta, Daniele Ceriotti, Eleonora Sabetta, Alessandra Colombini, Elena De Vecchi, Giuseppe Banfi, Massimo Locatelli, et al. Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests. *Clinical Chemistry and Laboratory Medicine (CCLM)*, 59(2):421–431, 2021.

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

Darlan S Candido, Ingra M Claro, Jaqueline G de Jesus, William M Souza, Filipe RR Moreira, Simon Dellicour, Thomas A Mellan, Louis Du Plessis, Rafael HM Pereira, Flavia CS Sales, et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science*, 369(6508):1255–1260, 2020.

Guocai Chai and Pasi Kangas. Recent developments of advanced austenitic and duplex stainless steels for oil and gas industry. *Energy Materials*, pages 703–709, 2014.

Maria de Lourdes Chauffaille, Irina Y Takihi, Wesley H Prieto, Pedro de Sá Tavares Russo, Alex F Sandes, Aline B Perazzio, Marçal CA Silva, and Matheus V Gonçalves. New reference values for the old erythrocyte sedimentation rate. *International Journal of Laboratory Hematology*, 2021.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Changjian Chen, Jun Yuan, Yafeng Lu, Yang Liu, Hang Su, Songtao Yuan, and Shixia Liu. Oodanalyzer: Interactive analysis of out-of-distribution samples. *IEEE Trans. Vis. Comput. Graph.*, 27(7):3335–3349, 2021.

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

Yi Chen, AH Wang, Bo Yi, KQ Ding, HB Wang, JM Wang, HB Shi, SJ Wang, and GZ Xu. Epidemiological characteristics of infection in covid-19 close contacts in ningbo city. *Zhonghua liu Xing Bing xue za zhi= Zhonghua Liuxingbingxue Zazhi*, 41(5):667–671, 2020.

Lijing Cheng and Jiang Zhu. Benefits of cmip5 multimodel ensemble in reconstructing historical ocean subsurface temperature variations. *Journal of Climate*, 29, 03 2016.

Annalisa Cherchi, Pier Giuseppe Fogli, Tomas Lovato, Daniele Peano, Doroteaciro Iovino, Silvio Gualdi, Simona Masina, Enrico Scoccimarro, Stefano Materia, Alessio Bellucci, et al. Global mean climate and main patterns of variability in the cmcc-cm2 coupled model. *Journal of Advances in Modeling Earth Systems*, 11(1):185–209, 2019.

Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, 7:e623, 2021.

Juan Pablo Conti. The day the samba stopped [power blackouts]. *Engineering & Technology*, 5(4):46–47, 2010.

Camila Ferreira Costa and Mario A. Nascimento. IDA 2016 industrial challenge: Using machine learning for predicting failures. In Henrik Boström, Arno J. Knobbe, Carlos Soares, and Panagiotis Papapetrou, editors, *Advances in Intelligent Data Analysis XV - 15th International Symposium, IDA 2016, Stockholm, Sweden, October 13-15, 2016, Proceedings*, volume 9897 of *Lecture Notes in Computer Science*, pages 381–386, 2016.

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. Underspecification presents challenges for credibility in modern machine learning, 2020.

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

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

Blair Davis, Robert Anderson, and Jan Walls. *Rashomon effects: Kurosawa, Rashomon and their legacies.* Routledge, 2015.

Joseph R. Davis and ASM International. Handbook Committee. *Stainless steels.* Materials Park, Ohio: ASM International, 1994.

Ralph M. Davison and James D. Redmond. Practical guide to using duplex stainless steels. *Materials Performance*, pages 57–62, 1990.

Joilson de Assis Cabral, Luiz Fernando Loureiro Legey, and Maria Viviana de Freitas Cabral. Electricity consumption forecasting in brazil: A spatial econometrics approach. *Energy*, 126:124–131, 2017.

Joilson de Assis Cabral, Maria Viviana de Freitas Cabral, and Amaro Olímpio Pereira Júnior. Elasticity estimation and forecasting: An analysis of residential electricity demand in brazil. *Utilities Policy*, 66:101108, 2020.

Matthew DeCamp and Charlotta Lindvall. Latent bias and the implementation of artificial intelligence in medicine. *Journal of the American Medical Informatics Association*, 27 (12):2020–2023, 2020.

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

B. Deo and R. Boom. *Fundamentals of Steelmaking Metallurgy.* Prentice and Hall, 1993.

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. Orientações sobre diagnóstico, tratamento e isolamento de pacientes com covid-19. *Journal Infection Control*, 9(2):56–75, 2020.

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

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

Michael Dorland. The rashomon effect and communication studies. *Canadian Journal of Communication*, 41(2):245–247, 2016.

Ralf Döscher, Mario Acosta, Andrea Alessandri, Peter Anthoni, Almut Arneth, Thomas Arsouze, Tommi Bergmann, Raffaele Bernadello, Souhail Bousetta, Louis-Philippe Caron, et al. The ec-earth3 earth system model for the climate model intercomparison project 6. *Geoscientific Model Development Discussions*, 1:2021, 2021.

Jean-Marie Dufour. Coefficients of determination. *McGill University*, pages 1–14, 2011.

JP Dunne, LW Horowitz, AJ Adcroft, P Ginoux, IM Held, JG John, JP Krasting, S Malyshev, V Naik, F Paulot, et al. The gfdl earth system model version 4.1 (gfdl-esm 4.1): Overall coupled model description and simulation characteristics. *Journal of Advances in Modeling Earth Systems*, 12(11):e2019MS002015, 2020.

Upol Ehsan and Mark O Riedl. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*, pages 449–466. Springer, 2020.

Francisco G Emmerich. 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):3371, 2021.

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

EPE, ONS, and CCEE. 2^a revisão quadrimestral das projeções da demanda de energia elétrica do sistema interligado nacional 2021-2025. *Estudos da Demanda*, 2021.

EPRI. Resource Adequacy for a Decarbonized Future: A Summary of Existing and Proposed Resource Adequacy Metrics. Technical Report 3002023230, Electric Power Research Institute, April 2022. URL https://www.epri.com/research/products/000000003002023230.

Veronika Eyring, Sandrine Bony, Gerald A Meehl, Catherine A Senior, Bjorn Stevens, Ronald J Stouffer, and Karl E Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016.

FAPESP. FAPESP COVID-19 Data Sharing/BR, 2020. https://repositoriodatasharingfapesp.uspdigital.usp.br.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8): 861–874, 2006.

Ana Ferraz. Economic impact of Brazil's energy crisis will last until 2023. [Online] https://brazilian.report/business/2021/09/19/economic-energy-crisis-2023/, 2021.

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

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

David Foldes, Richard Hinton, Siamak Arami, and Barbara J Bain. Plasmacytoid lymphocytes in sars-cov-2 infection (covid-19). *American Journal of Hematology*, 2020.

George Forman et al. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 3(Mar):1289–1305, 2003.

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

Caroline Freitas. Preços de produtos típicos do verão sobem junto com a temperatura. *A Gazeta*, 2020. URL `https://www.agazeta.com.br/es/economia/precos-de-produtos-tipicos-do-verao-sobem-junto-com-a-temperatura-1020`.

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

Mihir Gada, Zenil Haria, Arnav Mankad, Kaustubh Damania, and Smita Sankhe. Automated feature engineering and hyperparameter optimization for machine learning. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 981–986. IEEE, 2021.

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

Joseph L Gastwirth. The estimation of the lorenz curve and gini index. *The review of economics and statistics*, pages 306–316, 1972.

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

Memória Globo. Apagão no brasil em 2009. [Online] `https://memoriaglobo.globo.com/jornalismo/jornalismo-e-telejornais/jornal-nacional/reportagens-e-entrevistas/noticia/apagao-no-brasil-em-2009.ghtml`, 2022.

Paulo M. Gonçalves, Silas G.T. de Carvalho Santos, Roberto S.M. Barros, and Davi C.L. Vieira. A comparative study on concept drift detectors. *Expert Systems with Applications*, 41(18):8144–8156, 2014.

Henrik Grosskreutz. Cascaded subgroups discovery with an application to regression. In *Proc. ECML/PKDD*, volume 5211, page 33. Citeseer, 2008.

Wei-jie Guan, Zheng-yi Ni, Yu Hu, Wen-hua Liang, Chun-quan Ou, Jian-xing He, Lei Liu, Hong Shan, Chun-liang Lei, David SC Hui, et al. Clinical characteristics of coronavirus disease 2019 in china. *New England journal of medicine*, 382(18):1708–1720, 2020.

MARIA CAROLINA R GULLO. A economia na pandemia covid-19: algumas considerações. *Rosa dos Ventos*, 12(Esp. 3):1–8, 2020.

Robert N. Gunn. *Duplex Stainless Steels, Microstructure, Properties and Applications.* Abington Publishing, 1997.

Isabelle Guyon, Steve Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the nips 2003 feature selection challenge. *Advances in neural information processing systems*, 17:545–552, 2004.

Gerald J Hahn. The hazards of extrapolation in regression analysis. *Journal of Quality Technology*, 9(4):159–165, 1977.

Thomas Hale, Anna Petherick, Toby Phillips, and Samuel Webster. Variation in government responses to covid-19. *Blavatnik School working paper*, 2020.

J Hanley and B. McNeil. The meaning and use of the area under a receiver operating characteristic ROC curve. *Radiology*, 143:29–36, 1982.

Peter Hart, Nils Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968. doi: 10.1109/tssc.1968.300136. URL `https://doi.org/10.1109/tssc.1968.300136`.

CPM Hayward, KA Moffat, TI George, and M Proytcheva. Assembly and evaluation of an inventory of guidelines that are available to support clinical hematology laboratory practice. *International journal of laboratory hematology*, 37:36–45, 2015.

Daihai He, Guihong Fan, Xueying Wang, Yingke Li, and Zhihang Peng. The new SARS-CoV-2 variant and reinfection in the resurgence of COVID-19 outbreaks in Manaus, Brazil. *medRxiv*, 2021.

Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. Deep anomaly detection with outlier exposure. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview, 2019.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146 (730):1999–2049, 2020.

Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems*, 33:4778–4789, 2020.

James Hinns, Xiuyi Fan, Siyuan Liu, Veera Raghava Reddy Kovvuri, Mehmet Orcun Yalcin, and Markus Roggenbach. 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*, pages 323–335. Springer, 2021.

Andreas Holzinger, Peter Kieseberg, Edgar Weippl, and A Min Tjoa. Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable ai. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 1–8. Springer, 2018.

Giles Hooker. *Diagnostics and extrapolation in machine learning.* stanford university, 2004.

Ching-Lai Hor, Simon J Watson, and Shanti Majithia. Analyzing the impact of weather variables on monthly electricity demand. *IEEE transactions on power systems*, 20(4): 2078–2085, 2005.

Ben Hu, Hua Guo, Peng Zhou, and Zheng-Li Shi. Characteristics of sars-cov-2 and covid-19. *Nature Reviews Microbiology*, pages 1–14, 2020.

YJ Huang, R Ritschard, J Bull, and L Chang. Climatic indicators for estimating residential heating and cooling loads. Technical report, Lawrence Berkeley Lab., CA (USA), 1986.

Mallikarjun G Hudedmani, Vishwanath M Soppimath, Santosh V Hubballi, and Darshana Joshi. Dawn after black out: A hope of light–a review. *International Journal of Advance in Science and Technology*, 6(1):1264–1271, 2019.

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. 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*, 91:264–266, 2020.

Julian David Hunt, Andreas Nascimento, Carla Schwengber ten Caten, Fernanda Munari Caputo Tomé, Paulo Smith Schneider, André Luis Ribeiro Thomazoni, Nivalde José de Castro, Roberto Brandão, Marcos Aurélio Vasconcelos de Freitas, José Sidnei Colombo Martini, et al. Energy crisis in brazil: Impact of hydropower reservoir level on the river flow. *Energy*, 239:121927, 2022.

Thaer Idrees, Wesley H Prieto, Sabina Casula, Aswathy Ajith, Matthew Etthelson, Flavia A Andreotti Narchi, Pedro ST Russo, Fernando Fernandes, Julie Johnson, Anoop Mayampurath, et al. Use of statins among patients taking levothyroxine: An observational drug utilization study across sites. *Journal of the Endocrine Society*, 2021.

IEA. Hydropower has a crucial role in accelerating clean energy transitions to achieve countries' climate ambitions securely. [Online] `https://www.iea.org/news/hydropower-has-a-crucial-role-in-accelerating-clean-energy-transitions-to-achiev` 2021.

Anita S Iyer, Forrest K Jones, Ariana Nodoushani, Meagan Kelly, Margaret Becker, Damien Slater, Rachel Mills, Erica Teng, Mohammad Kamruzzaman, Wilfredo F Garcia-Beltran, et al. Persistence and decay of human antibody responses to the receptor binding domain of sars-cov-2 spike protein in covid-19 patients. *Science immunology*, 5(52), 2020.

Ping Jiang, Ranran Li, Ningning Liu, and Yuyang Gao. A novel composite electricity demand forecasting framework by data processing and optimized support vector machine. *Applied Energy*, 260:114243, 2020.

Ambika Kaul, Saket Maheshwary, and Vikram Pudi. Autolearn-automated feature generation and selection. In *2017 IEEE International Conference on data mining (ICDM)*, pages 217–226. IEEE, 2017.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3146–3154. Curran Associates, Inc., 2017a.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3146–3154. Curran Associates, Inc., 2017b.

Michael Kearns. Thoughts on hypothesis boosting. *Unpublished manuscript*, 45:105, 1988.

Tam Kemabonta. Grid resilience analysis and planning of electric power systems: The case of the 2021 texas electricity crises caused by winter storm uri. *The Electricity Journal*, 34(10):107044, 2021.

Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *VLDB*, volume 4, pages 180–191. Toronto, Canada, 2004.

Nicholas Kissel and Lucas Mentch. Forward stability and model path selection. *arXiv preprint arXiv:2103.03462*, 2021.

Arno Knobbe and Joris Valkonet. Building classifiers from pattern teams. In *Proceedings of the ECML PKDD'09 workshop LeGo*, pages 77–93. Citeseer, 2009.

R Krishnan, P Swapna, Ramesh Vellore, Sandeep Narayanasetti, AG Prajeesh, Ayantika Dey Choudhury, Manmeet Singh, TP Sabin, and J Sanjay. The iitm earth system model (esm): development and future roadmap. In *Current trends in the Representation of physical processes in weather and climate models*, pages 183–195. Springer, 2019.

Manish Kumar, Deepak Kumar Gupta, and Samayveer Singh. Extreme event forecasting using machine learning models. In *Advances in Communication and Computational Technology*, pages 1503–1514. Springer, 2021.

Akira Kurosawa. Rashomon [film]. *Producer: Daiei, Japan. Script: T. Matsuama*, 1950.

Miron B. Kursa and Witold R. Rudnicki. Feature selection with the Boruta package. *Journal of Statistical Software*, 36(11):1–13, 2010. URL `http://www.jstatsoft.org/v36/i11/`.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pages 2124–2132. AAAI Press, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6405–6416, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

Tim Leiner, Daniel Rueckert, Avan Suinesiaputra, Bettina Baeßler, Reza Nezafat, Ivana Išgum, and Alistair A Young. Machine learning in cardiovascular magnetic resonance: basic concepts and applications. *Journal of Cardiovascular Magnetic Resonance*, 21(1): 1–14, 2019.

Sebastian Lerch, Thordis L Thorarinsdottir, Francesco Ravazzolo, and Tilmann Gneiting. Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science*, pages 106–127, 2017.

Lijuan Li, Yongqiang Yu, Yanli Tang, Pengfei Lin, Jinbo Xie, Mirong Song, Li Dong, Tianjun Zhou, Li Liu, Lu Wang, et al. The flexible global ocean-atmosphere-land system model grid-point version 3 (fgoals-g3): description and evaluation. *Journal of Advances in Modeling Earth Systems*, 12(9):e2019MS002012, 2020.

Eunjung Lim, Jill Miyamura, and John J Chen. Racial/ethnic-specific reference intervals for common laboratory tests: a comparison among asians, blacks, hispanics, and white. *Hawai'i Journal of Medicine & Public Health*, 74(9):302, 2015.

Hui Liu, Jing Chen, Qin Yang, Fang Lei, Changjiang Zhang, Juan-Juan Qin, Ze Chen, Lihua Zhu, Xiaohui Song, Liangjie Bai, Xuewei Huang, Weifang Liu, Feng Zhou, Ming-Ming Chen, Yan-Ci Zhao, Xiao-Jing Zhang, Zhi-Gang She, Qingbo Xu, Xinliang Ma, Peng Zhang, Yan-Xiao Ji, Xin Zhang, Juan Yang, Jing Xie, Ping Ye, Elena Azzolini, Alessio Aghemo, Michele Ciccarelli, Gianluigi Condorelli, Giulio G. Stefanini, Jiahong Xia, Bing-Hong Zhang, Yufeng Yuan, Xiang Wei, Yibin Wang, Jingjing Cai, and Hongliang Li. Development and validation of a risk score using complete blood count to predict in-hospital mortality in COVID-19 patients. *Med*, 2021.

Si-Ming Liu, Yuan-Hao Chen, Jian Rao, Can Cao, Si-Yu Li, Mu-Han Ma, and Yao-Bin Wang. Parallel comparison of major sudden stratospheric warming events in cesm1-waccm and cesm2-waccm. *Atmosphere*, 10(11):679, 2019.

Yongyan Liu. Analysis of brazilian blackout on march 21st, 2018 and revelations to security for hunan grid. In *2019 4th International Conference on Intelligent Green Building and Smart Grid (IGBSG)*, pages 1–5. IEEE, 2019.

T Lovato, D Peano, M Butenschön, S Materia, D Iovino, E Scoccimarro, PG Fogli, A Cherchi, A Bellucci, S Gualdi, et al. Cmip6 simulations with the cmcc earth system model (cmcc-esm2). *Journal of Advances in Modeling Earth Systems*, 14(3):e2021MS002814, 2022.

Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363, 2019. doi: 10.1109/TKDE.2018.2876857.

Ning Lu, Guangquan Zhang, and Jie Lu. Concept drift detection via competence models. *Artificial Intelligence*, 209:11–28, 2014.

Scott Lundberg and Su-In Lee. An unexpected unity among methods for interpreting model predictions. *CoRR*, abs/1611.07478, 2016.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Annual Conf. on Neural Information Processing Systems*, pages 4768–4777, 2017.

Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760, 2018.

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):2522–5839, 2020.

David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

David Madras, James Atwood, and Alexander D'Amour. Detecting extrapolation with local ensembles. In *International Conference on Learning Representations*, 2020.

Debora Maia-Silva, Rohini Kumar, and Roshanak Nateghi. The critical role of humidity in modeling summer electricity demand across the united states. *Nature communications*, 11(1):1–8, 2020.

Hassan H Malik and John R Kender. Classification by pattern-based hierarchical clustering. In *From Local Patterns to Global Models Workshop, ECML/PKDD*, pages 1–18, 2008.

Charles T. Marx, Flavio Du Pin Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Medicaid Medicare. Clia programs; regulations implementing the clinical laboratory improvement amendments of 1988 (clia)-hcfa. final rule with comment period. *Fed Regist*, 57(40):7002–7186, 1992.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 2019.

Hanna Meyer and Edzer Pebesma. Predicting into unknown space? estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 12(9): 1620–1633, 2021.

W Greg Miller and Neil Greenberg. Harmonization and standardization: Where are we now? *The Journal of Applied Laboratory Medicine*, 6(2):510–521, 2021.

Leandro L Minku and Xin Yao. Ddd: A new ensemble approach for dealing with concept drift. *IEEE transactions on knowledge and data engineering*, 24(4):619–633, 2011.

Amin Moazami, Vahid M Nik, Salvatore Carlucci, and Stig Geving. Impacts of future weather data typology on building energy performance–investigating long-term patterns of climate change and extreme weather conditions. *Applied Energy*, 238:696–720, 2019.

Emilio F Moran, Maria Claudia Lopez, Nathan Moore, Norbert Müller, and David W Hyndman. Sustainable hydropower in the 21st century. *Proceedings of the National Academy of Sciences*, 115(47):11891–11898, 2018.

Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012.

Wolfgang A Müller, Johann H Jungclaus, Thorsten Mauritsen, Johanna Baehr, Matthias Bittner, R Budich, Felix Bunzel, Monika Esch, Rohit Ghosh, Helmut Haak, et al. A higher-resolution version of the max planck institute earth system model (mpi-esm1. 2-hr). *Journal of Advances in Modeling Earth Systems*, 10(7):1383–1413, 2018.

Fionn Murtagh. Multilayer perceptrons for classification and regression. *Neurocomputing*, 2(5-6):183–197, 1991.

Ahmet Nalbant, Tezcan Kaya, Ceyhun Varim, Selçuk Yaylaci, Ali Tamer, and Hakan Cinemre. Can the neutrophil/lymphocyte ratio (nlr) have a role in the diagnosis of coronavirus 2019 disease (COVID-19)? *Revista da Associação Médica Brasileira*, 66 (6):746–751, 2020.

N Nanava, M Betaneli, G Giorgobiani, T Chikovani, and N Janikashvili. Complete blood count derived inflammatory biomarkers in patients with hematologic malignancies. *Georgian Med News*, 302:39–44, 2020.

Felipe Naveca, Cristiano da Costa, Valdinete Nascimento, Victor Souza, André Corado, Fernanda Nascimento, Ágatha Costa, Débora Duarte, George Silva, Matilde Mejía, et al. Sars-cov-2 reinfection by the new variant of concern (voc) p. 1 in amazonas, brazil. *virological. org*, 2021.

Nirbhar Neogi, Dusmanta Mohanta, and Pranab Dutta. Review of vision-based steel surface inspection systems. *J Image Video Proc*, page 50, 2014.

Yilin Ning, Marcus Eng Hock Ong, Bibhas Chakraborty, Benjamin Alan Goldstein, Daniel Shu Wei Ting, Roger Vaughan, and Nan Liu. Shapley variable importance cloud for interpretable machine learning. *Patterns*, 3(4):100452, 2022.

Manuel Olave, Vladislav Rajkovic, and Marko Bohanec. An application for admission in public school systems. *Expert Systems in Public Administration*, 1:145–160, 1989.

B. C. O'Neill, C. Tebaldi, D. P. van Vuuren, V. Eyring, P. Friedlingstein, G. Hurtt, R. Knutti, E. Kriegler, J.-F. Lamarque, J. Lowe, G. A. Meehl, R. Moss, K. Riahi, and B. M. Sanderson. The scenario model intercomparison project (scenariomip) for cmip6. *Geoscientific Model Development*, 9(9):3461–3482, 2016. doi: 10.5194/gmd-9-3461-2016. URL `https://gmd.copernicus.org/articles/9/3461/2016/`.

ONS. Histórico da operação instalada, carga de energia. *Operador Nacional do Sistema Elétrico, Rio de Janeiro*, 2018.

Anton Orlov, Jana Sillmann, and Ilaria Vigo. Better seasonal forecasts for the renewable energy industry. *Nature Energy*, 5(2):108–110, 2020.

Guillermo Ortiz-Jiménez, Itamar Franco Salazar-Reque, Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. A neural anisotropic view of underspecification in deep learning. *CoRR*, abs/2104.14372, 2021.

Adam Palayew, Ole Norgaard, Kelly Safreed-Harmon, Tue Helms Andersen, Lauge Neimann Rasmussen, and Jeffrey V Lazarus. Pandemic publishing poses a new covid-19 challenge. *Nature Human Behaviour*, 4(7):666–669, 2020.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011a.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011b.

Tiago Pimentel, Adriano A. Veloso, and Nivio Ziviani. Fast node embeddings: Learning ego-centric representations. In *6th International Conference on Learning Representations*, 2018.

Stephen Prince. *The warrior's camera: the cinema of Akira Kurosawa*. Princeton University Press, 1999.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 6638–6648. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7898-catboost-unbiased-boosting-with-categorical-features.pdf.

Pushpalata Pujari and Jyoti Bala Gupta. Improving classification accuracy by using feature selection and ensemble model. *International Journal of Soft Computing and Engineering*, 2(2):380–386, 2012.

J. R. Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, March 1986.

Sergio Ramírez-Gallego, Bartosz Krawczyk, Salvador García, Michał Woźniak, and Francisco Herrera. A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing*, 239:39–57, 2017.

D Raj Reddy et al. Speech understanding systems: A summary of results of the five-year research effort. *Department of Computer Science. Camegie-Mell University, Pittsburgh, PA*, 17:138, 1977.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Intl Conf. on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI Conf. on Artificial Intelligence*, pages 1527–1535, 2018.

Carmen Ricós, Virtudes Álvarez, Carmen Perich, Pilar Fernández-Calle, Joana Minchinela, Fernando Cava, Carmen Biosca, Beatriz Boned, Mariví Doménech, José Vicente García-Lario, et al. Rationale for using data on biological variation. *Clinical Chemistry and Laboratory Medicine (CCLM)*, 53(6):863–870, 2015.

DR Ring, MK Harris, JA Jackman, and JL Henson. A fortran computer program for determining start date and base temperature for degree day models. Technical report, Texas Agricultural Experiment Station, the Texas A and M University System, 1983.

Hannah Ritchie, Edouard Mathieu, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Esteban Ortiz-Ospina, Joe Hasell, Bobbie Macdonald, Diana Beltekian, and Max Roser. Coronavirus pandemic (covid-19). *Our World in Data*, 2020.

Guangchun Ruan, Dongqi Wu, Xiangtian Zheng, Haiwang Zhong, Chongqing Kang, Munther A. Dahleh, S. Sivaranjani, and Le Xie. A cross-domain approach to analyzing the short-run impact of covid-19 on the us electricity sector. *Joule*, 4(11):2322–2337, 2020.

Ando Saabas. Interpreting random forests. *Diving into data*, 24, 2014.

Tetsuya Sakai. Statistical reform in information retrieval? *SIGIR Forum*, 48(1):3–12, 2014.

Katia Corrêa de Oliveira Santos, Daniela Bernardes Borges da Silva, Norio Augusto Sasaki, Margarete Aparecida Benega, Rebecca Garten, and Terezinha Maria de Paiva. Molecular epidemiology of influenza a (h1n1) pdm09 hemagglutinin gene circulating in sao paulo state, brazil: 2016 anticipated influenza season. *Revista do Instituto de Medicina Tropical de São Paulo*, 59, 2017.

Robert E Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

Lesia Semenova and Cynthia Rudin. A study in rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. *CoRR*, abs/1908.01755, 2019.

Tido Semmler, Sergey Danilov, Paul Gierz, Helge F Goessling, Jan Hegewald, Claudia Hinrichs, Nikolay Koldunov, Narges Khosravi, Longjiang Mu, Thomas Rackow, et al. Simulations for cmip6 with the awi climate model awi-cm-1-1. *Journal of Advances in Modeling Earth Systems*, 12(9):e2019MS002009, 2020.

Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2 (28):307–317, 1953.

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

Ben Shneiderman. Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4):1–31, 2020.

Daniel Silveira. Falha humana provocou apagão nas regiões norte e nordeste em março, diz ons. [Online] `https://g1.globo.com/economia/noticia/falha-humana-provocou-apagao-no-norte-e-nordeste-diz-ons.ghtml`, 2018.

Elena Caires Silveira. Prediction of COVID-19 from hemogram results and age using machine learning. *Frontiers in Health Informatics*, 9(1):39, 2020.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Anika Singanayagam, Monika Patel, Andre Charlett, Jamie Lopez Bernal, Vanessa Saliba, Joanna Ellis, Shamez Ladhani, Maria Zambon, and Robin Gopal. Duration of infectiousness and correlation with rt-pcr cycle threshold values in cases of covid-19, england, january to may 2020. *Eurosurveillance*, 25(32):2001483, 2020.

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. 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.

Robert G Steadman. The assessment of sultriness. part i: A temperature-humidity index based on human physiology and clothing science. *Journal of Applied Meteorology and Climatology*, 18(7):861–873, 1979.

Robert G. Steadman. A universal scale of apparent temperature. *Journal of Applied Meteorology and Climatology*, 23(12):1674 – 1687, 1984.

Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.

J. Stradomska, Z. Stradomski, and M. Soinski. The analysis of solidification process of ferritic-austenitic cast steel. *Archives of Foundry Engineering*, 9(1):83–86, 2009.

Alan Strahler and Arthur Strahler. *Physical geography.* John Wiley & Sons, 2007.

W Nick Street, William H Wolberg, and Olvi L Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization*, volume 1905, pages 861–870. International Society for Optics and Photonics, 1993.

Roland B Stull, C Donald Ahrens, et al. *Meteorology for scientists and engineers.* Brooks/Cole, 2000.

Neven Sumonja, Branislava Gemovic, Nevena Veljkovic, and Vladimir Perovic. Automated feature engineering improves prediction of protein–protein interactions. *Amino Acids*, 51(8):1187–1200, 2019.

Julia Swart and Lisa Brinkmann. *Economic Complexity and the Environment: Evidence from Brazil*, pages 3–45. Springer International Publishing, Cham, 2020. ISBN 978-3-030-30306-8. doi: 10.1007/978-3-030-30306-8_1. URL `https://doi.org/10.1007/978-3-030-30306-8_1`.

Neil C Swart, Jason NS Cole, Viatcheslav V Kharin, Mike Lazare, John F Scinocca, Nathan P Gillett, James Anstey, Vivek Arora, James R Christian, Sarah Hanna, et al. The canadian earth system model version 5 (canesm5. 0.3). *Geoscientific Model Development*, 12(11):4823–4873, 2019.

Li Tan, Qi Wang, Duanyang Zhang, Jinya Ding, Qianchuan Huang, Yi-Quan Tang, Qiongshu Wang, and Hongming Miao. Lymphopenia predicts disease severity of COVID-19: a descriptive and predictive study. *Signal transduction and targeted therapy*, 5(1):1–3, 2020.

Hiroaki Tatebe, Tomoo Ogura, Tomoko Nitta, Yoshiki Komuro, Koji Ogochi, Toshihiko Takemura, Kengo Sudo, Miho Sekiguchi, Manabu Abe, Fuyuki Saito, et al. Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in miroc6. *Geoscientific Model Development*, 12(7):2727–2765, 2019.

B. Thomas. Modeling of continuous-casting defects related to mold fluid flow. *Iron and Steel Technology (AIST Transactions)*, 3(7):128–143, 2006.

Jerry F Tjiputra, Jörg Schwinger, Mats Bentsen, Anne L Morée, Shuang Gao, Ingo Bethke, Christoph Heinze, Nadine Goris, Alok Gupta, Yan-Chun He, et al. Ocean biogeochemistry in the norwegian earth system model version 2 (noresm2). *Geoscientific Model Development*, 13(5):2393–2431, 2020.

Ian M Trotter, Torjus Folsland Bolkesjø, José Gustavo Féres, and Lavinia Hollanda. Climate change and electricity demand in brazil: A stochastic approach. *Energy*, 102: 596–604, 2016.

Faik Tursun, Mahmut E Cebeci, Osman B Tör, Aydın Sahin, Hacer G Taşkın, and A Nezih Güven. Determination of zonal power demand s-curves with ga based on top-to-bottom and end-use approaches. In *2016 4th International Istanbul Smart Grid Congress and Fair (ICSG)*, pages 1–5. IEEE, 2016.

Jussi Tuunanen et al. Modelling of changes in electricity end-use and their impacts on electricity distribution. 2015.

Frédéric Vitart and Andrew W Robertson. The sub-seasonal to seasonal prediction project (s2s) and the prediction of extreme events. *npj Climate and Atmospheric Science*, 1(1): 1–7, 2018.

E Volodin. The mechanisms of cloudiness evolution responsible for equilibrium climate sensitivity in climate model inm-cm4-8. *Geophysical Research Letters*, 48(24): e2021GL096204, 2021.

Evgeny Volodin. The mechanism of 60-year and 15-year arctic climate oscillations in climate model inm-cm5-0. In *EGU General Assembly Conference Abstracts*, page 7265, 2020.

Mark C Walters and Herbert T Abelson. Interpretation of the complete blood count. *Pediatric Clinics*, 43(3):599–622, 1996.

Bin Wang, Lihong Zheng, De Li Liu, Fei Ji, Anthony Clark, and Qiang Yu. Using multi-model ensembles of cmip5 global climate models to reproduce observed monthly rainfall and temperature with machine learning methods in australia. *International Journal of Climatology*, 38(13):4891–4902, 2018.

Dawei Wang, Bo Hu, Chang Hu, Fangfang Zhu, Xing Liu, Jing Zhang, Binbin Wang, Hui Xiang, Zhenshun Cheng, Yong Xiong, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in wuhan, china. *Jama*, 323 (11):1061–1069, 2020.

Zhe Wang, Tianzhen Hong, Han Li, and Mary Ann Piette. Predicting city-scale daily electricity consumption using data-driven models. *Advances in Applied Energy*, 2:100025, 2021.

Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. " do you trust me?" increasing user-trust by integrating virtual agents in explainable ai interaction design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 7–9, 2019.

Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

Jason Woods, Nelson James, Eric Kozubal, Eric Bonnema, Kristin Brief, Liz Voeller, and Jessy Rivest. Humidity's impact on greenhouse gas emissions from air conditioning. *Joule*, 6(4):726–741, 2022.

Joseph T Wu, Kathy Leung, and Gabriel M Leung. 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):689–697, 2020.

Tongwen Wu, Yixiong Lu, Yongjie Fang, Xiaoge Xin, Laurent Li, Weiping Li, Weihua Jie, Jie Zhang, Yiming Liu, Li Zhang, et al. The beijing climate center climate system model (bcc-csm): the main progress from cmip5 to cmip6. *Geoscientific Model Development*, 12(4):1573–1600, 2019.

Zunyou Wu and Jennifer M McGoogan. Characteristics of and important lessons from the coronavirus disease 2019 (covid-19) outbreak in china: summary of a report of 72 314 cases from the chinese center for disease control and prevention. *Jama*, 323(13): 1239–1242, 2020.

Laure Wynants, Ben Van Calster, Gary S Collins, Richard D Riley, Georg Heinze, Ewoud Schuit, Marc MJ Bonten, Darren L Dahly, Johanna AA Damen, Thomas PA Debray, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj*, 369, 2020.

Ai Tang Xiao, Yi Xin Tong, and Sheng Zhang. False-negative of rt-pcr and prolonged nucleic acid conversion in covid-19: rather than recurrence. *Journal of medical virology*, 2020.

Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the whole rashomon set of sparse decision trees. *arXiv preprint arXiv:2209.08040*, 2022.

Wei Xu. Toward human-centered ai: a perspective from human-computer interaction. *Interactions*, 26(4):42–46, 2019.

Senshan Yang, Joanne Logan, and David L Coffey. Mathematical formulae for calculating the base temperature for growing degree days. *Agricultural and Forest Meteorology*, 74 (1-2):61–74, 1995.

Seiji Yukimoto, Hideaki Kawai, Tsuyoshi Koshiro, Naga Oshima, Kohei Yoshida, Shogo Urakawa, Hiroyuki Tsujino, Makoto Deushi, Taichu Tanaka, Masahiro Hosaka, et al. The meteorological research institute earth system model version 2.0, mri-esm2. 0: Description and basic evaluation of the physical component. *Journal of the Meteorological Society of Japan. Ser. II*, 2019.

Yong Jie Zhao, Yun Hui Yan, and Ke Chen Song. Vision-based automatic detection of steel surface defects in the cold rolling process: considering the influence of industrial liquids and surface textures. *The International Journal of Advanced Manufacturing Technology*, 90(5–8):1665, 2017.

Florian Ziel. Modeling public holidays in load forecasting: a german case study. *Journal of Modern Power Systems and Clean Energy*, 6(2):191–207, 2018.

G. Zuin and A. Veloso. Learning a resource scale for collectible card games. In *2019 IEEE Conference on Games (CoG)*, pages 1–8, 2019. doi: 10.1109/CIG.2019.8847946.

G. Zuin, L. Chaimowicz, and A. Veloso. Deep learning techniques for explainable resource scales in collectible card games. *IEEE Transactions on Games*, pages 1–1, 2020. doi: 10.1109/TG.2020.3030742.

Gianluca Zuin, Adriano Veloso, João Cândido Portinari, and Nivio Ziviani. Automatic tag recommendation for painting artworks using diachronic descriptions. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

Gianlucca Zuin, Luiz Chaimowicz, and Adriano Veloso. Learning transferable features for open-domain question answering. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2018.

Gianlucca Zuin, Felipe Marcelino, Lucas Borges, João Couto, Victor Jorge, Mychell Laurindo, Glaucio Barcelos, Marcio Cunha, Valdeci Alvarenga, Henrique Rodrigues, et al. Predicting heating sliver in duplex stainless steels manufacturing through rashomon sets. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

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. Prediction of sars-cov-2-positivity from million-scale complete blood counts using machine learning. *Communications medicine*, 2(1):1–12, 2022a.

Gianlucca Zuin, Rob Buechler, Tao Sun, Chad Zanocco, Daniella Castro, Adriano Veloso, and Ram Rajagopal. Revealing the impact of extreme events on electricity consumption in brazil: A data-driven counterfactual approach. In *2022 IEEE Power & Energy Society General Meeting (PESGM)*, pages 1–5, 2022b.