Improving models performance in a data-centric approach applied to the healthcare domain

M. G. Valeriano[1][2], C. R. V. Kiffer[1], A. C. Lorena[2]

[1] Universidade Federal de São Paulo, Brazil
carlos.kiffer@unifesp.br
[2] Instituto Tecnológico de Aeronáutica
{valeriano, aclorena}@ita.br

Abstract. Machine learning systems heavily rely on training data, and any biases or limitations in datasets can significantly impair the performance and trustworthiness of these models. This paper proposes an instance hardness data-centric approach to enhance ML systems, leveraging the potential of contrasting the profiles of groups of easy and hard instances on a dataset to design classification problems more effectively. We present a case study with a COVID dataset sourced from a public repository that was utilized to predict aggravated conditions based on parameters collected on the patient's initial attendance. Our goal was to investigate the impact of different dataset design choices on the performance of the ML models. By adopting the concept of instance hardness, we identified instances that were consistently misclassified or correctly classified, forming distinct groups of hard and easy instances for further investigation. Analyzing the relationship between the original class, instance hardness level, and the information contained in the raw data source, we gained valuable insights into how changes in data assemblage can improve the performance of the ML models. Although the characteristics of the problem condition our analysis, the findings demonstrate the significant potential of a data-centric perspective in enhancing predictive models within the healthcare domain.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**.

Keywords: instance hardness, machine learning, healthcare

1. INTRODUCTION

Machine learning (ML) fundamentally involves extracting patterns from data, assuming that models will generalize to future instances. This process can be divided into two stages: first, the selection of algorithms and their adjustment to a dataset; second, revisiting the data to attempt to build better algorithms [Zha et al. 2023]. These stages describe two complementary approaches when developing ML models: model-centric and data-centric research. Although both stages are important, research on models is much more advanced than research on data. The role of data in ML is often neglected and undervalued, with models being the focal points of publications and conferences, becoming increasingly evident that research on data has been underdeveloped [Sambasivan et al. 2021]. However, the field of ML is undergoing a profound transformation. While the past focused on pursuing innovative algorithms and architectures, the present and future are increasingly centered on data.

As large models become the norm and real-world efficacy becomes paramount, the emphasis is shifting towards the entire data lifecycle, from collection and storage to transformation and integration of results into other systems. The importance of addressing societal issues through data further underscores this shift. In particular, there is growing recognition of the value of small datasets and the need for meticulous data curation and preprocessing to ensure the quality and representativeness of the data used in ML models [Oala et al. 2023]. The central role of data in advancing AI research is now widely recognized. For example, last year, a conference dedicated to data-centric ML research established a new journal in this field [Oala et al. 2023]. This shift highlights the disparity between the vast array of tools and strategies developed for model training - such as those popularized by scikit-learn [Pedregosa et al. 2011], and strategies for hyperparameter tuning [Bergstra et al. 2015] - and the still largely manual, time-consuming process of building robust datasets [Seedat et al. 2022]. The reliance on trial and error in data preparation is the norm in these cases with the risk of neglecting the quality and safety of the data underlying models [Sambasivan et al. 2021].

In this work, our goal is to contribute to data-centric ML research by providing a strategy to quantify the benefits of different decisions made during data preprocessing. We propose evaluating the *Instance Hardness* (IH), a metric proposed by [Smith et al. 2014] that assesses the difficulty level in correctly classifying individual instances within a dataset. Instances that are consistently misclassified by multiple ML techniques are deemed hard, while those consistently classified correctly are considered easy. Based on the average IH value for each class or a specific subgroup, we argue that it is possible to justify whether a particular preprocessing decision improves data quality and increases class separability. To support our proposal, we present a case study demonstrating how this approach can benefit the process of assembling datasets in healthcare.

By analyzing the IH values, we can evaluate problem design decisions, enhancing the resulting ML models' overall robustness and accuracy. Our approach emphasizes the importance of meticulous data preparation, highlighting that even minor adjustments in preprocessing can significantly impact model performance. This work offers a practical methodology for assessing data quality and contributes to a broader understanding of how data-centric practices can lead to more reliable and fair ML outcomes. As the field evolves, our findings underscore the necessity of integrating data-centric perspectives into ML development and deployment, ensuring that models are not only technically sound but also ethically and socially responsible.

2. MEASURING THE DIFFICULTY OF A PROBLEM

Each dataset has different characteristics and levels of difficulty. This might happen for various reasons and can be analyzed from different perspectives. For example, overlapping regions in the feature space can make it difficult to separate classes [Hüllermeier and Waegeman 2021]. Another reason is the presence of outliers that may interfere in delineating a precise decision boundary among classes [Napierala and Stefanowski 2016]. In this way, how can we measure data adequacy for a classification task? This concern has led to the recent focus on Instance Hardness (IH) analysis [Smith et al. 2014; Liu et al. 2024; Paiva et al. 2022; Lorena et al. 2024; Seedat et al. 2024], by identifying which instances are systematically misclassified by ML models.

This approach was explored by [Chatzimparmpas et al. 2022] in selecting instances for oversampling when dealing with imbalanced classification problems. [Smith et al. 2014] and [Seedat et al. 2022] investigated the impact of removing difficult instances, noting an increase in predictive performance, although this improvement is not always consistent. [Seedat et al. 2024] examined the correlation between the difficulty level and the underlying mechanism that generates hardness, noting that it is not always possible to clearly identify the type of generation mechanism.

In previous work, we explored how the performance of models on difficult instances could serve as an explainability strategy, facilitating collaborative work between domain experts and data scientists [Valeriano et al. 2024; Valeriano et al. 2024]. Our goal here is to investigate the impact of different dataset design choices on the performance of ML models. By utilizing the concept of instance hardness, we identified instances that were consistently misclassified or correctly classified, forming distinct groups of easy and hard instances for further investigation.

Most methods for assessing hardness at the instance level depend on the adopted algorithm, meaning the model itself can influence the characterization of individual data samples [Seedat et al. 2022].

Misclassification depends on the learning algorithm and the relation of the instance with the entire training dataset. In this way, the probability of an instance being misclassified is relative. Ideally, we seek a measure inherent to the data difficulty independent of any specific model. The goal is to ensure that the measure reflects the inherent difficulty of the data.

Each ML technique adopts a specific strategy to identify and learn patterns within the data and may be more adequate for certain types of learning tasks while potentially being less effective for others. Consequently, if an instance consistently receives incorrect classifications from multiple ML techniques with different biases, it can be considered difficult to classify or hard [Smith et al. 2014]. Building on this, an empirical definition based on the classification behaviour of a set of algorithms is proposed by [Smith et al. 2014]. At the same time this is a quantitative measure, it also can support subjective evaluations [Valeriano et al. 2023; Valeriano et al. 2024].

In order to define IH, as proposed by [Smith et al. 2014], consider D as a dataset containing n pairs of observations $(\mathbf{x_i}, y_i)$. Each $\mathbf{x_i} \in X$ is an instance described by m input features that belong to the class specified by $y_i \in Y$, the instance label. In addition, let $h : X \to Y$ denote a classification hypothesis, that is, an ML predictive model generated from D. In practice, h is determined by a learning algorithm l trained on a dataset D using specific hyperparameters β. In this way, the hardness level (IH) of the instance $\mathbf{x_i}$ with respect to h can be expressed as:

$$IH_h\big(\mathbf{x_i}, y_i\big) = 1 - p(y_i \mid \mathbf{x_i}, l(D, \beta)), \tag{1}$$

where p denotes the probability the learning algorithm l assigns $\mathbf{x}_i$ to its expected class y_i.

To obtain a more robust measure of instance hardness, consider a set of representative learning algorithms denoted as $\mathcal{L}$ [Smith et al. 2014]. This set consists of ML algorithms with different biases. A comprehensive measure of IH can be derived by evaluating the performance of the instance under consideration across multiple algorithms in $\mathcal{L}$. The IH measure can then be expressed as:

$$IH_{\mathcal{L}}\big(\mathbf{x}_i, y_i\big) = 1 - \frac{1}{|\mathcal{L}|} \sum_{j=1}^{|\mathcal{L}|} p\big(y_i \mid \mathbf{x}_i, l_j(D, \beta)\big) \tag{2}$$

This equation expresses that if an instance consistently gets misclassified by a diverse pool of learning algorithms, denoted as $\mathcal{L}$, it can be considered hard to classify. Conversely, easy instances are expected to be correctly classified by any learning algorithm.

3. CASE STUDY

The dataset analyzed in this study was obtained from a publicly available repository[1][Mello et al. 2020]. It consisted of the results of laboratory blood tests collected during the COVID-19 pandemic from a large hospital in São Paulo, Brazil. Raw data also includes patient demographics such as age and sex, place of hospitalization, and dates of hospitalization and discharge. The dataset was framed as a binary classification problem, aiming to predict severe cases of hospitalized COVID-19 from the blood tests collected on the first day of attendance. We have adopted as severity criteria 14 or more days of hospitalization or death. All decisions were made in consultation with a data specialist from the medical field. We refer to a previous work for a detailed description of the dataset assembling [Valeriano et al. 2022].

The resulting dataset contains 1432 instances, among them 36.7% belonging to the severe class, and 17 input features. We conducted an initial evaluation of the predictive performance of models in these dataset. The algorithms adopted were Random Forest, Gradient Boosting and Support Vector Machine. Performance was assessed with a ten-fold cross-validation approach. Hyper-parameters

[1] https://repositoriodatasharingfapesp.uspdigital.usp.br/

were tuned in grid-search strategy. Despite challenges such as inconsistent data, missing values, and class imbalance, our ML models achieved promising results with an AUC of 0.75. Details about models training and a deep discussion on the methodology and results can be found in our previous work [Valeriano et al. 2022]. In the present work, our objective was a better understanding of the performance of the ML models. Particularly on how the definition adopted to consider a severe case could negatively impact the predictive results registered. We conduct our investigation by analyzing hard and easy instances of the dataset.

3.1 Selecting hard and easy instances

To calculate the instance hardness level, we adopted the measure as implemented in the PyHard library [Lorena et al. 2024]. The error probability, when predicting the label of each instance averaged across multiple ML models, is assessed using a five-fold cross-validation procedure with five repetitions. The adopted algorithms are Bagging, Support Vector Machines (with linear and RBF kernels), Multilayer Perceptron, Gradient Boosting, Logistic Regression, and Random Forest. They were all trained with the default hyperparameter values of the scikit-learn Python package. The resulting instance hardness (IH) levels consist of values in the interval [0,1], where higher values are attributed to harder instances.

To determine easy and hard instances, we utilized the 10th and 90th percentiles of the IH values, stratified by class. This stratification resulted in four groups: the hard and easy groups of the *non-severe* class contain 89 instances each, while groups of the *severe* class contain 52 instances each.

3.1.1 *Formulating hypothesis with the data specialist.* We aim to assess the impact of our dataset assembly decisions on the performance of the ML models. These decisions were made in collaboration with an expert to ensure consistent preprocessing, although they remain subject to debate. Once we identified hard and easy instances, we sought to understand if our dataset design choices contributed to a higher difficulty level for specific instances. We discussed potential factors that could have influenced the models' performance and formulated two research questions.

RQ1: How is the distribution of hospitalization days among hard and easy patients? As mentioned earlier, we used a 14-day criterion to differentiate between severe and non-severe patients. We hypothesize that hard patients would exhibit hospitalization length closer to the classification boundaries (between 12 and 16 days) as they possess characteristics similar to those of the opposite (non-severe) class, making them more challenging to classify.

RQ2: Are patients progressing to death easier or harder to classify than patients with extended hospitalization? Death among COVID-19 patients is a rare event compared to the disease recovery rates. When assembling our dataset, we included patients with extended hospitalizations as severe cases to address potential class imbalance bias in our ML models. Thus, our proxy for severity comprises two criteria: death and extended hospitalization. We aim to determine if either of these events is easier to classify in the absence of the other.

3.2 Results

Our goal is to understand how our dataset assemblage decisions impact the predictive performance of ML models. To test our hypotheses, we explored the difference between hard and easy instances.

RQ1: How is the distribution of hospitalization days between hard and easy patients? We adopted 14 days as the criterion to split data into *severe* and *non-severe* patients. To verify if instances with a hospitalization length of around 14 days are harder to classify, we inspected the distribution of hospitalization days inside our four groups: easy and *severe*, easy and *non-severe*, hard and *severe*, and hard and *non-severe*. Figure 1 presents the histograms representing these distributions. Our analysis confirmed the hypothesis that instances with hospitalization lengths of around 14 days tend to be more challenging to classify. In the *non-severe* class, easy patients predominantly exhibited hospitalization lengths of fewer than 8 days, while hard patients generally had lengths exceeding 10 days. Conversely, within the *severe* class, hard instances were concentrated around fifteen days. Easy instances in the *severe* class demonstrated a wide range of hospitalization days. This group includes a significant proportion of death cases for which we do not have hospitalization length information.

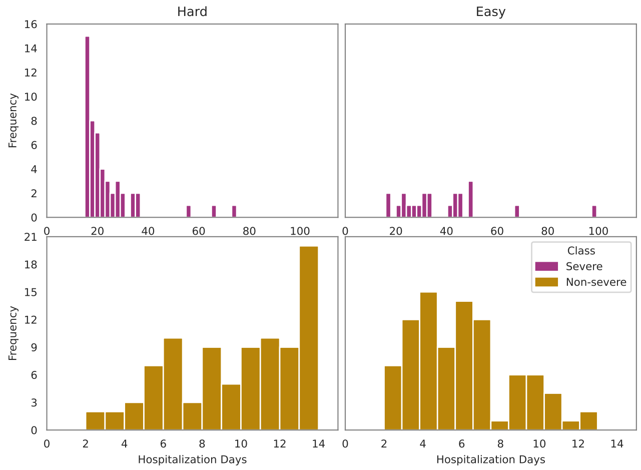

Fig. 1. The distribution of hospitalization days according to class and hardness level. *Non-severe* cases are represented in yellow, while the *severe* class is depicted in dark pink. Hard instances present hospitalization days close to the value adopted to split classes.

RQ2: Patients progressing to death are easier or harder to classify than patients presenting an extended hospitalization? Since our proxy to severity consisted of two criteria (extended hospitalization or death), we would like to understand if any of them is easier to classify. We analyzed the presence of death cases inside our groups of severe instances. Among easy and *severe* patients, there are 29 death cases, while one is placed in the hard *severe* group. This indicates that death is an outcome easier to predict than extended hospitalization. Next, we present an investigation to understand if predicting only death results in an easier problem.

Exploring Different Definitions of Extended Hospitalization: Based on the insights gained so far, we have identified opportunities to enhance the performance of our ML models. Given that instances close to the threshold for defining an extended hospitalization were more challenging to classify, we sought to identify the optimal value for defining this cutoff. We tested cutoff values ranging from 7 to 30 days and measured the performance of our models. To evaluate the impact of different cutoff values, the instance hardness value was assessed again in each new definition of severity. We also conducted a performance evaluation using a five-fold cross-validation strategy. We employed the same set of seven algorithms adopted to measure IH. To deal with class imbalance, a random subsampling in the majority class was performed within each training set generated in the cross-validation process.

In addition to investigating the impact of different definitions of extended hospitalization, we also assessed the performance of our models when considering death as the sole criterion for determining aggravated conditions, given the evidence suggesting that death is relatively easier to predict. Figure 2 illustrates the evolution of mean instance hardness values and model performance metrics across different day thresholds used to split classes. We evaluate instance hardness both overall and separately by class. To understand the decrease in precision, we present additional plots showing the evolution of false positives and true positives relative to the actual number of instances in each class.

Analyzing Figure 2, we observe that as the cutoff value for splitting classes increases, the metrics evolve in a generally consistent manner. When considered across the entire dataset, instance hardness

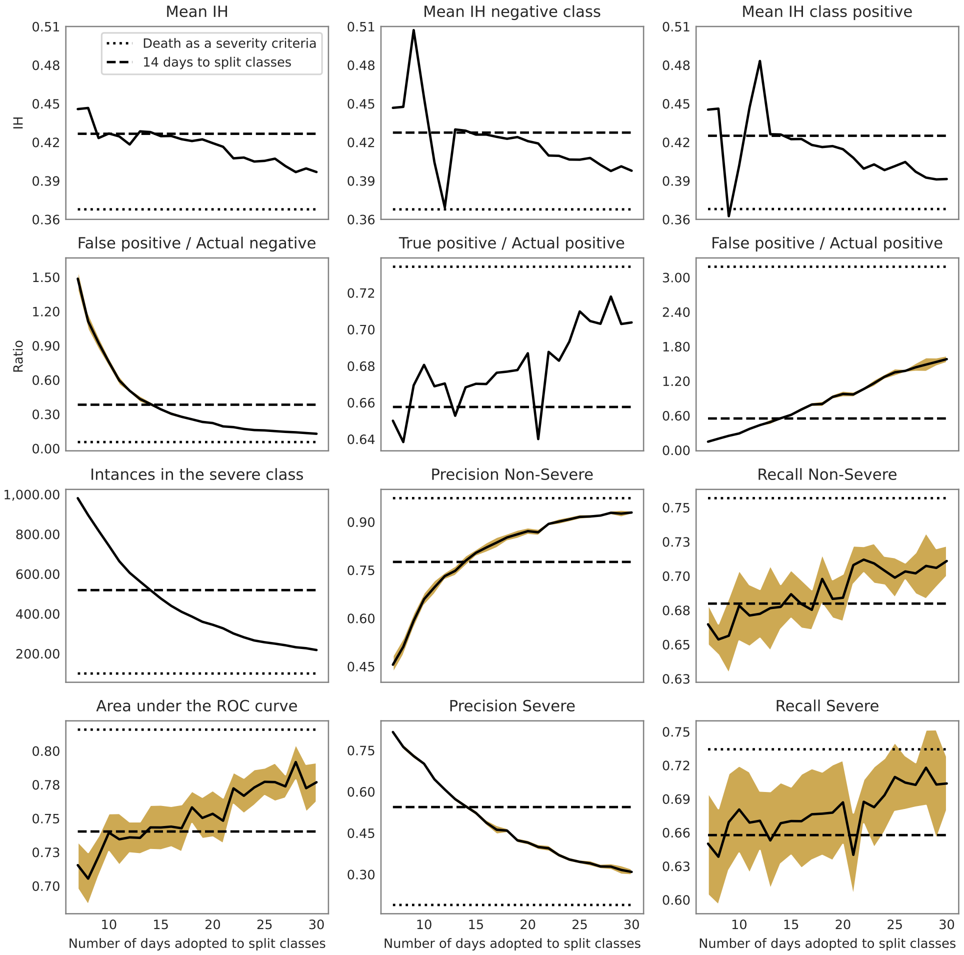

Fig. 2. Evolution of several metrics as the number of days considered for classifying *severe* and *non-severe* cases increases from 7 to 30 days. The dotted line represents values when considering only death as the severity criterion. The dashed line represents the current dataset (14 days of hospitalization). The continuous line in the last three rows represents the average of seven ML models, and the shaded area indicates the standard deviation. Results were obtained in a five-fold cross-validation procedure.

values decrease almost linearly. When examining IH values by class, we see a peak in difficulty for the negative class around 10 days. This peak decreases quickly, reaching the lowest level before 15 days, then rises again and follows a linear decrease pattern. Conversely, IH values for the positive class exhibit the opposite behaviour, indicating that classifying one class may be easier at the cost of misclassifications in the other. In all three scenarios, considering only death as a severity criterion presents the lowest level of IH while considering 14 days presents an intermediate profile. Regarding performance metrics, we observe a consistent increase in AUC, recall for both classes and precision in the *non-severe* class. When considering only death as the *severity* criterion, we achieve the highest predictive performance values, while considering 14 days of hospitalization results in intermediate performance values. However, precision in the positive class shows the opposite behaviour; performance declines as the cutoff value to separate classes increases, with the poorest performance occurring when considering only death as the *severity* criterion.

To better understand these values, we plot the ratios of false positives to actual negatives, true positives to actual positives, and false positives to actual positives. Analyzing these plots in the second row of Figure 2, we realize that the number of false positives decreases in proportion to the size of the negative class, with the lowest rate achieved when considering only death as the *severity* criterion. The number of true positives increases relative to the actual size of the positive class, with the highest level achieved when considering only death as the severity criterion. Thus, models correctly classify positive samples but also present more false positives. This is why precision in the *severe* class decreases while all other metrics increase. In the context of predicting whether a patient will be *severe* or *non-severe*, achieving high accuracy in the positive class is the most important metric, even if it means classifying some *non-severe* patients as *severe*. The consequences of misclassifying a *severe* case are much more serious than misclassifying a *non-severe* case. Therefore, considering only death as the criterion for a *severe* condition is the best option for distinguishing patients despite the lower precision achieved.

4. DISCUSSION

In this case study, we adopted a data-centric approach to explore the potential of leveraging instance hardness values analysis to enhance the performance of ML models. Our focus was on a classification task to predict aggravated conditions associated with COVID-19. Our investigation aimed to assess how the decisions made during the dataset assembly process have influenced the models' performance. To accomplish this, we leverage instance hardness values to distinguish between challenging and straightforward instances for each class.

In collaboration with a data specialist, we explored hypotheses that could influence the difficulty level of instances in the dataset. Our analysis confirmed our intuition that instances characterized by hospitalization lengths around 14 days (the initial cutoff for distinguishing *severe* from *non-severe* COVID cases) pose a more significant classification challenge. Notably, the distribution of hospitalization days varied significantly between easy and hard instances in both *severe* and *non-severe* classes. Furthermore, we investigated the distinction between patients who succumbed to the disease and those with prolonged hospital stays concerning their hardness level. Our analysis revealed that patients who progressed to death were classified as *severe* easily compared to those who were hospitalized for a long period. This suggests that death is a strong indicator of *severity*, and it might be easier to distinguish death cases from *non-severe* outcomes than extended hospitalizations. Drawing from these insights, we conducted experiments to explore different cutoff limits for defining prolonged hospitalizations. The results indicated that extending the duration considered for classifying hospital stays as prolonged led to improved predictive performance of the ML models. Moreover, focusing solely on mortality as a criterion for *severity* prognosis yielded the best predictive outcomes. However, this led to a decrease in precision in the positive class, because the number of false positive instances increased. This trade-off is justified, as misclassifying a *non-severe* case as *severe* is less detrimental than the opposite scenario.

These results underscore the significance of adopting a data-centric approach in designing and analysing ML systems. By focusing on the characteristics and limitations of the training data, we gain valuable insights into the factors contributing to instance hardness and the performance of predictive models. This, in turn, informs the refinement of problem design and feature selection, ultimately enhancing the accuracy and reliability of ML models, particularly in the healthcare domain.

As a case study, this work does not aim to present a framework or methodology applicable in a general context. This is not a limitation, as data-centric solutions are inherently specific to the data they address. The approach can be easily extended to other domains and problems, as the criterion for identifying instances that are easy and hard to classify per class is generic and not domain-specific. However, the hypothesis to be investigated regarding the profiles of the easy and hard instances per class will naturally vary based on the domain characteristics. In addition to the specific contributions to the problem at hand, we also showcase how model performance can be enhanced through improvements in data quality. Furthermore, we illustrate how instance hardness can support this process. This does not exclude adopting more robust models and optimising hyperparameters, but in a better-designed problem, any model engineering will succeed more effectively.

Acknowledgements

To the Brazilian research agencies CAPES (grant 88887.507037/2020-00) and FAPESP (21/06870-3).

REFERENCES

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., and Cox, D. D. Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science & Discovery* 8 (1): 014008, 2015.

Chatzimparmpas, A., Paulovich, F. V., and Kerren, A. Hardvis: Visual analytics to handle instance hardness using undersampling and oversampling techniques. *arXiv preprint arXiv:2203.15753*, 2022.

Hüllermeier, E. and Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning* 110 (3): 457–506, 2021.

Liu, C., Smith-Miles, K., Wauters, T., and Costa, A. M. Instance space analysis for 2d bin packing mathematical models. *European Journal of Operational Research* 315 (2): 484–498, 2024.

Lorena, A. C., Paiva, P. Y., and Prudêncio, R. B. Trusting my predictions: on the value of instance-level analysis. *ACM Computing Surveys* 56 (7): 1–28, 2024.

Mello, L. E., Suman, A., Medeiros, C. B., Prado, C. A., Rizzatti, E. G., Nunes, F. L., Barnabé, G. F., Ferreira, J. E., Sá, J., Reis, L. F., et al. Opening brazilian covid-19 patient data to support world research on pandemics. *Zenodo*, 2020.

Napierala, K. and Stefanowski, J. Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems* vol. 46, pp. 563–597, 2016.

Oala, L., Maskey, M., Bat-Leah, L., Parrish, A., Gürel, N. M., Kuo, T.-S., Liu, Y., Dror, R., Brajovic, D., Yao, X., et al. Dmlr: Data-centric machine learning research–past, present and future. *arXiv preprint arXiv:2311.13028*, 2023.

Paiva, P. Y. A., Moreno, C. C., Smith-Miles, K., Valeriano, M. G., and Lorena, A. C. Relating instance hardness to classification performance in a dataset: a visual approach. *Machine Learning*, 2022.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *Journal of machine learning research* 12 (Oct): 2825–2830, 2011.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L. M. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. pp. 1–15, 2021.

Seedat, N., Crabbé, J., Bica, I., and van der Schaar, M. Data-iq: Characterizing subgroups with heterogeneous outcomes in tabular data. *arXiv preprint arXiv:2210.13043*, 2022.

Seedat, N., Imrie, F., and van der Schaar, M. Dc-check: A data-centric ai checklist to guide the development of reliable machine learning systems. *arXiv preprint arXiv:2211.05764*, 2022.

Seedat, N., Imrie, F., and van der Schaar, M. Dissecting sample hardness: A fine-grained analysis of hardness characterization methods for data-centric ai. *arXiv preprint arXiv:2403.04551*, 2024.

Smith, M. R., Martinez, T., and Giraud-Carrier, C. An instance level analysis of data complexity. *Machine learning* 95 (2): 225–256, 2014.

Valeriano, M., Matran-Fernandez, A., Kiffer, C., and Lorena, A. C. Understanding the performance of machine learning models from data-to patient-level. *ACM Journal of Data and Information Quality*, 2024.

Valeriano, M. G., Kiffer, C. R. V., and Lorena, A. C. Supporting decision making in health scenarios with machine learning models. In *Anais do simposio brasileiro de pesquisa operacional*, 2022.

Valeriano, M. G., Paiva, P. Y. A., Kiffer, C. R. V., and Lorena, A. C. A framework for characterizing what makes an instance hard to classify. In *Brazilian Conference on Intelligent Systems*. Springer, pp. 353–367, 2023.

Valeriano, M. G., Pereira, J. L. J., Kiffer, C. R. V., and Lorena, A. C. Explaining instances in the health domain based on the exploration of a dataset's hardness embedding. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '24 Companion)*. ACM, Melbourne, VIC, Australia, 2024.

Zha, D., Bhat, Z. P., Lai, K.-H., Yang, F., Jiang, Z., Zhong, S., and Hu, X. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*, 2023.