Understanding the performance of machine learning models from data-to patient-level

Understanding the performance of machine learning models from data-to patient-level

MARIA GABRIELA VALERIANO, Instituto Tecnológico de Aeronáutica and Universidade Federal de São Paulo, Brazil
ANA MATRAN-FERNANDEZ, University of Essex, United Kingdom
CARLOS ROBERTO VEIGA KIFFER, Universidade Federal de São Paulo, Brazil
ANA CAROLINA LORENA, Instituto Tecnológico de Aeronáutica, Brazil

Machine Learning (ML) models have the potential to support decision-making in healthcare by grasping complex patterns within data. However, decisions in this domain are sensitive and require active involvement of domain specialists with deep knowledge of the data. In order to address this task, clinicians need to understand how predictions are generated so they can provide feedback for model refinement. There is usually a gap in the communication between data scientists and domain specialists that needs to be addressed. Specifically, many ML studies are only concerned with presenting average accuracies over an entire dataset, losing valuable insights that can be obtained at a more fine-grained patient-level analysis of classification performance. In this paper, we present a case study aimed at explaining the factors that contribute to specific predictions for individual patients. Our approach takes a data-centric perspective, focusing on the structure of the data and its correlation with ML model performance. We utilize the concept of *Instance Hardness*, which measures the level of difficulty an instance poses in being correctly classified. By selecting the hardest and easiest to classify instances, we analyze and contrast the distributions of specific input features and extract meta-features to describe each instance. Furthermore, we individually examine certain instances, offering valuable insights into why they offer challenges for classification, enabling a better understanding of both the successes and failures of the ML models. This opens up the possibility for discussions between data scientists and domain specialists, supporting collaborative decision-making.

ACM Reference Format:
Maria Gabriela Valeriano, Ana Matran-Fernandez, Carlos Roberto Veiga Kiffer, and Ana Carolina Lorena. 2024. Understanding the performance of machine learning models from data-to patient-level. 1, 1 (May 2024), 19 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Machine learning (ML) models can efficiently extract patterns from data and make accurate predictions for future observations. In many cases, these models have shown superior performance compared to traditional strategies like logistic regression [5, 9, 20]. However, despite their advantages, the adoption of ML models in healthcare presents challenges due to the potential risk of erroneous predictions. One key challenge lies in the interpretability of these models, as their decision-making processes can be complex. This lack of interpretability can hinder trust in the predictions

Authors' addresses: Maria Gabriela Valeriano, Instituto Tecnológico de Aeronáutica and Universidade Federal de São Paulo, Praça Marechal Eduardo Gomes, São José dos Campos, Brazil, maria.valeriano@ga.ita.br; Ana Matran-Fernandez, University of Essex, Wivenhoe Park, Colchester, United Kingdom; Carlos Roberto Veiga Kiffer, Universidade Federal de São Paulo, Rua Botucatu, São Paulo, Brazil; Ana Carolina Lorena, Instituto Tecnológico de Aeronáutica, Praça Marechal Eduardo Gomes, São José dos Campos, Brazil.

Manuscript submitted to ACM

of ML models, thereby limiting their utilization in healthcare settings [23]. To address this issue, an interdisciplinary approach is necessary, involving both machine learning experts and healthcare professionals [2]. By combining their expertise, such collaborative efforts can help develop ML systems that are not only accurate but also interpretable and trustworthy.

In order to enhance the effectiveness of ML models, it is crucial to actively involve domain specialists who possess rich expert knowledge of the data. By integrating human expertise into the modeling and decision-making processes, ML models can be developed to be more accurate, robust, and reliable [7]. This collaboration between domain specialists and ML experts has the potential to yield significant advantages. However, there is a lack of results reporting the outcomes of such interactions in the healthcare domain. To achieve this task, it is important for clinicians to comprehend how predictions are made so that they can provide valuable feedback for model refinement. A study exploring the issue of model explainability from the perspective of clinicians, revealing their concerns, is presented in [23]. Clinicians expressed a desire to understand why model predictions may deviate from clinical protocols and which factors contribute to specific predictions for individual patients. This understanding is crucial for establishing trust in ML models and facilitating effective collaboration between clinicians and data scientists.

In this work, we present a case study addressing the discussed issues. Our approach focuses on the data itself, utilizing a data-centric perspective to understand its structure and correlate it with the performance of the models. To identify instances that are currently misclassified, we adopt the concept of Instance Hardness (IH) as defined by Smith *et al.* [19]. IH represents the likelihood of an instance being misclassified. By employing a pool of different algorithms, we can calculate the IH value for each instance, where instances with high levels of IH can be considered *hard*, while those with low levels are *easy*. Additionally, we extract a set of hardness measures known as meta-features for each instance. These meta-features provide further insights into the characteristics of the instances and aid in understanding their classification difficulty [1].

The analysis of instance hardness, with or without the assistance of meta-features, has already been employed to investigate the reasons behind misclassifications and to evaluate the reliability of a model's predictions. For instance, Paiva and colleagues [14] introduced frameworks that project instances onto a bi-dimensional space based on their hardness level and meta-feature values. Another approach proposed by Seedat *et al.* [16] utilizes predictive confidence and examines instances' behavior during the training process to identify hard and ambiguous instances. They demonstrate that removing hard instances can lead to improved model performance. In the healthcare domain specifically, Houston *et al.* [8] employed meta-feature values to show that model failures were attributable to biases in the dataset rather than a lack of sensible assumptions by the models. Additionally, Chatzimparmpas *et al.* [3] provided a visual comparison of hard instances, enabling users to inspect instances for potential oversampling issues.

Building upon these advancements, our study delves deeper into the analysis of hard instances, aiming to extract valuable insights from them. In healthcare scenarios, removing hard instances, although it may improve model performance, can be risky. Therefore, a more appropriate approach to address models' performance is to rely on the knowledge and expertise of domain specialists to decide if they should be removed, corrected, or kept inside the dataset. By identifying and analyzing hard instances and contrasting them with the easy ones, our study aims to provide a better explanation for the decisions made by the models, facilitating fruitful discussions between data scientists and clinicians.

To conduct our analysis, we utilize a COVID-related dataset obtained from a Brazilian public repository[1]. This dataset was specifically curated, with data domain specialists, to predict the progression of patients towards aggravated

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

conditions using routinely collected tests from the first day of hospitalization. Predicting such outcomes is particularly challenging due to imbalanced data, missing values, and the fact that the data was not originally collected for research purposes (i.e., it is observational data). Through assembling this dataset, we sought to explore the potential of leveraging routinely collected data to build predictive models. In previous work [25], we presented the performance of ML models on this dataset, achieving an AUC of 0.75. We compared this performance to related studies and concluded that the inclusion of additional features would be necessary to effectively distinguish between the classes.

Here, we deeper explore data distribution to understand models' performance. Three meta-features were inspected through visualizations, showing trends between feature values, meta-features, and IH levels. Furthermore, we show how the meta-features can provide a better understanding of the performance of the models from the data perspective and how this information can be interpreted at the instance (i.e., patient) level, thus facilitating discussions between data scientists and domain specialists.

2 METHODS

This work offers a dataset analysis based on data distribution and meta-feature values indicating possible reasons for misclassifications. Starting from the concept of Instance Hardness, we select the most difficult instances of the dataset and oppose them with those instances that are easily classified. Our methodology encompasses: (i) the assembling and preparation of the dataset; (ii) the definition of IH and how we measure it; (iii) the analysis of instances through feature values; and (iv) the definition and measures of meta-feature values.

2.1 Data

To develop our analyses, we employed a publicly available dataset assembled to classify hospitalized patients with COVID from a large private hospital in Brazil into *severe* and *non-severe* cases. Patients were considered severe when they had an extended hospitalization (14 days or more) or died within 14 days of hospitalization. This dataset was extracted from a large repository containing 2,891,301 results from blood tests from 14,673 patients between March 2020 and May 2021. To assemble our dataset from the original data available, all decisions were broadly discussed with the support of an infectious disease specialist. We refer to our previous work for a deeper discussion about these criteria [25].

The inclusion criteria for this study were:

- Patients with a positive result for COVID test within the first four days from admission.
- Blood tests taken within the first four days of hospitalization. Where more than one test was taken within this window, the results from the first test were used. Most patients (92.9%) had the blood test done in the first day.
- Patients older than 16 years.
- Patients with blood count records.

After applying the criteria above, we included results from blood tests only when they were available for at least 50% of patients in each class. In the present analysis, we also added some exclusion criteria. Since we are interested in analyzing the hardness level of instances, we removed those that would present a different pattern. This included patients who had been previously hospitalized within the last six months; patients who underwent surgery during their hospitalization; and patients who were re-hospitalized after recovering from COVID-19.

The resulting dataset consisted of $n = 1076$ instances, with 5.59% of missing values. Rather than filling in the missing values, which could potentially introduce bias and alter the distribution and distances between instances, we made the

decision to remove instances with missing values. This choice was motivated by the fact that some of the meta-features we intended to explore relied on distance calculations. We identified two features with a substantial number of missing values when compared to the other features (46.7 and 35%). Based on previous work, we knew that they were deemed less critical for predicting severity [24]. Therefore, we removed these two features from the dataset. Following this step, any instance with a missing value was subsequently dropped. As a result, the final dataset size was reduced to $n = 885$ instances, ensuring a more complete and reliable dataset for our analysis.

At this stage, we encountered a class imbalance ratio of 1:0.4 in the dataset. It is well-known that class imbalance can significantly impact the performance of machine learning models. This dataset is no exception, as there is a noticeable difference in model performance when using class balancing techniques compared to not employing them. Particularly, the imbalanced dataset exhibits considerably lower performance in terms of sensitivity, with values hovering around 40% across all tested datasets. On the other hand, when a random sub-sampling technique is applied to achieve a balanced representation of both classes, sensitivity values improve to over 60% for almost all tested datasets. Since our main objective is not to train a model for making predictions but rather to analyze the hard instances, particularly in the minority class, we randomly sampled instances from the non-severe class in order to balance the dataset. As a result, the final dataset size was reduced to $n = 526$. We provide a comprehensive comparison of model performances between the balanced and imbalanced datasets in the Appendix (see A).

Table 1 summarizes the dataset adopted in this study. The feature values in the dataset represent the results of blood tests conducted within the first four days of hospitalization and the sex and age of the patients.

Table 1. Summary of the analysed dataset including the number of patients, percentage of female patients, age statistics, and blood tests.

	Severe	Non-severe
Number of instances	263	263
Female ratio	0.33	0.34
Age (mean ± stdev)	68.47 ± 12.71	58.14 ± 15.50
Blood tests	sodium, potassium, urea, C-reactive protein, D-dimer, GOT, GPT, creatinine, blood count	

2.2 Assessing Instance Hardness

In any given dataset, certain observations can offer more challenges for classification than others due to various factors [19]. For example, an observation located near the decision boundary, where instances from different classes overlap, tends to be more difficult to classify compared to an instance situated in a dense region far away from the decision boundary, surrounded by instances from the same class. The concept of instance hardness, introduced by Smith and colleagues [19], quantifies the level of challenge associated with classifying a particular instance within a dataset. This measure is determined by evaluating the performance of different ML algorithms when classifying the specific instance under consideration.

Each machine learning technique adopts a specific strategy to identify and learn patterns within the data, which introduces a bias towards certain types of learning tasks while potentially being less effective for others. As a result, if an instance consistently receives incorrect classifications from multiple ML techniques with different biases, it can be considered difficult to classify or "hard".

In order to define IH, 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 and belongs to the class specified by $y_i \in Y$, the instance label. In addition, let $h : X \rightarrow Y$ denote a classification hypothesis, that is, an ML predictive model generated from D. The hardness level of the instance $\mathbf{x}_i$ with respect to h can be expressed as:

$$IH_h(\mathbf{x}_i, y_i) = 1 - p(y_i|\mathbf{x}_i, h), \tag{1}$$

In practice, h is determined by a learning algorithm l trained on a dataset D using specific hyperparameters β [19]. To obtain a more robust measure of instance hardness, the authors consider a set of representative learning algorithms denoted as $\mathcal{L}$. This set consists of machine learning algorithms with different biases. By evaluating the performance of the instance under consideration across multiple algorithms in $\mathcal{L}$, a comprehensive measure of instance hardness can be derived. The IH measure can then be expressed as:

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

This equation expresses the notion 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.

Based on this definition, the loss value of each instance was assessed to measure the level of hardness. We adopt in this step a five-fold cross-validation strategy. This step was necessary because, to measure the performance of a predictive model, testing data is required, which should be distinct from the training dataset. Cross-validation allows the generation of different train and test data partitions from the same dataset. Initially, the dataset is divided into approximately equal-sized folds, with $k = 5$ folds being created. During each iteration, $k - 1$ folds are used for training a predictive model, while the remaining fold is reserved for testing, simulating the introduction of new data to the model. This process is repeated k times, with each fold being utilized as the test set just once.

It is worth noting that each model has hyperparameters that need to be chosen. In many cases, a common practice is to employ techniques for hyperparameter optimization to achieve the best model performance. However, such techniques can be time-consuming. Given the focus of this paper, which is to understand how the data structure impacts model performance, no hyperparameter optimization was conducted, and default values were used instead.

For each class, we identified the instances that were the hardest and easiest to classify, which we will refer to as "hard" and "easy" instances from this point onward. To determine these instances, we utilized the tenth and ninetieth percentiles of the IH values, resulting in four groups, each containing 26 instances.

2.3 Feature values analysis

We focused our analysis on four specific features: three blood tests that are known to be influenced by the presence of COVID-19 and age, which is a known risk factor for the severity of this disease. These tests have demonstrated patterns in their values among patients who develop an aggravated condition, as reported in the medical literature [6, 18, 21]. Additionally, our previous research using a similar dataset from the same data source has indicated that these particular features are among the most important for predicting the development of an aggravated condition [24].

Table 2 provides the names of these blood tests, along with their reference range values observed in healthy individuals. It also outlines how the values of these blood tests are typically affected by the presence of COVID-19 (i.e., whether

they increase or decrease) as reported in studies such as [4, 13]. Moving forward, these reference values will serve as the standard when referring to high or low blood test values in our analysis.

Table 2. Blood tests analysed and reference values in a healthy patient.

Feature names	Reference values	How the values are affected by the COVID-19
Lymphocytes (%)	20–40 %	Decrease
C-Reactive protein (CRP)	lower than 1 mg/dL	Increase
Urea	13–43 mg/dL	Increase

We conducted our analysis of feature values by examining their distribution within the hardest instances and contrasting them with the patterns observed in the easiest instances and observing differences or trends between these two groups.

2.4 Meta-features

Smith *et al.* [19] and Arruda and colleagues [1] have defined a set of measures aimed at offering pausible explanations for instances hard to classify. These measures, known as *Hardness Measures* (HM), offer valuable insights into the challenges faced by classification algorithms when dealing with specific instances. While the inner workings of machine learning models can be complex and non-intuitive, HM can be visually represented and easily interpreted.

In this study, we calculated the HM values based on the implementation described by Paiva *et al.* [14], where higher values of HM indicate instances that are more challenging to classify. These measures are computed using all instances in the dataset and have a polynomial computational cost relative to the number of features and observations. From a set of 13 meta-features, we selected three that most correlated with the IH.

Next, we describe these three meta-features. The remaining meta-features and their correlation values with the IH measure are briefly discussed in the Appendix (see B.2).

2.4.1 Class likelihood ($CL(\mathbf{x_i})$). The Class likelihood (CL) measure represents the likelihood of an instance $\mathbf{x_i}$ belonging to its assigned class. It can be calculated by multiplying the conditional probability $P(\mathbf{x_i}|y_i)$, which represents the probability of $\mathbf{x_i}$ belonging to class y_i, by the prior probability of the class $P(y_i)$. In our case, the prior probability of each class is set as $\frac{1}{C}$, being C the number of classes. To calculate the conditional probability, we assume that each feature is independent of the others and multiply the individual conditional probabilities. The CL measure follows a naive Bayes classifier approach, assuming equal importance for all features when estimating the occurrence of an event.

The probability of feature values is estimated independently in each dimension. The overall data distribution is then estimated by assuming conditional independence, where the estimation is obtained by multiplying the probabilities of each feature. Instances that have a higher likelihood of belonging to their assigned classes will have lower CL values. These instances can be considered easier to classify compared to those with higher CL values. In the Appendix, we provide a visual representation of the class likelihood measure in a synthetic two-dimensional problem (see B.4).

2.4.2 k-Disagreeing Neighbors ($kDN(\mathbf{x_i})$). The kDN measure represents the fraction of neighbors of $\mathbf{x_i}$ (i.e., observations that are closer to $\mathbf{x_i}$ in the input space) that do not belong to the same class as $\mathbf{x_i}$. The number of neighbors was set to $k = 10$. Instances with high values of kDN are more challenging to classify due to their proximity to members of other classes. To find the neighborhood of an instance, the distance between data points was calculated using the

range-normalized Manhattan distance, where the distance between two points is the sum of the absolute differences of their Cartesian coordinates.

In the Appendix, we provide a graphical representation of the *kDN* metric on a small dataset, along with a visual depiction of the Manhattan distance (see B.4).

2.4.3 Disjunct Class Percentage (DCP($\mathbf{x_i}$)). This measure involves constructing a Decision Tree (DT) using all instances in the dataset *D*. A DT is a tree-like structure where each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node holds a class label. The feature and threshold values are chosen to generate nodes with the lowest impurity, measured using the Gini Index. The tree is then post-pruned using the cost complexity parameter to remove splits that do not significantly reduce impurity.

The disjunct refers to the leaf node where $\mathbf{x_i}$ is classified. Within this node, the ratio between the number of instances that do not belong to the same class as $\mathbf{x_i}$ and the total number of instances in the disjunct is calculated. Easier instances tend to have few examples of other classes within their disjunct. Figure 8(c) illustrates the *DCP* measure, where easier instances generally exhibit a low presence of examples from other classes in their corresponding disjunct.

Each of these HM encompasses different perspectives of the data. Feature *kDN* is a measure of similarity, mapping whether the instance is located far or close to other instances of their own class, and with all features receiving the same weight. Feature *CL* also gives the same weight to each feature, although taking into account how those values are correlated with the ground truth class. On the other hand, *DCP* selects features and thresholds that are more important in order to differentiate classes.

3 RESULTS

In this study, we conducted a data-centric analysis of ML model performance focusing on the concept of instance hardness. We utilized IH values to select easy and hard instances, and compared feature and meta-feature values of those instances. The IH levels were calculated as the likelihood of an instance being misclassified. The performance of the models was accessed through 5-fold cross-validation using a pool of seven algorithms with different biases. The distributions of IH values are similar across both classes, and both have instances that are hard to classify. The severe class presents a histogram slightly shifted to the right with higher levels of IH. The histograms are shown in the Appendix (see Section B.1). For each class we selected ten percent of the easiest and hardest instances. The mean IH for the hard instances selected is 0.662 and 0.242 for the easy ones. Table 4 presents the mean and standard deviation values of IH, according to class, in each group.

3.1 Feature distributions

Figure 1 presents the distributions of the lymphocytes (LYM%), age, CRP, and urea values for both easy and hard instances in the severe and non-severe classes. In general, easy instances follow the expected pattern of the disease (please refer to Table 2). Examining the distributions of easy instances, we can observe distinct patterns for each class. Severe patients tend to be older with lower lymphocyte values, while non-severe patients are generally younger with lower CRP and urea values. However, the distribution of hard instances highlights the pattern expected for the opposite class. Especially for age and lymphocytes, hard and non-severe patients present the distribution expected for severe and easy instances. Whilst for CRP and Urea the hard and severe shows a similar distribution to the opposite class. Besides that, while easy instances have a clearly distinct pattern between the two classes, hard instances have large intervals of overlap.

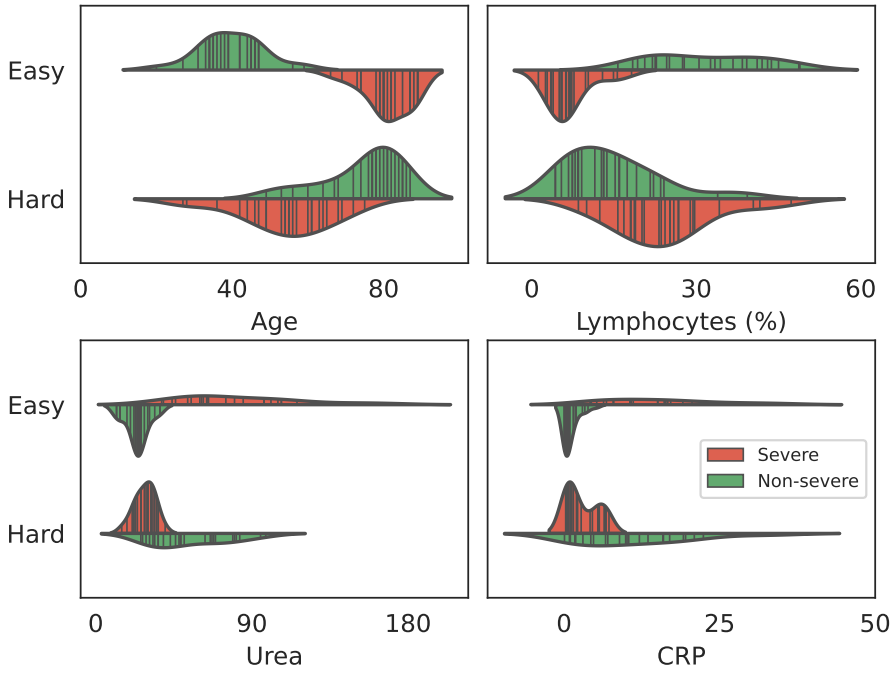


Fig. 1. Distribution of Age, Lymphocytes (%), Urea and CRP, separately for hard and easy instances and severe and non-severe cases.

The original database contains the results of all blood tests collected during patients' stay in the hospital. As lymphocytes and CRP are tests routinely measured through hospitalization, we followed how those values evolve in our four groups. Figure 2 presents the mean and standard deviations for each of them.

By analyzing Figure 2, we can observe a general pattern that helps to explain the behavior of patients who are difficult to classify. Despite the large standard deviations, we can describe each group based on the evolution of their test results:

- Easy-severe patients: These patients are hospitalized with low values of LYM% and high values of CRP, which is expected for severe cases. Due to their critical initial condition, the recovery process takes more than 14 days.
- Hard-severe patients: These patients are admitted with normal values of LYM% and high (but not critically high) values of CRP. Their condition worsens in the following days, explaining their unexpectedly prolonged hospitalization.
- Easy-non-severe patients: These patients are hospitalized with high values of LYM% and low values of CRP. They are discharged before completing 14 days of hospitalization, indicating a relatively smooth recovery process.
- Hard-non-severe patients: These patients are admitted with normal values of LYM%, although they present high values of CRP. However, their condition quickly improves in the next few days, leading to their release in less than 14 days.

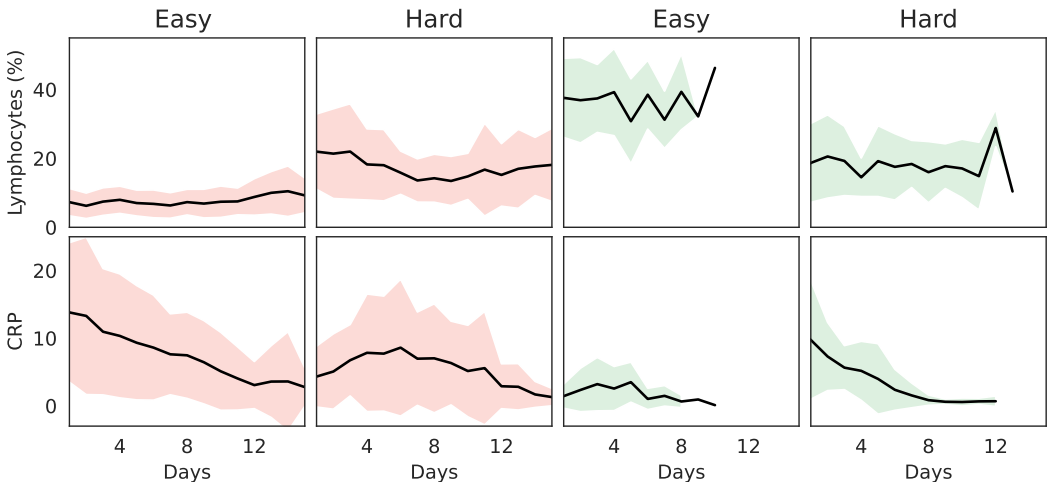


Fig. 2. Evolution of Lymphocytes (%) and CRP values during hospital stay for each class. The dark line is the mean and the standard deviation is represented by the shaded area. The shaded area is coloured according to the original class, red corresponds to severe patients while green indicates non-severe patients. The easy non-severe patients are discharged before the 14-day limit, justifying the shortened number of days.

It is possible to gain some insights into the conditions of each group of patients and why some instances do not exhibit the expected behavior based on their hospitalization status. Generally, the viral load peak of the disease is reached within 5-6 days of symptom onset [22]. Patients categorized as severe and easy were probably hospitalized during this period when the body's immune response was fully activated. On the other hand, severe and hard patients reached their peak a few days after hospitalization. In this case, considering the number of days since the first symptoms can potentially enhance the performance of the predictive model, although it is not possible to retrieve this information inside the data source.

Regarding lymphocyte values, our groups align with patterns reported in medical literature. Tan *et al.* [21] analyzed blood test values in COVID patients classified into three conditions: death cases, patients with moderate outcomes, and those with severe outcomes. The results revealed that LYM% is a reliable marker for classifying patients among these outcomes. In death cases, the percentage of lymphocytes was consistently lower than 5% after two weeks of the disease course. In severe patients, lymphocytes initially decreased and then reached 10% before discharge, while in moderate patients, LYM% fluctuated around normal values and was over 20% upon discharge.

The description of death cases corresponds to the easy-severe patients who, upon hospitalization, already exhibited critical levels of blood test values. In fact, 10 instances within this group correspond to patients who died, while none of the other groups contains death cases. The severe, but not fatal, cases correspond to our severe-hard patients, displaying a decrease in LYM% values after hospitalization. The moderate cases align with our hard and non-severe patients, showing some variability in LYM% but ultimately reaching normal values within 14 days of hospitalization. The easy and non-severe patients represent mild cases, typically released shortly after clinical intervention [21].

Regarding CRP values, we can observe distinct curve shapes between the easy-severe patients and hard-non-severe patients. Both groups initially exhibit high CRP values upon admission, but while one group's values decrease linearly, the other's decrease exponentially. A similar trend is noticed between hard-severe patients and easy-non-severe patients,

although their CRP values increase, the curves display different inclinations and peaks. These results support the findings of Sharifpour *et al.* [17], who suggest that the rate of increase in CRP levels during the first week of hospitalization serves as an indicator of disease progression. The authors report that this rate of increase in CRP level is more predictive than the maximum CRP value. Besides that, we evidence the importance of the rate of decrease, while high CRP levels are characteristic of the severe class, they can also be high in the non-severe class but with a quick decrease.

Although the objective of the trained models was to predict severity at the moment of hospitalization, these findings suggest that considering the CRP value of the second day of hospitalization can potentially improve their performance. This possibility will be investigated in future works.

Finally we consider relevant to describe the sex distribution inside our groups, since studies report sex disparity in severe outcomes of COVID-19 [10]. Usually, studies conclude that men are more likely to evolve to aggravated disease conditions. In our dataset, males and females are equally distributed between the two classes. Among our groups, in both classes, the distribution of sex among easy instances is the same (m = 14 | f =12). Conversely, in the hard groups, female patients have a greater presence (m = 8 | f = 18 and m = 9 | f = 17). It appears that accurately classifying women poses a greater challenge in both classes. Research indicates a sex-based difference in immunological responses to COVID-19 [12]. Female patients exhibit a more robust immunological response, potentially contributing to the difficulty in predicting their outcomes. This heightened cellular response may be a gender-specific reaction rather than a clear indication of greater infection-related impairment. Mukherjee and Pahan [12] also highlight societal expectations that sometimes discourage men from actively seeking medical assistance or consulting physicians. Consequently, men often delay seeking treatment, potentially making it easier to accurately classify them as their body's response has already progressed further.

3.2 Meta-features distribution

Now we move to a second part of our analysis, in order to explain how ML models made their decisions. Since models had access only to test results of the first day, they took patients' initial condition to make a prediction. Each model has a different approach to identify data patterns, this approach usually involves complex mathematical formulation. Our proposal is to analyze data distribution, inside our four groups, through meta-features values. These meta-features, already established in the literature, are more intuitive and together can offer a broad comprehension of data characteristics that are influencing models' results.

From a set of 13 original meta-features [14] we only look at the three that, in this dataset, presented the greatest correlation with the IH. Each of these hardness measures provides a different perspective on the data, offering insights into why certain instances may be difficult to classify. Figure 3 presents the distribution of these three meta-features according to class and level of hardness.

3.2.1 Class Likelihood: In our analysis, we observed that easy instances generally exhibit low *CL* values, as depicted in the top portion of Figure 3. This is in line with our expectations, as easy instances are more confidently classified by the model. However, hard instances, particularly those in the non-severe class, exhibit a wide range of *CL* values. This indicates that some hard instances have a moderate probability of belonging to their assigned class. However, most part of those instances that present low values of *CL* and are still hard to classify show high values of *DCP*. This is an example of how the meta-features can show different reasons why instances are hard to classify. The correlation between *CL* and instance hardness is 0.685.

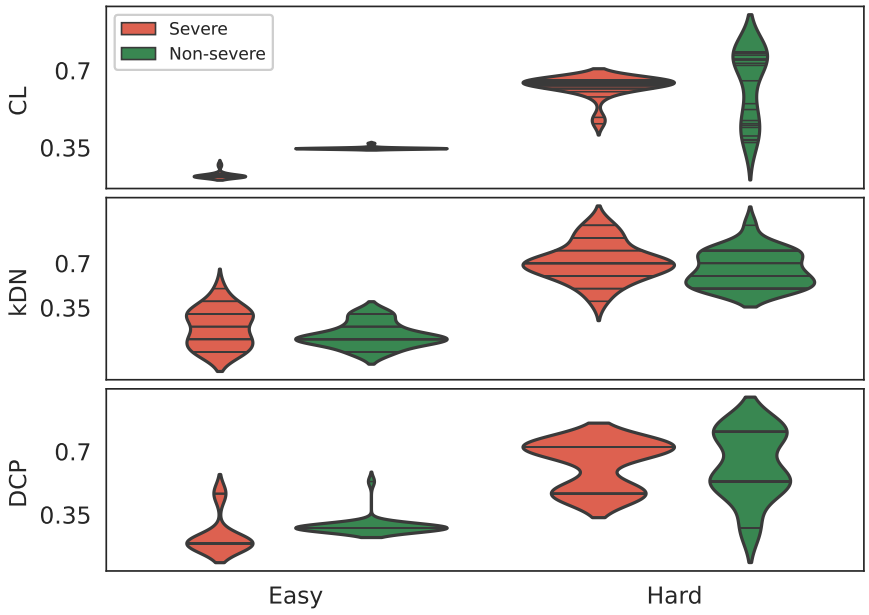
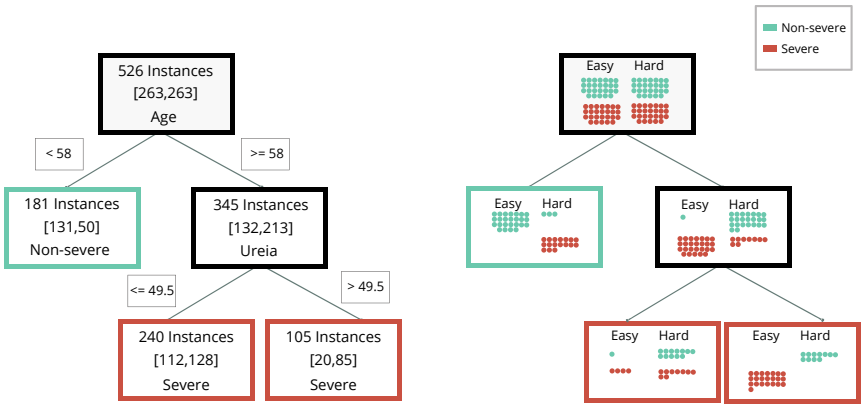


Fig. 3. Distribution of meta-features: *CL*, *kDN* and *DCP* in the four groups of selected instances.

3.2.2 k-disagreeing neighbors: Analyzing feature *kDN* values in the middle of Figure 3, we can see that few instances (in both classes) have zero disagreeing neighbours. This is an evidence of overlapping between classes, since even the easiest instances have neighbours of the opposite class. However, while hard instances overlap the opposite class in central values, easy instances overlap in a peripheral interval. Figure 9 in the supplementary material shows how the overlapping interval changes between easy and hard instances. The correlation between *kDN* and instance hardness is 0.638.

3.2.3 Disjunct class percentage: This meta-feature has the strongest correlation with IH (0.710). In the bottom image of Figure 3 is possible to see that easy instances are classified with few members of the opposite class in the same node. On the other hand, hard instances are classified inside a node with instances mainly belonging to the other class. Figure 4a presents the decision tree built in order to extract *DCP* values. Two features were used to split instances and the tree has three leaf nodes. In Figure 4(b) we show the distributions of our four groups inside the tree. Easy instances are mainly placed together, in leaf nodes corresponding to their own class. Conversely, hard instances are frequently missclassified, and placed in a disjunct with more instances of the opposite class than of their own class.

Now we move to the last step of our analysis, where we use insights from both feature distributions and meta-features values to analyse individual instances that have been classified as hard.



(a) Features and thresholds chosen to split nodes.

(b) Hard and easy distributions inside the DT.

Fig. 4. Decision tree generated to calculated feature *DCP* measures.

3.3 Patient level

Within our hard groups, three instances are surrounded by instances of the opposite class, a phenomenon that intrigues us. To understand why these instances belong to the opposite class of all their neighbors, we singled out them for closer analysis. These particular instances exhibit the maximum value of the meta-feature *kDN*. We have labeled them as *Outlier 1*, *Outlier 2*, and *Outlier 3*. Figure 5 presents a comprehensive summary of all the information, including features and meta-features, for each of these instances, and subsequently, we delve into the analysis of each instance individually.

- *Outlier 1:* This instance belongs to a 49-year-old severe male patient. The instance has a high value of *DCP*, since it is incorrectly classified at the first level of the decision tree, with a large mixture of classes, as shown at the bottom of Figure 5. Additionally, the instance exhibits low values of CRP and Urea, which are associated with the non-severe class. This discrepancy between the feature values and the assigned class is reflected in the high value of *CL*. At the time of hospitalization, the patient's initial condition did not indicate an aggravated condition. However, the level of CRP rose quickly in the next couple of days, at the same time the level of lymphocytes decreases, leading to a prolonged hospitalization until recovery. The patient was released after 17 days of hospitalization.
- *Outlier 2:* This outlier corresponds to a 63-year-old male patient that belongs to the severe class. The classification of this instance is accurate at the second level of the decision tree, resulting in intermediate values of *DCP*. However, the instance exhibits low values of Urea and CRP, which are typically associated with non-severe instances. This discrepancy contributes to the high value of *CL*. It is worth noting that there is no available information or records in the database for this patient for the last two days of hospitalization, despite registering a total stay of 15 days. Given this context, we can speculate that the patient may have been released on the 13^{th} day, potentially leading to a change in the class label from severe to non-severe.

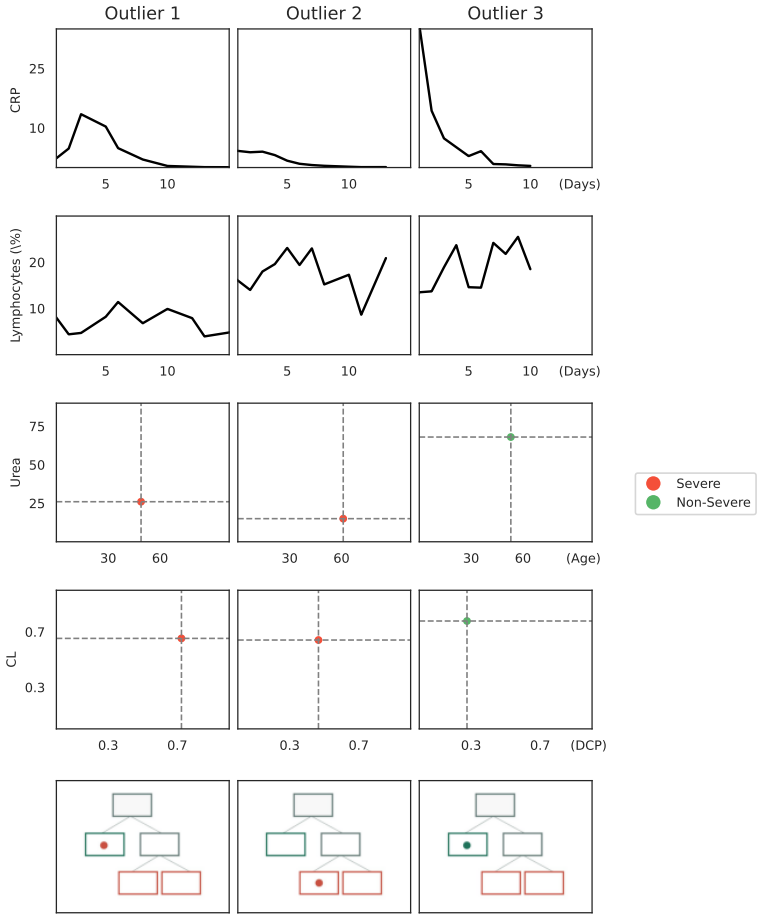


Fig. 5. The top two rows illustrate the progression of C-reactive protein (CRP) and lymphocyte percentage (LYM %) throughout the patients' hospitalization. The third row presents the levels of urea and the age of patients at the time of hospitalization. In the fourth row, meta-feature values are presented. Finally, the last row represents the DT generated to calculate *DCP* values and shows in which leaf node the instance was placed. It is noteworthy that all three instances exhibit $kDN = 1$.

- *Outlier 3:* This outlier belongs to the non-severe class and is a 53-year-old male patient. The classification at the first level of the decision tree is correct, with mainly instances of the same class. However, the instance exhibits a pattern of feature values that are typically associated with the severe class. Specifically, the patient has intermediate values of Lymphocytes, while CRP and Urea are high. This discrepancy between the feature values and the assigned class contributes to the high value of *CL*. The patient's condition improves in the following days, leading to a release after 11 days of hospitalization.

4 DISCUSSION

In this paper, we analyzed a Brazilian COVID-19 dataset that was assembled to predict a serious condition, defined as "death or hospitalization equal to or longer than 14 days". Our aim was to demonstrate how analyzing hard instances

can shed light on issues within a dataset. We adopted the concept of instance hardness as the likelihood of an instance being misclassified by a set of different ML algorithms. By selecting the hardest and easiest instances, we inspected the distributions of certain features and extracted a set of meta-features to describe each instance. We closely analyzed three meta-features that were highly correlated with IH and compared their values between easy and hard instances in both classes. Furthermore, we individually examined certain instances based on their features and meta-feature values.

In studies with a selected population, the influence of the disease on the patient's health condition can be more rigorously analysed. Our work, on the other hand, are based on an anonymized source of observational data, where confounding factors may be present. Adopting this source of data we can rely on a greater number of instances, usually larger than is possible to do in retrospective and follow-up studies. Routinely collected data can be a great source of information to feed ML models in order to improve health systems. However, since data was not collected for the purpose of a study, the interaction with the data expert is even more important in designing problems and deploying models.

Our analyses offered some interesting explanations for why certain instances were challenging to classify, enabling a better understanding of both the successes and failures of the models by considering the distributions of feature and meta-feature values. This opens up the possibility for discussions between data scientists and domain specialists. The question of how to address hard instances is domain-specific, and it is important to provide the specialist with helpful support to determine if these instances are so atypical that they should be removed from the dataset or whether they can be subject to correction.

Our research is centered on numerical data; however, the same analysis principles can be applied to different data types. Although the library used for measuring Instance Hardness and meta-feature values is designed exclusively for numerical data, it is possible to convert data in other formats to numerical representations. This strategy has previously been employed with image data [11], using a pre-trained neural network as a numerical feature extractor, and can similarly be applied to analyze text data. It is important to note that this technique may result in lower interpretability, as numerical features may not directly correspond to real-world characteristics. In terms of data types, our study focused on medical data. We believe that the medical field is an excellent context for such investigations, given the significance of visualization procedures in high-stakes areas [15]. Nevertheless, the analysis framework can be extended to support human-AI development in various other domains.

Naturally, there are limitations to our approach, and we would like to highlight two main ones. First, we examined only a few features and meta-features from small groups of instances, which undoubtedly do not encompass all data issues. Although the choice of discussing only three instances may be seen as limited, we intended to exemplify how the knowledge we obtained can be applied to individual instances, particularly when a discussion with the domain expert is needed. With our contribution, we offer valuable insights and data perspectives that serve as a starting point for establishing a dialogue between the data science team building the ML models and the domain experts who will use these models in the future. While the information we have acquired may not be complete, it is still interesting and useful.

Secondly, a limitation relates to the applicability of this approach to new instances. Since our analysis here was based on the ground truth labels, as IH and meta-feature values rely on this information, it would not be possible to conduct a similar analysis for new instances encountered by the models. However, our intention here is to take a step back and facilitate the understanding of the models' decision-making and their deployment. In the future, an adapted version of meta-features that is not reliant on true labels can be proposed to investigate model decisions for unlabeled instances.

ACKNOWLEDGMENTS

This work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – Financing Code 001. This work is supported by CAPES under Grant No.: 88887.507037/2020-00 and by FAPESP under Grant No.: 2021/06870-3. The authors thank FAPESP for the COVID data-sharing initiative.

REFERENCES

[1] José LM Arruda, Ricardo BC Prudêncio, and Ana C Lorena. 2020. Measuring instance hardness using data complexity measures. In *Brazilian Conference on Intelligent Systems*. Springer, 483–497.

[2] André Calero Valdez, Martina Ziefle, Katrien Verbert, Alexander Felfernig, and Andreas Holzinger. 2016. Recommender systems for health informatics: state-of-the-art and future perspectives. In *Machine learning for health informatics*. Springer, 391–414.

[3] Angelos Chatzimparmpas, Fernando V Paulovich, and Andreas Kerren. 2022. HardVis: Visual Analytics to Handle Instance Hardness Using Undersampling and Oversampling Techniques. *arXiv preprint arXiv:2203.15753* (2022).

[4] Guang Chen, DI Wu, Wei Guo, Yong Cao, Da Huang, Hongwu Wang, Tao Wang, Xiaoyun Zhang, Huilong Chen, Haijing Yu, et al. 2020. Clinical and immunological features of severe and moderate coronavirus disease 2019. *The Journal of clinical investigation* 130, 5 (2020), 2620–2629.

[5] Alexander Decruyenaere, Philippe Decruyenaere, Patrick Peeters, Frank Vermassen, Tom Dhaene, and Ivo Couckuyt. 2015. Prediction of delayed graft function after kidney transplantation: comparison between logistic regression and machine learning methods. *BMC medical informatics and decision making* 15 (2015), 1–10.

[6] Menglu Gao, Qianying Wang, Jianhao Wei, Zhaoqin Zhu, and Haicong Li. 2020. Severe Coronavirus disease 2019 pneumonia patients showed signs of aggravated renal impairment. *Journal of Clinical Laboratory Analysis* 34, 10 (2020), e23535.

[7] Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3, 2 (2016), 119–131.

[8] Andrew Houston, Georgina Cosma, Phillipa Turner, and Alexander Bennett. 2021. Predicting surgical outcomes for chronic exertional compartment syndrome using a machine learning framework with embedded trust by interrogation strategies. *Scientific Reports* 11, 1 (2021), 1–15.

[9] Grey Leonard, Charles South, Courtney Balentine, Matthew Porembka, John Mansour, Sam Wang, Adam Yopp, Patricio Polanco, Herbert Zeh, and Mathew Augustine. 2022. Machine learning improves prediction over logistic regression on resected colon cancer patients. *Journal of Surgical Research* 275 (2022), 181–193.

[10] Jing Li, Yinghua Zhang, Fang Wang, Bing Liu, Hui Li, Guodong Tang, Zhigang Chang, Aihua Liu, Chunyi Fu, Jing Gao, et al. 2020. Sex differences in clinical findings among patients with coronavirus disease 2019 (COVID-19) and severe condition. *MedRxiv* (2020), 2020–02.

[11] Camila Castro Moreno, Pedro Yuri Arbs Paiva, Gustavo H Nunes, and Ana Carolina Lorena. 2021. Contrasting the profiles of easy and hard observations in a dataset. In *Proc. NeurIPS DCAI Workshop*.

[12] Shreya Mukherjee and Kalipada Pahan. 2021. Is COVID-19 gender-sensitive? *Journal of Neuroimmune Pharmacology* 16 (2021), 38–47.

[13] Fesih Ok, Omer Erdogan, Emrullah Durmus, Serkan Carkci, and Aggul Canik. 2021. Predictive values of blood urea nitrogen/creatinine ratio and other routine blood parameters on disease severity and survival of COVID-19 patients. *Journal of medical virology* 93, 2 (2021), 786–793.

[14] Pedro Yuri Arbs Paiva, Camila Castro Moreno, Kate Smith-Miles, Maria Gabriela Valeriano, and Ana Carolina Lorena. 2022. Relating instance hardness to classification performance in a dataset: a visual approach. *Machine Learning* (2022), 1–39.

[15] Markus Plass, Michaela Kargl, Patrick Nitsche, Emilian Jungwirth, Andreas Holzinger, and Heimo Müller. 2022. Understanding and Explaining Diagnostic Paths: Toward Augmented Decision Making. *IEEE Computer Graphics and Applications* 42, 6 (2022), 47–57.

[16] Nabeel Seedat, Jonathan Crabbé, Ioana Bica, and Mihaela van der Schaar. 2022. Data-IQ: Characterizing subgroups with heterogeneous outcomes in tabular data. *arXiv preprint arXiv:2210.13043* (2022).

[17] Milad Sharifpour, Srikant Rangaraju, Michael Liu, Darwish Alabyad, Fadi B Nahab, Christina M Creel-Bulos, Craig S Jabaley, and Emory COVID-19 Quality & Clinical Research Collaborative. 2020. C-Reactive protein as a prognostic indicator in hospitalized patients with COVID-19. *PloS one* 15, 11 (2020), e0242400.

[18] Anil Shrestha, Gaurav Jung Shah, Sagar Neupane, and Richa Shrestha. 2022. C-Reactive Protein as a Prognostic Marker in Hospitalized Patients with COVID-19. *Journal of Nepalgunj Medical College* 20, 1 (2022), 70–73.

[19] Michael R Smith, Tony Martinez, and Christophe Giraud-Carrier. 2014. An instance level analysis of data complexity. *Machine learning* 95, 2 (2014), 225–256.

[20] Herdiantri Sufriyana, Atina Husnayain, Ya-Lin Chen, Chao-Yang Kuo, Onkar Singh, Tso-Yang Yeh, Yu-Wei Wu, Emily Chia-Yu Su, et al. 2020. Comparison of multivariable logistic regression and other machine learning algorithms for prognostic prediction studies in pregnancy care: systematic review and meta-analysis. *JMIR medical informatics* 8, 11 (2020), e16503.

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

[22] Matthew Zirui Tay, Chek Meng Poh, Laurent Rénia, Paul A MacAry, and Lisa FP Ng. 2020. The trinity of COVID-19: immunity, inflammation and intervention. *Nature Reviews Immunology* 20, 6 (2020), 363–374.

[23] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*. PMLR, 359–380.

[24] Maria Gabriela Valeriano, Carlos RV Kiffer, Giane Higino, Paloma Zanão, Dulce A Barbosa, Patrícia A Moreira, Paulo Caleb JL Santos, Renato Grinbaum, and Ana Carolina Lorena. 2022. Let the data speak: analysing data from multiple health centers of the São Paulo metropolitan area for COVID-19 clinical deterioration prediction. In *2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*. IEEE, 948–951.

[25] Maria Gabriela Valeriano, Carlos Roberto Veiga Kiffer, and Ana Carolina Lorena. 2022. Supporting Decision Making in Health Scenarios with Machine Learning Models. In *Anais do simposio brasileiro de pesquisa operacional*. https://proceedings.science/sbpo/sbpo-2022/trabalhos/supporting-decision-making-in-health-scenarios-with-machine-learning-models?lang=pt-br

A ALGORITHMS ADOPTED AND THEIR PERFORMANCE

To assess Instance Hardness levels we adopted seven ML techniques, with different approaches to classify data. Next, Table 3 presents models performance when adopting, or not, a subsampling approach. The values reported are the mean among a 5-fold cross-validation loop.

Table 3. Mean performance and standard deviation of ML models across the 5-fold cross-validation loop. SVM refers to Support Vector Machines that were adopted with Linear and RBF kernels. MLP refers to Multilayer Perceptron.

	Imbalanced			After subsampling		
Algorithm	AUC	Sensitivity	Specificity	AUC	Sensitivity	Specificity
Linear SVM	0.716 ± 0.026	0.190 ± 0.033	0.960 ± 0.020	0.724 ± 0.020	0.670 ± 0.055	0.643 ± 0.069
RBF SVM	0.712 ± 0.025	0.243 ± 0.065	0.960 ± 0.020	0.742 ± 0.021	0.685 ± 0.075	0.658 ± 0.082
Random Forest	0.762 ± 0.013	0.349 ± 0.083	0.908 ± 0.024	0.734 ± 0.010	0.628 ± 0.076	0.658 ± 0.090
Gradient Boosting	0.734 ± 0.021	0.357 ± 0.083	0.889 ± 0.016	0.721 ± 0.038	0.635 ± 0.065	0.647 ± 0.073
Logistic regression	0.734 ± 0.024	0.296 ± 0.064	0.903 ± 0.035	0.731 ± 0.018	0.654 ± 0.057	0.670 ± 0.070
Bagging	0.712 ± 0.025	0.308 ± 0.059	0.879 ± 0.022	0.664 ± 0.034	0.476 ± 0.089	0.742 ± 0.075
MLP	0.703 ± 0.048	0.399 ± 0.058	0.860 ± 0.035	0.724 ± 0.014	0.643 ± 0.089	0.654 ± 0.074
mean	0.725	0.306	0.909	0.719	0.627	0.667

B INSTANCE HARDNESS AND META-FEATURES

B.1 Instance Hardness values

The Instance Hardness level was assessed for all instances in the dataset. Figure 6 presents the distribution of IH values according to the original class.

Table 4 presents the mean and standard deviation values of IH inside the four groups of selected instances.

Table 4. IH mean and stdv for each group.

	Severe		Non-Severe	
Instance Hardness	Hard	Easy	Hard	Easy
Mean	0.665	0.242	0.658	0.241
stdev	0.044	0.023	0.032	0.028

B.2 Complete description of meta-features employed

In total 13 meta-features were measured. In the main article we explained three of them. Next we give a brief explanation of the other ones.

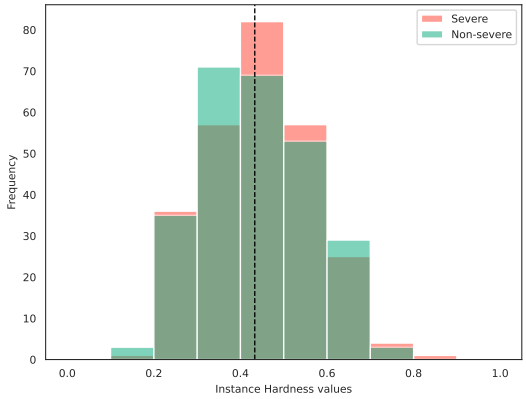


Fig. 6. Distribution of Instance Hardness for severe and non-severe patients. Regions of overlap of severe and non-severe values are in dark green.

Tree Depth $TD(\mathbf{x}_i)$: also builds a DT using D and calculates the ratio between the depth of the leaf node where $\mathbf{x}_i$ is classified and the maximum tree depth. Easier instances will tend to present low values of TD because they are classified early in the DT. There are two versions of this measure, using pruned ($TD_P(\mathbf{x}_i)$) and unpruned ($TD_U(\mathbf{x}_i)$) decision trees.

Class Likelihood Difference $CLD(\mathbf{x}_i)$: measures the difference between the likelihood that $\mathbf{x}_i$ belongs to its class and the maximum likelihood it has to belong to any of the other classes. The likelihood is calculated in the same way as described in CL. The complement of this measure is taken as an output. When the difference of likelihoods is low, there is doubt on the label of the instance and the CLD value is higher, which is more difficult to classify (specifically between the two classes involved in the likelihood difference computation).

Fraction of features in overlapping areas $F1(\mathbf{x}_i)$: calculates the percentage of features of $\mathbf{x}_i$ with values in the overlapping region of the classes. The overlapping region is the interval within which we can find feature values from more than one class. A low percentage means that the instances are easily distinguished from the instances of the other classes according to the input features' values.

Fraction of nearby instances of different classes $N1(\mathbf{x}_i)$: fits a Minimum Spanning Tree (MST) with the dataset observations. An MST is an acyclic graph with weighted edges and all vertices connected. The vertices are always connected to result in the lower sum of edges weights. The weights here are given by the pairwise distances between different instances. The $N1(\mathbf{x}_i)$ returns the number of instances from a different class that are connected to $\mathbf{x}_i$ in the MST, normalized by the total number of instances connected to $\mathbf{x}_i$ in the MST. Easier instances will have more connections with observations of the same class in the MST, resulting in lower $N1$ values.

Ratio of the intra-class and extra-class distances $N2(\mathbf{x}_i)$: takes the ratio between the distance of the nearest neighbor of the same class of $\mathbf{x}_i$, and the nearest enemy, which is the nearest neighbor from a different class. Next, the complement is taken. The idea is that instances closer to their nearest enemies than to their nearest neighbor from the same class are more difficult to classify.

Local Set Cardinality $LSC(\mathbf{x}_i)$: the local set of an instance $\mathbf{x}_i$ is composed of the instances in D that are closer to $\mathbf{x}_i$ than its nearest enemy (ne - nearest neighbor from another class), normalized by the total number of instances that belong to the class y_i. Easier instances will tend to present a large local set cardinally, which means that the nearest enemy is distant and the instance is surrounded by many others sharing its class. The complement of this measure is taken as an output.

Local Set Radius $LSR(\mathbf{x}_i)$: takes the radius of the local set of $\mathbf{x}_i$, normalized by the distance between $\mathbf{x}_i$ and the farthest instance of the same class as $\mathbf{x}_i$. The complement of this measure is taken as an output. LSR presents low values when $\mathbf{x}_i$ has a distant nearest enemy and all the instances belonging to the same class of $\mathbf{x}_i$ are closer to it.

Usefulness $U(\mathbf{x}_i)$: refers to the number of instances having $\mathbf{x}_i$ in their local sets. This value is normalized by the number of instances from the same class of $\mathbf{x}_i$, except from it. The idea is that an easy instance possesses many examples from its class in their neighborhood, being more useful. The complement of this measure is taken as an output.

Harmfulness $H(\mathbf{x}_i)$: is the fraction of instances having $\mathbf{x}_i$ as their nearest enemy. If $\mathbf{x}_i$ is easy to classify, it will be far from instances from another class and consequently it will not be the nearest enemy of many instances.

B.3 Correlation between Instance Hardness and meta-features values

In order to choose a small set of meta-features to explore we performed a correlation between the level of instance hardness and the meta-feature values. Table 5 presents these results.

Table 5. Frequency of Special Characters

Meta-feature	Correlation with Intance Hardness
Disjunct class percentage	0.710
Class likelihood	0.685
Class likelihood difference	0.685
k-Disagreeing neighbours	0.638
Local set cardinality	0.441
Harmfulness	0.419
Ratio of the intra-class and extra-class distances	0.402
Usefulness	0.391
Fraction of nearby instances of different classes	0.357
Local set radius	0.258
Tree depth	0.174
Fraction of features in overlapping areas	0.043

B.4 Graphical representation of meta-features adopted

For the three meta-features explored we offer a graphical representation of the measures, aiming to facilitate their understanding by data experts.

B.5 Overlap between classes for easy and hard instances

Figure 9 shows how the overlap between classes is different for easy and hard instances. Although in both cases there are feature overlapping, in hard instances it occurs in central regions of the distribution. Conversely, easy instances overlap in peripheral intervals.

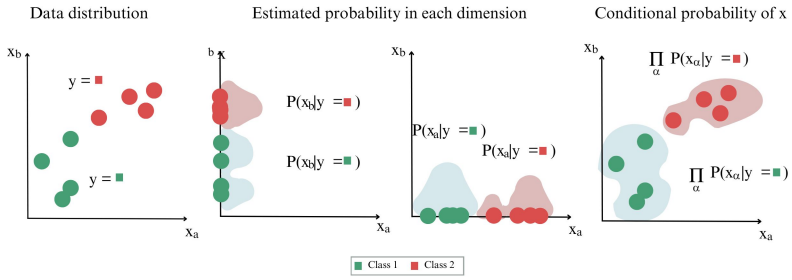
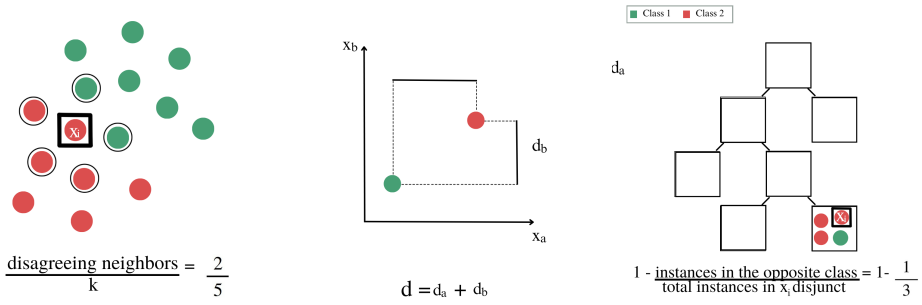


Fig. 7. Visual representation of feature *CL*.



(a) Feature kDN. (b) Manhattan distance. (c) Feature DCP.

Fig. 8. Visual representation of feature *kDN*, Manhattan distance and feature *DCP*.

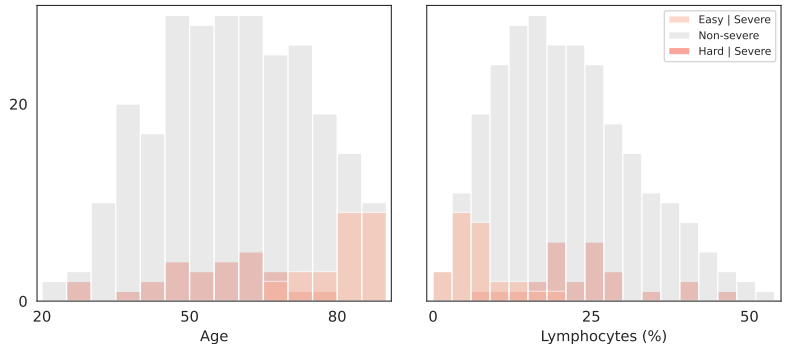


Fig. 9. Overlap between classes in easy and hard instances for features Age and lymphocytes (%).