Everything is Varied: The Surprising Impact of Instantial Variation on ML Reliability

Andrea Campagner[a], Lorenzo Famiglini[b], Anna Carobene[c], Federico Cabitza[a,b]

[a] *IRCCS Istituto Ortopedico Galeazzi, Milano, Italy*
[b] *Dipartimento di Informatica, Sistemistica e Comunicazione, University of Milano-Bicocca, Milano, Italy*
[c] *IRCCS Ospedale San Raffaele, Milano, Italy*

Abstract

Instantial variation (IV) refers to variation that is due not to population differences or errors, but rather to within-subject variation, that is the intrinsic and characteristic patterns of variation pertaining to a given instance or the measurement process. Although taking into account IV is critical for the proper analysis of the results, this source of uncertainty and its impact on robustness have so far been neglected in Machine Learning (ML). To fill this gap, we look at how IV affects ML performance and generalization, and how its impact can be mitigated. Specifically, we provide a methodological contribution to formalize the problem of IV in the statistical learning framework. To prove the relevance of our contribution, we focus on one of the most critical domains, healthcare, and take individual (analytical and biological) variation as a specific kind of IV; in this domain, we use one of the largest real-world laboratory medicine datasets for the task of COVID-19 detection, to show that: 1) common state-of-the-art ML models are severely impacted by the presence of IV in data; and 2) advanced learning strategies, based on data augmentation and soft computing methods (data imprecisiation), and proper study designs can be effective at improving robustness to IV. Our findings demonstrate the critical relevance of correctly accounting for IV to enable safe deployment of ML in real-world settings.

Keywords: Instantial Variation, Uncertainty, Robustness, Medical Machine Learning, Soft Computing

1. Introduction

In recent years, the interest toward the application of Machine Learning (ML) methods and systems to the development of decision support systems has been steadily increasing.This interest has been mainly driven by the promising

Email address: a.campagner@campus.unimib.it (Andrea Campagner)

results obtained and reported by these systems in academic research for different tasks [1, 2, 3, 4]. Despite these promising results, the adoption of ML-based systems in real-world critical settings has been lagging behind [5], with these systems often failing to meet the expectations and requirements needed for safe deployment [6, 7], a concept that has been termed the *last mile of implementation* [8]. While reasons behind the gaps in this "last mile" are numerous, among them we recall the inability of ML systems to reliably generalize in new contexts and settings [9, 10], as well as their lack of robustness to variation in data, leading to poorer performance in real settings [11] and, ultimately, to what has been called the *replication crisis* of ML [12].

In the ML literature, the notion of variation has usually been associated with variance of the population data distribution, that is, with the variance of the reference population, or of smaller samples taken from this latter, due to the presence of outliers or anomalies [13], out-of-distribution instances [14, 15], evolvable feature sets [16] or concept/co-variate shifts and drifts [17, 18]. While these forms of variation are certainly relevant, however they are not the only ones that can arise in real-world settings. Indeed, another source of variation in data is the so-called *instantial variation* (IV) [19]: this denotes variation that is not due to population differences or errors [20], but rather to the intrinsic and characteristic (that is, individual) patterns of variation pertaining to single instances, that is to within-subject variation [21] that can affect both the target as well as the features[1]. Such patterns of variations arise from the complex interplay between two distinct but intrinsically inseparable forms of variation, namely *measuring variation* (i.e., exogenous variation that is related to the measurement instrument and procedure) and *measurand variation* (i.e., endogenous variation that is intrinsic to the measured instances and features[2]). This complex interplay makes dealing with IV particularly difficult: indeed, while many works have focused on how to account for target variability and uncertainty in the labels (e.g., [22, 23, 24, 25]), little research has been so far devoted to IV affecting the data features. To put it in somewhat extreme but frank terms, it is as if one of the most important assumptions of supervised ML methodologies (albeit one of the most neglected) regarded the *invariance of the phenomena and objects to be classified*. However, the feature values of an

[1]The term 'within-instance variation' refers to the variation of data points within a single instance or group of instances, while 'between-instance variation' refers to the variation between different instances or groups of instances. In other words, within-instance variation measures how much each instance varies internally, while between-instance variation measures how much each instance differs from other instances.

[2]In many cases, what we call an instance, or object, is the perceivable manifestation and expression of a set of highly complex and tightly coupled processes (as an example think of biological entities -such as plants and animals- which are systems comprising numerous apparatuses that interact with each other in complex and continuous patterns of exchange of energy, matter and information). In a sense, what we can call "instantiality" is an emergent property that depends on the level of analysis, description and relationship, and on the characteristics (such as "time constants", dimensions, perceptual capabilities) of an external observer.

object *rarely* do not change over time, or are measured in the same way and observed in the same conditions [26]. This discrepancy is particularly evident in medical contexts and in laboratory data or other physiological signals and biomarkers, and more generally in every phenomenon whose manifestations can exhibit time-varying patterns.

In this article, we will focus on these latter settings as prime examples of the problem of instantial variation, as well as of the potential impact of this issue on the development of technological supports for real-world decision making: thus this article will describe our research as a case study and hopefully as a source of ideas about instantial variation that can be used in different settings, including those that present less critical issues or have less impact on the lives and well-being of individuals.

In medical settings, in particular, IV has been studied under the name of individual variation, and the two related forms of measurand and measuring variation manifest themselves in the form of, respectively, *biological variation* (BV) [21], i.e. the intrinsic distribution of feature values for a given subject or patient, and *analytical variation* (AV), i.e. the variation in the measurement process and instrument itself. The presence of IV entails [27] that for each individual one can identify a "subject average" or central tendency (*homeostatic point*) arising from such factors as personal characteristic of the individuals themselves (e.g., genetic characteristics, age, phenotypic elements such as diet and physical activity) or of the measurement instrument (e.g., calibration), and a distribution of possible values, whose uncertainty is represented by the extent of the IV: crucially, only a snapshot (i.e., a sample) from this instantial distribution can be accessed at any moment.

While the potential impact of IV on computer-supported diagnosis has been known for a while [20] (for instance, in [28] authors reported that "computer interpretations of electrocardiograms recorded 1 minute apart were significantly (grossly) different in 4 of 10 cases"[3]), only conjectures have so far been produced to estimate its extent. Nonetheless, IV has two strong implications for ML applications. First, ML models trained on data affected by IV, even highly accurate ones, can fail to be robust and properly generalize not only to new patients, but also to the same patients observed in slightly different conditions: for example, an healthy patient could indeed be classified as healthy with respect to the features actually observed for them, while they could have been classified as non-healthy for a slightly different set of feature values, which nevertheless

[3]This simple result, if only for its evocative power, should not be underestimated and indeed inspire similar controls in SOTA models in the automated diagnosis based on biological signals and biomarkers. By assuming ML models sufficiently generalizable, such a result might suggest to us that the nature of certain instances changes over time so quickly that deciding on the basis of a picture taken at T_0 might lead to very different conclusions if instead that picture were taken at T_1, a short time later. The assumption of constancy and low time-variance could be wrong for many domains where things appear to be much more stable and regular than they actually are.

would still be totally compatible with the distribution due to IV[4]. Second, differently from distribution-related variation, collecting additional data samples, which has been considered a primary factor in the continued improvement of ML systems, can help only marginally in reducing the impact of IV [20], unless specific study designs are adopted that allow to capture multiple observations for each individuals across time [30, 31].

Despite these apparently relevant characteristics, the phenomenon of IV has been so far largely overlooked in the ML literature: On a superficial analysis, the two components of IV could remind of other sources of uncertainty. For instance, AV (and, by extension, measuring variation) could be considered assimilable to "attribute noise", which has been widely studied in machine learning [32]. However, AV and attribute noise represent two intrinsically different notions, in that attribute noise is usually interpreted as the result of a measurement error, while AV is an intrinsic pattern of variation that is characteristic of both the given measurement instrument used and the phenomenon of interest. For this reason, to our knowledge no previous work has really investigated the impact of IV on ML systems, nor has proposed viable techniques to improve robustness and manage this source of perturbations.

In this article, we thus attempt to fill the above-mentioned gaps in the specialized literature. To this aim, this paper will consist of three parts. In the first part we will address the theoretical structure of the problem of learning from data affected by IV, by proposing a generalization of the statistical learning theoretic framework to this setting. The second part will focus on the research question "can instantial variation significantly affect the accuracy, and hence the robustness, of a machine model on a diagnostic task grounding on laboratory medicine data" (H_1)? Due to the pervasiveness of IV, proving this hypothesis could suggest that most ML models could be seriously affected by lack of robustness on real-world and external data. To this aim, we will apply an expertise-grounded, generative model to simulate the effects of IV on data, and we will show how commonly used classes of ML models fail to be robust against it. More in particular, to provide a more self-contained and detailed discussion, we will focus our experimental analysis on a specific setting, the medical one, which is of particular relevance due to its critical characteristics as well as due to it being one of the fields of applications of ML in which the problem of IV has been more frequently acknowledged. Finally, the last part of the paper will aim to build on the rubble left by the first part, and it will focus on the hypothesis whether more advanced learning and regularization methods (grounding on, either, data augmentation [33] or data imprecisiation [34]) will achieve increased robustness in face of the same perturbations (H_2). To address these two research questions, and motivated by the lack of datasets that rep-

[4]As we show in the following, this setting is a generalization of the usual one adopted in ML theory [29]: not only we assume that the best model could have less than perfect accuracy, but we also assume that any instance is represented as a distribution of vectors possibly lying in opposite sides of the decision boundary.

resent and allow to investigate this complex form of uncertainty, we will rely on a large gold-standard medical dataset that had been proposed for the task of COVID-19 diagnosis, a major impactful concern, which was specifically constructed with the help of clinical laboratory medicine to study IV, grounding on previous knowledge in this domain [30, 35, 36]. A graphical summary of the role and impact of IV in ML, as well as of the general structure of this article, is given in Figure 1.

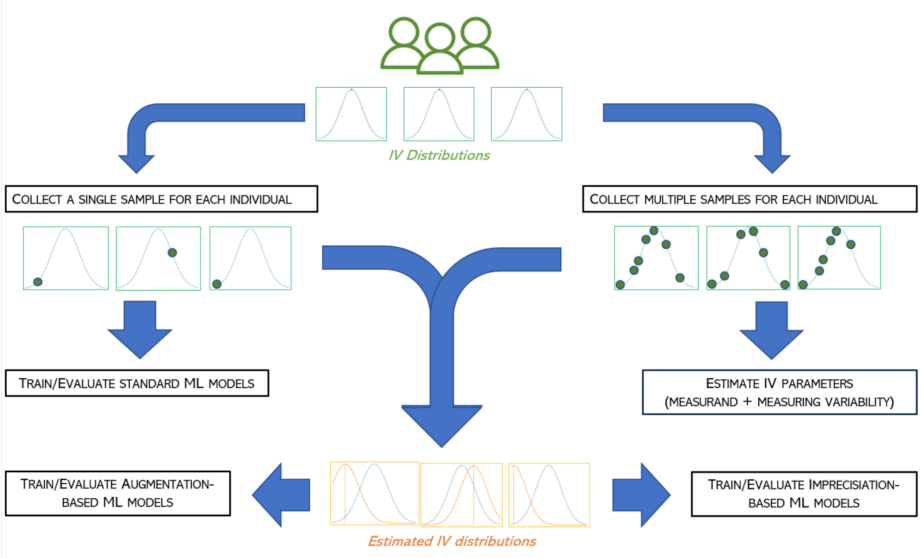

Figure 1: A graphical representation of the problem of instantial variation (IV) in ML. Each instance (represented here as a stylized person) can be associated with a distribution, which describes the uncertainty about the data for that instance due to IV. Then, since the above mentioned distributions are unknown, in any experimental setting and for each individual instance, we can either collect multiple samples (as shown in the rightmost part of figure) or just a single sample (as shown in the leftmost part of the figure, and as is typically done in ML studies, see also Section 3.2). Collecting multiple samples for each individual instance allows to estimate the extent of IV and its components measurand and measuring variabilities (see Section 2.1). The single samples, along with the estimated IV parameters, can be used to obtain an empirical approximation of the unknown IV distribution for each patient, which can then be used to train or evaluate data augmentation-based or imprecisiation-based ML models, such as those described in Section 3.3.

2. Background

As discussed in the previous section, the aim of this article is to evaluate and address the potential impact of IV on ML models' robustness. In this section, we first provide basic background on IV, its importance in clinical settings, and methods to compute it.

2.1. Instantial Variation in Medical Data

IV is considered one of the most important sources of uncertainty in clinical data [21] and recent research has highlighted the need to take IV properly into account in any use of medical data [27, 37] . IV can be understood as encompassing three main components: pre-analytical variation, analytical variation and (within-subject) biological variation [21]. Pre-analytical variation denotes uncertainty due to patients' preparation (e.g., fasting, physical activity, use of medicaments) or sample management (including, collection, transport, storage and treatment) [38]; it is usually understood that pre-analytic variation can be controlled by means of careful laboratory practice [19]. AV, by contrast, describes the un-eliminable uncertainty which is inherent to every measurement technique, and is characterized by both a random component (i.e., variance, that is the agreement between consecutive measurements taken with the same instrument); and a systematic component (i.e., bias, that is the differences in values reported by two different measurement instruments). Finally, BV describes the uncertainty arising from the fact that features or biomarkers can change through time, contributing to a variance in outcomes from the same individual that is independent of other forms of variation.

As already mentioned, IV can influence the interpretation and analysis of any clinical data: for this reason, quantifying IV, also in terms of its components, is of critical importance. However collecting reliable data about IV is not an easy task [39, 40]. To this aim, standardized methodologies have recently been proposed [30, 31]: intuitively, IV can be estimated [41, 39, 42] by means of controlled experimental studies that monitor *reference individuals*[5] [43] by collecting multiple samples over time.

Formally speaking, let us assume that a given feature of interest x_i has been monitored in n patients for m time steps. At each time step, $k \quad 2$ repeated measurements should be performed, so as to determine the AV component of IV. Then, the IV of feature x_i, for patient p, is estimated as $IV_p(x_i) = StDev(x_i^p)$, while the AV component is defined as $AV_p(x_i) = \frac{1}{m}\sum_{t=1}^{m} StDev(x_i^p(t))$, where x_i^p denotes the collection of values of x_i for patient p, and $x_i^p(t)$ denotes the collection of values of x_i for patient p at the t-th time step. Then, the BV component of IV is computed as $BV_p(x_i) = \sqrt{IV_p(x_i)^2 \quad AV_p(x_i)^2}$. The overall variations IV, AV, BV, finally, can be computed as the average of the variations across the population of patients. While the above representation of IV, AV and BV is given in absolute terms, typically these quantities are express in relative (or, percent) terms, defining the so-called coefficients of individual (resp., analytical, biological) variation, that is $CVT(x_i) = \frac{IV(x_i)}{\overline{x_i}}$, $CVA(x_i) = \frac{AV(x_i)}{\overline{x_i}}$ and $CVI(x_i) = \frac{BV(x_i)}{\overline{x_i}}$, where $\overline{x_i}$ is the average of value of x_i across all patients and all m time steps. The value of CVT, can then be used to model the uncertainty about the observations obtained for any given patient p, for any

[5]The term reference individual denotes an individual that, for some reasons, can be considered representative of the population of interest (e.g., healthy patients).

given set of features $x^p = (x_1^p, ..., x_d^p)$, : indeed, any patient p, as a consequence of the uncertainty due to IV, can be represented by a d-dimensional Gaussian $N_p(\hat{x}^p, \ ^p)$, where $\hat{x}^p$ is a d-dimensional vector characteristic representation of patient p, called *value at the homeostatic point*, and ^{p} is the diagonal covariance matrix given by $^p_{i,i} = CVT(x_i) \ \hat{x}_i^p$. More generally, having observed a realization x^p of $N_p(\hat{x}^p, \ ^p)$ for patient p, its distribution can be estimated as $N_p(x^p, \ ^{-p})$, where $^{-p}_{i,i} = CVT(x_i) \ x_i^p$. We illustrate the computation of IV (and its components AV and BV) according to the above-mentioned procedure, along with the use of Gaussian distributions based on the IV as a way to model uncertainty, through the following example, which considers a simple clinical setting in which a single feature is monitored.

Example 1. *Assume that we want to compute the IV of a given parameter of interest (e.g., the White Blood Count, WBC) for a collection of two patients p_1, p_2. To this aim, the value of WBC has been measured across three different time steps (to estimate the BV component of IV), with two measurements per time step (to estimate the AV component of IV). A collection of such data, based on the dataset described in Section 3.3, is represented in Table 1: here the data for patient 1 was generated from a distribution for which $\hat{WBC}^{p_1} = 12.2$ (the value at the homeostatic point) and $IV_{p_1}(WBC) = \ ^{p_1} = 1$, while the data for patient 2 was generated from a distribution for which $\hat{WBC}^{p_2} = 6.6$ and $IV_{p_2}(WBC) = \ ^{p_2} = 1$.*

Patient	Time step	WBC	
		Observation 1	Observation 2
p_1	t_1	9.92	13.59
	t_2	11.83	12.45
	t_3	11.52	4.64
p_2	t_1	6.76	6.75
	t_2	8.55	7.19
	t_3	5.46	5.44

Table 1: An example dataset to illustrate the computation of individual variability.

For each patient p_i, we can compute the IV for the WBC parameter as $IV_{p_i}(WBC) = StDev(WBC^{p_i})$. Thus, $IV_{p_1} = 1.51$ and $IV_{p_2} = 1.06$. Similarly, we can then compute the AV for each patient as

$$AV_{p_i}(WBC) = \frac{StDev(WBC^{p_i}(t_1)) + StDev(WBC^{p_i}(t_2)) + StDev(WBC^{p_i}(t_3))}{3}.$$

Thus, $AV_{p_1}(WBC) = 1.23$, while $AV_{p_2}(WBC) = 0.23$ Finally, we can compute the BV, for each patient p_i, by applying the formula

$$BV_{p_i}(WBC) = \sqrt{IV_{p_i}(WBC) \quad AV_{p_i}(WBC)}.$$

Thus, $BV_{p_1}(WBC) = 0.87$ and $BV_{p_2}(WBC) = 1.07$. The average variations AV, IV and BV can then be computed as $IV = 1.27, AV = 0.73, BV = 0.95$:

since the mean value of WBC is $\overline{WBC} = 9.51$*, we can derive the coefficients of variation as* $CVT = 0.14, CVA = 0.08, CVI = 0.10$*. We can estimate the value of WBC at the homeostatic point for patient* p_1 *as* $\hat{WBC}^{p_1} = 12.33$*, and for* p_2 *as* $\hat{WBC}^{p_2} = 6.69$*, and model the uncertainty about their observations through the 1-dimensional Gaussian distributions* $N_{p_1}(12.33, 1.67)$ *and* $N_{p_2}(6.69, 0.90)$*.*

Given an observation for a new patient p_3*, with* $WBC^{p_3} = 10.82$*, we can model the IV uncertainty for* p_3 *by the Gaussian distribution* $N_{p_3}(10.82, 1.46)$*.*

Due to the complexity of design studies to obtain reliable IV estimates, a few compiled sources of IV data, for healthy patients, are available: the largest existing repositories in this sense, are the data originating from the European Biological Variation Study (EuBIVAS) and the Biological Variation Database (BVD) [44, 45], both encompassing data about commonly used laboratory biomarkers. In the following sections, we will rely on data available from these sources in the definition of our experiments.

3. Methods

In this section, we describe the main proposed methodology. We first introduce a theoretical contribution to frame the modeling of instantial variation within the framework of statistical learning theory. Then, we will describe two different experiments: in the first experiment, we evaluate how commonly used ML models fare when dealing with data affected by IV; then, in the second experiment, we evaluate the application of more advanced ML approaches to improve robustness to IV.

3.1. Instantial variation and Statistical Learning

One of the most simple yet remarkable results in Statistical Learning Theory (SLT) is the *error decomposition theorem* [29] (also called bias-variance tradeoff, or bias-complexity tradeoff), which states that the true risk $L_D(h)$ of a function h from a family H w.r.t. to a distribution D on the instance space $Z = X \quad Y$ can be decomposed as:

$$L_D(h) = \epsilon^{Bayes} + \epsilon^{Bias} + \epsilon^{Est} \tag{1}$$

where $\epsilon^{Bayes} = min_{f2F} L_D(f)$ is the *Bayes error*, i.e. the minimum error achievable by any measurable function; $\epsilon^{Bias} = min_{h^0 2H} L_D(h^0) \quad min_{f2F} L_D(f)$ is the *bias*, i.e. the gap between the Bayes error and the minimum error achievable in class H; $\epsilon^{Est} = L_D(h) \quad min_{h^0 2H} L_D(h^0)$ is the *estimation error*, i.e. the gap between the error achieved by h and the minimum error achievable in H. This latter term can be further characterized by noting that, with probability $1 \quad \delta$ over the selection of a training set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, the estimation error can be bounded by $\epsilon^{Est} \quad L_S(h) + O(Complexity_{\delta,m}(H)) \quad min_{h^0 2H} L_D(h^0)$, where $L_S(h)$ where is the error achieved by h on the training set S and $Complexity_{\delta,m}(H)$ is a measure of the capacity of a class of functions, such as its Rademacher complexity or VC dimension [29].

A striking consequence of IV for ML tasks regards a generalization of the error decomposition theorem due to the impossibility of accessing the true distributional-valued representation of instances but only a sample drawn from the respective distributions. To formalize this notion, as in the previous section, denote with $f_p = N(x^p, \ ^p)$ the distributional representation due to IV for instance p. Then, the learning task can be formalized through the definition of a *random measure* [46] η defined over the Borel σ-algebra (Z, B) on the instance space $Z = X \quad Y$, which associates to each instance (x, y) a probability measure $N(x, \ ) \quad \delta_y$, where δ_y is the Dirac measure at $y \ 2 \ Y$. A training set $S = f(x^1, y^1), \ldots, (x^n, y^n)g$ is then obtained by first sampling n random measures $f_1, \ldots, f_n$ from η^n, and then, for each p, by sampling a random element $(x^p, y^p) \quad f_p$. Then, the IV-induced generalization of the error decomposition theorem can be formulated as:

$$L_\eta(h) = \epsilon_\eta^{Bayes} + \epsilon_\eta^{Bias} + \epsilon_\eta^{Est} + \epsilon_\eta^{IV} \tag{2}$$

Indeed, the true error of h w.r.t. η can be expressed as

$$L_\eta(h) = E_{F \ \ \eta^m} \left[\frac{1}{m} \sum_{f_p 2 F} E_{(x^p, y^p) \ \ f_p} l(h, (x^p, y^p)) \right]. \tag{3}$$

Letting D be the probability measure over $X \quad Y$ obtained as the *intensity measure* [47] of η, and $L_D(h) = E_{S \ \ D^m} L_S(h)$ be the expected error of h w.r.t. to the sampling of a training set S from the product measure D^m, then the above expression can be derived by setting $\epsilon_\eta^{Bayes} = min_{f 2 F} L_\eta(f)$, $\epsilon_\eta^{Bias} = min_{h' 2 H} L_\eta(h') \quad min_{f 2 F} L_\eta(f)$, $\epsilon_\eta^{Est} = L_D(h) \quad min_{h' 2 H} L_\eta(h')$ and ϵ_η^{IV} is defined as $E_{F \ \ \eta^m, S \ \ D} \left[\frac{1}{m} \sum_p E_{(x^p, y^p), (x^{p'}, y^{p'}) \ \ f_p} l(h, (x^p, y^p)) \quad l(h, (x^{p'}, y^{p'})) \right]$.

Thus, compared with Eq (1), Eq (2) includes an additional error term ϵ^{IV} which measures the gap in performance due to the inability to use the IV-induced distributional representation of the instances, bur rather only a single instantiation of such distributions. This aspect is also reflected in the estimation error component in which the reference $min_{h' 2 H} L_\eta(h')$ is compared not with the true error $L_\eta(h)$ but rather with the expected error over all possible instantiations $L_D(h)$. In the following sections, we will show, through an experimental study, that the impact of IV can be significant and lead to an overestimation of any ML algorithm's performance and robustness.

3.2. Measuring the Impact of Instantial Variation on Machine Learning Models

In order to study whether and how the performance of a ML model could be impacted by IV, we designed an experiment through which we evaluated several commonly adopted ML models in the task of COVID-19 diagnosis from routine laboratory blood exams, using a public benchmark dataset. Aside from its practical relevance [35], we selected this task for three additional reasons. First, blood exams are considered one of the most stable panels of exams [48]: this allows us to evaluate the impact of IV in a conservative scenario where the

features of interest are affected by relatively low levels of variability. Second, validated data about IV for healthy patients who underwent blood exams are available in the specialized literature [49, 50, 51] and these exams have high predictive power for the task of COVID-19 diagnosis [35]. Third, the selected dataset was associated with a companion longitudinal study [36] that has been used to estimate IV data for the COVID-19 positive patients: we believe this to be particularly relevant since no information of this kind is available for non-healthy patients, due to the complexity of designing studies for the collection of IV data, which could nonetheless exhibit disease-specific patterns. Although the estimation of IV is of paramount importance, both in medicine and other safety-critical domains, the striking lack of datasets presenting information to assess IV makes it a priority to devote further efforts and initiatives to make such resources available to the ML research community to make their models more robust and reliable. For this reason, the considered dataset was specifically commissioned to clinical laboratory experts for the purpose of studying the impact of IV in ML, as well as with the aim of developing a first public benchmark dataset for further studies in this setting. Furthermore, we remark that, while other longitudinal datasets have been made available for the specific setting of COVID-19 diagnosis (see e.g. [52]), when taking into account data coming from different populations and settings (such as the Italian and Chinese ones), one must also consider the potential impact of covariate shifts and out-of-distribution-related variability [53], as well as potential issues of harmonization and pre-analytical variability (whereas both can instead be considered negligible when considering data collected in the same setting and with the same instrumentation): since in this article we were interested in assessing the specific impact of IV, we focused only on the above-mentioned dataset.

More in particular, we used a dataset of patients who were admitted at the emergency departments of the IRCCS Ospedale San Raffaele and IRCCS Istituto Ortopedico Galeazzi, two of the major research hospitals in Italy, and underwent a COVID-19 test [35, 53]. The dataset was collected between February and May 2020 and encompasses 18 continuous features and 3 binary features (including the target). Since the dataset was affected by missing data, in order to limit the bias due to data imputation, we discarded all instances having more than 25% missing values: the resulting dataset encompasses 1422 instances, pertaining to an equal number of different patients, and is described in Table 2.

Complete Blood Count data (i.e. features WBC, RBC, HCT, NE, LY, MO, EO, BA) was obtained by analysis of whole blood samples by means of a Sysmex XE 2100 haematology automated analyser. Biochemical data (ALT, AST, ALP, GGT, LDH, CK, CA, GLU, UREA, CREA) was obtained by analysis of serum samples by means of a Cobas 6000 Roche automated analyser. For each of the considered patients, COVID-19 positivity was determined based on the result of the molecular test for SARS-CoV-2 performed by RT-PCR on nasopharyngeal swabs: on a set of 165 cases for which the RT-PCR reported uncertain results, chest radiography and X-rays were also used to improve over the sensitivity of the RT-PCR test by combination testing.

To evaluate the impact of IV, we used a biologically-informed generative

Table 2: The list of features, along with the target. Mean and standard deviation are reported for continuous features, distribution of values is reported for discrete feature. For the discrete features we report the distribution of values. For the laboratory blood data, we also report the analytical (CVA) and biological (CVI) variation, differentiated by healthy vs non-healthy patients, and missing rate.

Features	Acronym	Units	Mean	Std	Missing	CVA	$CVI_{y=0}$	$CVI_{y=1}$
Alanine Transaminase	ALT	U/L	39.87	42.26	0.07	0.04	0.093	0.051
Aspartate Transaminase	AST	U/L	46.90	51.90	0.14	0.04	0.095	0.52
Alkaline Phosphatase	ALP	U/L	88.61	72.09	16.24	0.05	0.054	0.045
Gamma Glutamyl Transferase	GGT	U/L	67.48	140.52	17.09	0.035	0.089	0.036
Lactate Dehydrogenase	LDH	U/L	332.52	218.43	8.02	0.03	0.052	0.024
Creatine Kinase	CK	U/L	184.47	382.02	56.19	0.05	0.145	0.062
Calcium	CA	mg/dL	2.20	0.17	0.84	0.03	0.018	0.018
Glucosium	GLU	mg/dL	119.12	55.80	0.42	0.028	0.047	0.026
Urea	UREA	mg/dL	48.64	42.69	31.01	0.03	0.141	0.035
Creatinine	CREA	mg/dL	1.19	1.01	0.07	0.025	0.044	0.022
Leukocytes	WBC	10^9/L	8.65	4.77	0.00	0.019	0.111	0.033
Erythrocytes	RBC	10^{12}/L	4.55	0.72	0.00	0.009	0.018	0.010
Hematocrit	HCT	%	39.47	5.57	0.00	0.018	0.024	0.019
Neutrophils	NE	%	72.48	13.35	8.51	0.03	0.146	0.014
Lymphocytes	LY	%	18.58	11.11	8.51	0.036	0.11	0.043
Monocytes	MO	%	7.76	3.86	8.51	0.063	0.134	0.033
Eosinophils	EO	%	0.82	1.59	8.51	0.079	0.156	0.098
Basophils	BA	%	0.34	0.27	8.51	0.031	0.128	0.056
Sex	-	Female Male	42% 58%	-	-	-	-	-
Age	-	Years	61.19	18.89	-	-	-	-
Target	-	Positive Negative	53% 47%	-	-	-	-	-

model whose aim was to simulate the effect of biological and analytical variation on the measured features of the patients in the dataset. More in detail, based on the definition and computation of IV described previously, the generative model is defined by a case-dependent, class-conditional, multi-variate Gaussian distribution $N(x, \ ^{x,y})$, where we recall $^{x,y} = diag(x \ \sqrt{CVA^2 + CVI_y^2})$. We note that the adopted generative model grounds on two weak assumptions: first, that the percent BV coefficients of healthy and non-healthy patients are

different; second, that the distributions of individual features are conditionally independent having observed the patients' feature values and class label: these assumptions are widely adopted in the specialized IV literature as well as implicitly in the release format of the available IV data sources. More in particular, for CVA and $CVI_{y=0}$ we considered values previously reported in the literature [49, 51, 50], while the values of $CVI_{y=1}$ were estimated from the longitudinal observation of the COVID-19 positive patients considered in this study [36]: the considered dataset (see [36] and [54] for further details) encompassed multiple observations for staying day (for up to 31 staying days) for 1104 COVID-19 positive patients, which encompassed the popularion considered in this article. Based on this data, estimates for CVT, CVA and CVI were compute using the same methodology as described in the previous section and illustrate in Example 1. All data, including the CVA and CVI estimates, is publicly available at `anonymizedurl`.

We considered 7 different ML models, commonly used in medical settings on tabular data, namely: Support Vector Machine (with RBF kernel) (SVM), Logistic Regression (LR), k-Nearest Neighbors (KNN), Naive Bayes (NB), Random Forest (RF), Gradient Boosting (GB), ExtraTrees (ET). We evaluated, in particular, the scikit-learn implementations of the previous models, with default hyper-parameters. We did not evaluate deep learning models, due to several limitations of these latter for tabular data tasks, such as the one we consider in this paper. Indeed, as a first limitation, deep learning models have been shown to require much extensive hyper-parameter optimization compared to simpler models in order to achieve acceptable performance [55, 56]. Most importantly, several recent studies [55, 57, 56] and surveys [58], have shown deep learning models to be out-performed by other models on tabular data. Furthermore, in previous studies [35, 53, 36, 54] we showed that standard ML models, such as those mentioned above, are able to achieve state-of-the-art performance for the task of COVID-19 diagnosis from routine laboratory blood exams.

All code was implemented in Python v. 3.10.4, using numpy v. 1.23.0, scikit-learn v. 1.1.1 and scikit-weak v. 0.2.0. For the above-mentioned ML models, we considered the default hyper-parameter values as defined in scikit-learn v. 1.1.1, with the exception of the random_state seed, which was set to 99 for all evaluated models to ensure reproducibility, and the max_depth hyperparameters for Random Forest and Gradient Boosting, which were set to 10 to avoid over-fitting and reduce the running time.

The impact of IV on the performance of the above-mentioned ML models was evaluated by means of a repeated cross-validation evaluation procedure: for a total of 100 iterations, a 3-fold cross-validation procedure was applied. More in detail, in each 3-fold cross-validation the two training folds were used to train the ML models, while the test fold Te was used to obtain a perturbed fold Te_p as follows: for each instance (x, y) *2* Te, a perturbed instance (x^0, y) was obtained to simulate the effect of instantial variation, by sampling x^0 from $N(x, \ ^{x,y})$. The trained ML model was then evaluated on both Te and Te_p to measure the impact of instantial variation, if any, by comparing the distribution of average performance on the original test folds with that of the perturbed test folds. In

terms of performance metrics, we considered the accuracy, the AUC and the F1 score. The robustness of the ML models to IV was evaluated by comparing the average performance on the non-perturbed and IV perturbed data: in particular, we considered a model to be robust to IV if the 95% confidence intervals for the above-mentioned quantities overlapped (equivalently, the C.I. of the difference included the value 0).

3.3. Data Augmentation and Imprecisiation Methods to Manage Instantial Variation

In light of the results for the traditional ML models, which show the lack of robustness of standard ML models w.r.t. IV (see Section 4.1), we investigated the application of more advanced methods that attempt to directly address the representation of IV in data and hence tackle the error decomposition show in Eq (2). In particular, we consider approaches based either on *data augmentation*, a popular solution in modern ML to deal with data issues, or *data imprecisiation*, a soft computing-inspired approach to deal with uncertainty in data. In both cases, we adopted the experimental protocol described in the previous section.

Data augmentation [59, 33] refers to regularization techniques whose aim is to increase the stability and robustness of a ML model by enriching the training set with new instances. In our setting, the idea is to inject further information related to the IV distribution within the model to improve generalization. Since in the considered setting a generative model of IV was available, this latter was used to simulate synthetic data to augment the original training set. For each instance (x, y) in the training folds, we generated $n = 100$ new samples from the distribution $N(x, \ ^{x,y})$, so as to simulate the effect of having multiple observations, perturbed by IV, for each patient. We considered, in particular, the application of the above-mentioned data augmentation strategy to the SVM (denoted as ACS) and Gradient Boosting (denoted as ACG) ML models, since these latter two were shown to be more robust to IV (see previous section). The pseudo-code for evaluating the data augmentation models is reported in Algorithm 1.

By contrast, data imprecisiation [24, 34] refers to soft computing approaches by which data affected by some form of uncertainty are transformed into imprecise (soft) observations, that is distributions over possible instances, which are then used to train specialized ML algorithms. Formally speaking, an *imprecisiation scheme* is a function $is : X \quad Y \ ? \ [0, 1]^{X \ \ Y}$, where X is the feature space. In the experiments, we considered two commonly adopted imprecisiation schemes grounding on, respectively, probability theory and fuzzy set theory [60], namely:

$$is_{prob} : (x, y) \ ? \ (N(x, \ ^{x,y}), y) \tag{4}$$

$$is_{poss} : (x, y) \ ? \ (Gauss(x, \ ^{x,y}), y) \tag{5}$$

where $Gauss(a, b)$ denotes the Gaussian fuzzy vector, whose q-component is defined as $Gauss(a, b)_q(x) = e^{\frac{(x_q \ a)^2}{b^2}}$. Intuitively, is_{prob} represents each instance

Algorithm 1 The procedure to evaluate the impact of IV on the data augmentation-based ML models.

procedure data_augmentation_eval(h: ML model, S: dataset, M: metric, a : number of augmented instances)
  for all iterations it = 1 to 100 **do**
    Split S in 3 class-stratified folds
    for all Tr: training fold, Te : test fold **do**
      Tr_a = ;
      for all (x, y) 2 Tr **do**
        for all iteration j = 1 to a **do**
          Add to Tr_a (x^0, y), x^0 $N(x,$ $^{x,y})$
        end for
      end for
      Te^0 = ;
      for all (x, y) 2 Te **do**
        Add to Te^0 (x^0, y), x^0 $N(x,$ $^{x,y})$
      end for
      Train h on Tr_a
      Eval h on Te $(M(h, Te))$, Te^0 $(M(h, Te^0))$
    end for
  end for
  return The distributions of $M(h, Te)$ and $M(h, Te^0)$
end procedure

affected by IV as a Gaussian probability distribution over possible instances, while is_{poss} represents each instance affected by IV as a Gaussian possibility distribution (i.e., a Gaussian fuzzy set) over possible instances. Thus, the general idea of applying data imprecisiation in our setting is to model the uncertainty due to IV by representing each instance as a soft cloud of points in the feature space, whose distribution is determined by the IV parameters, as a form of soft computing-inspired regularization.

We considered three ML algorithms proposed in the learning from imprecise data literature, namely: k-Nearest Distributions (KND) [61], Support Measure Machine (SMM) [62], Weighted re-Sampling Forest (WSF) [63].

KND denotes the generalization of kNN to distribution-valued instances, namely we used the is_{prob} scheme[6] and Mahalanobis distance:

$$(x^p \quad x^q)^T \frac{^{x^p,y^p} \ ^1 + \ ^{x^q,y^q} \ ^1}{2} (x^p \quad x^q) \tag{6}$$

SMM, by contrast, refers to the generalization of SVM to instances represented as probability distributions (thus, only the is_{prob} imprecisiation scheme

[6]Since Mahalanobis' distance takes into account only the mean and scale, using is_{poss} scheme would result in the same algorithm.

was considered). The SMM model grounds on the notion of a *kernel mean embedding* [62], that is a generalization of the notion of kernel in ML to the the space of probability distributions, which could thus be seen as a measure of similarity between two imprecise instances. For computational complexity reasons, we considered the RBF kernel, which for normally distributed imprecise instances can be expressed in closed form as [62]:

$$RBF_{\gamma,is_{prob}} = \frac{e^{\frac{(x^p \quad x^q)^T(\quad x^p,y^p + \quad x^q,y^q + \frac{1}{\gamma} I) \quad ^1(x^p \quad x^q)}{2}}}{\sqrt{det(\gamma \quad x^p,y^p + \gamma \quad x^q,y^q + I)}} \tag{7}$$

Finally, the WSF model is an approximation algorithm to solve the generalized risk minimization problem [24], a commonly adopted approach to deal with imprecise. WSF is based on a generalization of bootstrapped tree ensembles to instances represented as possibility distributions (thus, only the is_{poss} imprecisiation scheme was considered): in addition to the randomization w.r.t. the split point selection and the bootstrap re-sampling of the instances, an additional randomization on the feature values is considered. Specifically, for each tree in the ensemble, each imprecise instance $is_{poss}(x, y)$ in the corresponding bootstrap set is used to sample an instance (x^0, y^0), by means of a two-step procedure [64]: first, a number $\alpha \; 2 \; [0, 1]$ is selected uniformly at random, then a random value is drawn from the α-cut $is_{poss}(x, y)^\alpha = f(x^0, y^0) \; 2 \; X \quad Y : is_{poss}(x, y)(x^0, y^0) \quad \alpha g$. A pseudo-code description of WSF is reported in Algorithm 2. Further details on the computational and statistical properties of WSF for the IV setting are reported in Appendix A.

Algorithm 2 The WSF algorithm.

procedure WSF(S: dataset, ens: ensemble size, H model class)
  $Ensemble$;
  for all iterations $it = 1$ to ens **do**
    Draw a boostrap sample S^0 from S
    Tr_{it} ;
    for all $(x, y) \; 2 \; S^0$ **do**
      Sample α $U[0, 1]$
      Add (x^0, y^0) $is_{poss}(x, y)^\alpha$ to Tr_{it}
    end for
    Add model $h_{it} \; 2 \; H$ trained on Tr_{it} to $Ensemble$
  end for
  return $Ensemble$
end procedure

The imprecisiation-based models were evaluated in a setup similar to the one adopted for the data augmentation-based ML models, as shown in Algorithm 3.

All data augmentation-based and imprecisiation-based models were implemented in Python v. 3.10.4, using numpy v. 1.23.0, scikit-learn v. 1.1.1 and scikit-weak v. 0.2.0. The full code for the algorithms and evaluation procedures

Algorithm 3 The procedure to evaluate the impact of IV on the data imprecisiation-based ML models.

procedure `data_imprecisiation_eval`(h: ML model, S: dataset, M: metric, is: imprecisiation scheme)
  for all iterations it = 1 to 100 **do**
    Split S in 3 class-stratified folds
    for all Tr: training fold, Te : test fold **do**
      Tr_a = ;; Te_b = ;; Te^0 = ;
      for all (x, y) 2 Tr **do**
        $Tr_a.append(is((x, y)))$
      end for
      for all (x, y) 2 Te **do**
        $Te_b.append(is((x, y)))$
        Sample (x_p, y) $N(x,$ $^{x,y})$
        $Te^0.append(is((x^0, y)))$
      end for
      Train h on Tr_a
      Eval h on Te_b ($M(h, Te_b)$), Te^0 ($M(h, Te^0)$)
    end for
  end for
  return The distributions of $M(h, Te_b)$ and $M(h, Te^0)$
end procedure

is available on GitHub at `anonymizedurl`. In regard to the hyper-parameter settings, for the data augmentation models we set the number of augmentation rounds to 100: for ACS we used as base model a SVC with rbf kernel and default hyper-parameters, while for ACG we used a GradientBoostingClassifier with max_depth set to 0 and random_state set to 99 for consistency with the classical case. For SMM we used as kernel the RBF kernel defined in (7) with $\gamma = \frac{1}{\text{num. features}}$, while for WSF we used ExtraTreeClassifier as base classifier, we set the number of ensembled models to 100 and the random_state seed to 99. Finally, for KND we set the number of neighbors k to 5.

4. Results and Discussion

In the next sections we report on the results of the experiments described in Sections 3.2 and 3.3.

4.1. Measuring the Impact of Instantial Variation on Machine Learning Models

First of all, we assessed whether the perturbed data obtained by means of the considered generative model was significantly different from the original data. Ideally, to be realistic, IV-based perturbations should not influence too much the overall data distribution.To this purpose, we considered a subset of 4 predictive features (namely LY, WBC, NE and AST), which were previously

shown to be among the most predictive features for the considered task [35]. We compared the distributions of the above-mentioned features before and after the IV perturbations, by means of the Kolmogorov-Smirnov test with $\alpha = 0.01$. The obtained p-values were, respectively, 1 (for LY, WBC and NE) and 0.104 (for AST): thus the null hypothesis of equal distributions for the IV perturbed and non-perturbed data could not be rejected.

The impact of IV on the ML models is reported in Figure 2. The difference in performance (baseline vs perturbed) was significant for all algorithms: indeed, for all algorithms, the confidence intervals on the baseline and IV perturbed data did not overlap. The best algorithms on the non-perturbed data were RF and ET, w.r.t. all considered metrics (AUC: 0.87, Accuracy: 0.8, F1: 0.8); while the best algorithms on the the IV perturbed data were SVM (w.r.t. AUC: 0.69, and Accuracy: 0.5) and GB (w.r.t. F1: 0.5).

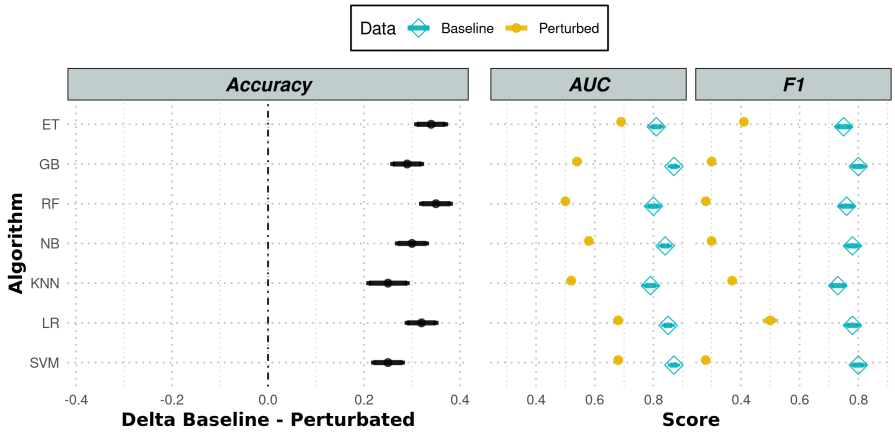

Figure 2: Results of the experiments for measuring the impact of IV on the performance of standard ML models. For each algorithm and metrics, we report the average and 95% confidence interval for both baseline (that is, non-perturbed) and IV perturbed data.

These results highlight how, even though the distributions of highly predictive feature were not significantly affected by IV, IV nonetheless had a significant impact on the performance of all the considered ML algorithms, that were therefore not robust to IV-related uncertainty. Algorithms, however, were not equal in their robustness (or lack thereof). In particular, the more robust models were SVM (w.r.t. Accuracy, with average performance decrease 0.25, and AUC, with average decrease 0.12) and GB (w.r.t. F1 score, with average performance decrease 0.28). While this latter observation can be given a learning theoretical justification[7], we note that even SVM and GB had a significant decrease

[7]Both SVM and GB are margin-based classifiers [65, 66]. It is not hard to see that the existence of a large margin on the non-perturbed data is a necessary (but not sufficient) condition for robustness to IV: indeed, if the margin of the optimal

in performance on the IV-perturbed data. Thus, even models that are usually considered to be robust can nevertheless be strongly impacted by IV.

4.2. Data Augmentation and Imprecisiation Methods to Manage Instantial Variation

The results for data augmentation and imprecisiation-based ML models are reported in Figure 3. For all models except SMM, the difference in performance on baseline and IV perturbed data was not significant. The best models on the non-perturbed data were SMM, WSF (w.r.t. AUC: 0.87) and WSF, ACG (w.r.t. Accuracy: 0.8, F1: 0.81), while the best models on the IV perturbed data were ACG and WSF (AUC: 0.86, Accuracy: 0.79, F1: 0.8). Comparing these results with those shown in the previous section, it is easy to observe that both data augmentation and data imprecisiation-based ML models were much more robust to IV perturbations than the standard ML models. Indeed, the most robust models (w.r.t. AUC: WSF and ACS, with average difference 0.003; w.r.t. Accuracy and F1: WSF and ACG, with average difference 0.006) were hardly impacted by IV. Even the least robust model (SMM) was much more robust than the standard ML models (average differences w.r.t. AUC: 0.08; w.r.t. Accuracy: 0.09, w.r.t. F1: 0.09).

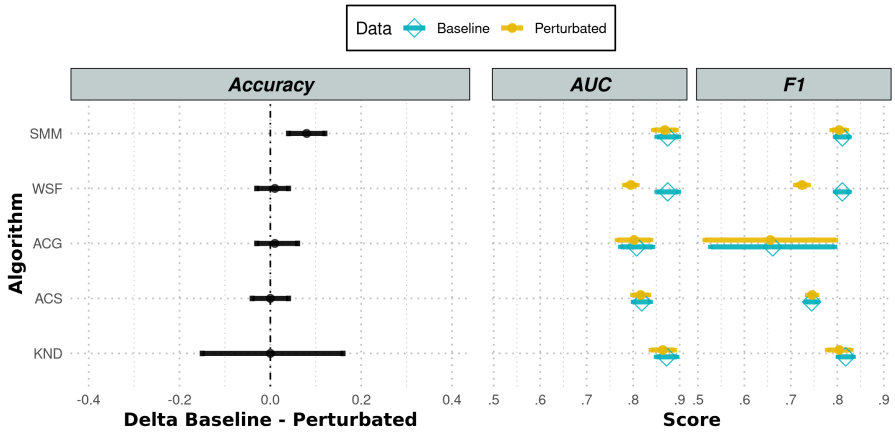

Figure 3: Results of the experiments for measuring the impact of IV on the performance of data augmentation-based and data imprecisiation-based ML models. For each algorithm and metrics, we report the average and 95% confidence interval for both baseline (that is, non-perturbed) and IV perturbed data.

In light of these results, we claim that data augmentation and imprecisiation can be helpful to improve robustness under IV perturbations. We conjecture

classifier on the training set is smaller than Δ, then the existence of any two patients $(x_1, y_1), (x_2, y_2)$ with different diagnosis and whose IV-induced distributions are s.t. $Pr\left[d(x'_1, x'_2) \leq \Delta | x'_1 \sim N(x_1, \Sigma^{x_1,y_1}), x'_2 \sim N(x_2, \Sigma^{x_2,y_2})\right] \geq \epsilon$, implies that the model is not robust to IV.

this to be due to directly taking into account information about IV in data representation and model training, which allows to strike a trade-off among the various components of the generalized error decomposition shown in Eq. (2) (see also the Appendix for a more detailed theoretical justification for the WSF algorithm). We note that these two approaches, while performing similarly in terms of accuracy and robustness, have different characteristics that may influence their suitability in practical scenarios. Data augmentation methods allow to use out-of-the-box ML models, since IV management is implemented as a preprocessing step: this is not the case for data imprecisiation-based approaches, which require specialized ML algorithms. By contrast, imprecisiation-based approaches have lower computational cost and may thus scale better on larger datasets: e.g., if m is the training set size, d the number of features, and r the augmentation rounds then, the time costs of SMM and WSF are $O(m^2 d^3)$ and $O(d\, m \log(m))$; while those of ACS and ACG are, respectively, $O(m^2 r^2)$ and $O(d\, m\, r \log(mr))$.

5. Conclusion

In this article we studied the impact of IV, an often neglected type of uncertainty affecting data (as representations and proxies [67] of entities' features and behaviors), on the performance and robustness of ML models. Through a realistic experiment on COVID-19 diagnosis, a problem of significant practical interest which we take as paradigmatic of the class of applications with high risk and impact on human subjects, we showed how standard ML algorithms can be strongly impacted by the presence of IV, failing to generalize properly. Such an issue can severely limit the applicability and safety of ML methods in tasks where data are expected to be affected by IV, that is most applications in real-world domains where the manifestations of the phenomena of interest could exhibit varying patterns. Our results then imply that out-of-the-box methods cannot be naively applied in such domains. Crucially, even though our example is medical, we note that our analyses and methods are domain-independent and apply to each and any setting in which our theoretical framework adequately describes the structure and nature of data. Nonetheless, every cloud has a silver lining, and we showed that more advanced learning methods, grounding on data augmentation and soft computing-inspired data imprecisiation, can achieve better robustness with respect to IV: this highlights the need to employ models that take into account the *generative history* underlying the data acquisition process, including the uncertainty due to IV, in their learning algorithms. Furthermore, we believe that our results highlight the importance of adopting proper algorithmic and experimental designs for ML studies in medicine: due to the potential impact of IV on the performance of ML models, data collection studies should be designed so as to enable the estimation of IV values which could then be used in the ML development phase. Thus, increasing emphasis should be placed on longitudinal studies, or otherwise studies in which multiple samples are collected for each involved patients under controlled conditions: as we described in Section 2.1 and illustrated in Example 1 such longitudinal studies can be used to

obtain precise and reliable estimation of IV, which can then be incorporated in ML algorithm such as those we described in Section 3.3, that can harness such information about the data generating mechanism to obtain more robust and uncertainty-aware ML models (see Example 2).

Example 2. *Assume that the data illustrated in Example 1 was collected and used to obtain estimates of IV, as well as to train a data augmentation-based or imprecisiation-based ML model h. Assume we observe new observations for three patients p_3, p_4, p_5, with $WBC^{p_3} = 10.82, WBC^{p_4} = 6.45$ and $WBC^{p_5} = 8.67$. The previously computed IV can then be used to apply the probability-based and possibility-based imprecisiation schemes: $is_{prob}(p_3) = \mathcal{N}(10.82, 1.46), is_{poss}(p_3) = Gauss(10.82, 1.46)$, $is_{prob}(p_4) = \mathcal{N}(6.45, 0.87), is_{poss}(p_4) = Gauss(6.45, 0.87)$ and $is_{prob}(p_5) = \mathcal{N}(8.67, 1.17), is_{poss}(p_5) = Gauss(8.67, 1.17)$. These imprecisified instances can then be given be as input to an already trained data augmentation-based or imprecisiation-based ML model h, such as those described in Section 3.3, to obtain a prediction that takes into account the uncertainty due to IV.*

We believe that these results could pave the way for the investigation of IV and its effects on the safety and robustness of ML models deployed in real-world clinical settings, also to meet high-level requirements expressed in regulatory principles in laws and regulations (e.g. the EU AI Act). To this purpose, our study has been based on a large dataset specifically collected for the purpose of studying the impact of IV on ML development, which is publicly available and could thus be used as benchmark for future studies dealing with this problem. Concluding, we summarize in what follows the open problems that might be of interest to the ML community:

- In the introduction, we have defined IV as a combination of two different terms: measurand and measuring variation. In the medical setting, we have seen that these two forms of variation can be associated with the notions of, respectively, biological and analytical (or, analytical and pre-analytical) variation, and describe ways to compute such a decomposition of IV in its two components. We believe it would be interesting to extend such a decomposition in general settings, and to further develop the theoretical analysis introduced in Section 3.1 to account for the decomposition of the ϵ_η^{IV} term (see Eq. (2)) into two terms that correspond to measurand and measuring variation;

- In our experiments, we assumed the IV distributions to be Gaussian with diagonal covariance. While this model is commonly adopted in the literature, we believe that further research should explore the relaxation of this assumption, by considering more general models of IV accounting for non-linear or causal relationships among features: to this aim, the use of deep generative models [68] or causal models [69] could enable the construction of more informative and expressive IV models;

- While we proposed and discussed a framework to model IV in SLT, the theoretical side of this issue merits further study. In particular, even

though the problem of learning from distributional data has recently been investigated [70, 71, 62], this area is still in its infancy;

- Last, but not least, in this work we showed the impact of IV on COVID-19 diagnosis from blood tests. Future work should extend our work to a broader spectrum of applications. We believe this to be of primary importance to advance the development of robust and sound ML systems: in this sense, we hope and believe that our results would foster the collection and sharing of datasets that allow to account for this important characteristic of data in future research studies.

Acknowledgments

This research has been supported by the Italian Ministry of Health through project "Ricerca Corrente".

References

[1] R. Aggarwal, V. Sounderajah, G. Martin, D. S. Ting, A. Karthikesalingam, D. King, H. Ashrafian, A. Darzi, Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis, NPJ digital medicine 4 (2021) 1–23.

[2] F. Fahimi, S. Dosen, K. K. Ang, N. Mrachacz-Kersting, C. Guan, Generative adversarial networks-based data augmentation for brain–computer interface, IEEE transactions on neural networks and learning systems 32 (2020) 4039–4051.

[3] L. Jiao, R. Zhang, F. Liu, S. Yang, B. Hou, L. Li, X. Tang, New generation deep learning for video object detection: A survey, IEEE Transactions on Neural Networks and Learning Systems (2021).

[4] D. W. Otter, J. R. Medina, J. K. Kalita, A survey of the usages of deep learning for natural language processing, IEEE transactions on neural networks and learning systems 32 (2020) 604–624.

[5] J. Wilkinson, K. F. Arnold, E. J. Murray, M. van Smeden, K. Carr, R. Sippy, M. de Kamps, A. Beam, S. Konigorski, C. Lippert, Time to reality check the promises of machine learning-powered precision medicine, The Lancet Digital Health 2 (2020) e677–e680.

[6] C. L. Andaur Navarro, J. A. Damen, T. Takada, S. W. Nijman, P. Dhiman, J. Ma, G. S. Collins, R. Bajpai, R. D. Riley, K. G. Moons, Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review., bmj 375 (2021) n2281.

[7] J. Futoma, M. Simons, T. Panch, F. Doshi-Velez, L. A. Celi, The myth of generalisability in clinical research and machine learning in health care, The Lancet Digital Health 2 (2020) e489–e492.

[8] E. Coiera, The last mile: where artificial intelligence meets reality, Journal of Medical Internet Research 21 (2019) e16323.

[9] A. L. Beam, A. K. Manrai, M. Ghassemi, Challenges to the reproducibility of machine learning models in health care, Jama 323 (2020) 305–306.

[10] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, B. Van Calster, A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models, Journal of clinical epidemiology 110 (2019) 12–22.

[11] X. Li, S. Zhang, Q. Zhang, X. Wei, Y. Pan, J. Zhao, X. Xin, C. Qin, X. Wang, J. Li, Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study, The Lancet Oncology 20 (2019) 193–201.

[12] E. Coiera, E. Ammenwerth, A. Georgiou, F. Magrabi, Does health informatics have a replication crisis?, Journal of the American Medical Informatics Association 25 (2018) 963–968.

[13] L. Akoglu, Anomaly mining: Past, present and future, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 1–2.

[14] D. Adila, D. Kang, Understanding out-of-distribution: A perspective of data dynamics, in: I (Still) Can't Believe It's Not Better! Workshop at NeurIPS 2021, PMLR, 2022, pp. 1–8.

[15] P. Morteza, Y. Li, Provable guarantees for understanding out-of-distribution detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 8, 2022.

[16] B.-J. Hou, L. Zhang, Z.-H. Zhou, Learning with feature evolvable streams, Advances in Neural Information Processing Systems 30 (2017).

[17] J. Liu, Z. Shen, P. Cui, L. Zhou, K. Kuang, B. Li, Y. Lin, Stable adversarial learning under distributional shifts, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 8662–8670.

[18] S. Rabanser, S. Günnemann, Z. Lipton, Failing loudly: An empirical study of methods for detecting dataset shift, Advances in Neural Information Processing Systems 32 (2019).

[19] C. G. Fraser, Biological variation: from principles to practice, American Association for Clinical Chemistry, 2001.

[20] L. Naranjo, C. J. Pérez, Y. Campos-Roca, M. Madruga, Replication-based regularization approaches to diagnose reinke's edema by using voice recordings, Artificial Intelligence In Medicine 120 (2021) 102162.

[21] M. Plebani, A. Padoan, G. Lippi, Biological variation: back to basics, Clinical Chemistry and Laboratory Medicine (CCLM) 53 (2015) 155–156.

[22] F. Cabitza, A. Campagner, V. Basile, Toward a perspectivist turn in ground truthing for predictive computing, Proceedings of the AAAI Conference on Artificial Intelligence (2023).

[23] F. Cabitza, A. Campagner, M. Mattioli, The unbearable (technical) unreliability of automated facial emotion recognition, Big Data & Society 9 (2022) 20539517221129549.

[24] E. Hüllermeier, Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization, International Journal of Approximate Reasoning 55 (2014) 1519–1534.

[25] H. Song, M. Kim, D. Park, Y. Shin, J.-G. Lee, Learning from noisy labels with deep neural networks: A survey, IEEE Transactions on Neural Networks and Learning Systems (2022).

[26] F. Cabitza, A. Campagner, D. Albano, A. Aliprandi, A. Bruno, V. Chianca, A. Corazza, F. Di Pietto, A. Gambino, S. Gitto, et al., The elephant in the machine: Proposing a new metric of data reliability and its application to a medical case to assess classification reliability, Applied Sciences 10 (2020) 4014.

[27] T. Badrick, Biological variation: Understanding why it is so important?, Practical Laboratory Medicine 23 (2021) e00199.

[28] D. H. Spodick, R. L. Bishop, Computer treason: intraobserver variability of an electrocardiographic computer system, The American journal of cardiology 80 (1997) 102–103.

[29] S. Shalev-Shwartz, S. Ben-David, Understanding machine learning: From theory to algorithms, Cambridge university press, 2014.

[30] A. K. Aarsand, T. Røraas, P. Fernandez-Calle, C. Ricos, J. Díaz-Garzón, N. Jonker, C. Perich, E. González-Lao, A. Carobene, J. Minchinela, The biological variation data critical appraisal checklist: a standard for evaluating studies on biological variation, Clinical chemistry 64 (2018) 501–514.

[31] W. A. Bartlett, F. Braga, A. Carobene, A. Coşkun, R. Prusa, P. Fernandez-Calle, T. Røraas, N. Jonker, S. Sandberg, B. V. W. Group, A checklist for critical appraisal of studies of biological variation, Clinical Chemistry and Laboratory Medicine (CCLM) 53 (2015) 879–885.

[32] M. Mannino, Y. Yang, Y. Ryu, Classification algorithm sensitivity to training data with non representative attribute noise, Decision Support Systems 46 (2009) 743–751.

[33] D. A. Van Dyk, X.-L. Meng, The art of data augmentation, Journal of Computational and Graphical Statistics 10 (2001) 1–50.

[34] J. Lienen, E. Hüllermeier, Instance weighting through data imprecisiation, International Journal of Approximate Reasoning 134 (2021) 1–14.

[35] F. Cabitza, A. Campagner, D. Ferrari, C. Di Resta, D. Ceriotti, E. Sabetta, A. Colombini, E. De Vecchi, G. Banfi, M. Locatelli, A. Carobene, Development, evaluation, and validation of machine learning models for covid-19 detection based on routine blood tests, Clinical Chemistry and Laboratory Medicine (CCLM) 59 (2021) 421–431.

[36] L. Famiglini, G. Bini, A. Carobene, A. Campagner, F. Cabitza, Prediction of icu admission for covid-19 patients: a machine learning approach based on complete blood count data, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2021, pp. 160–165.

[37] H. Fröhlich, R. Balling, N. Beerenwinkel, O. Kohlbacher, S. Kumar, T. Lengauer, M. H. Maathuis, Y. Moreau, S. A. Murphy, T. M. Przytycka, From hype to reality: data science enabling personalized medicine, BMC medicine 16 (2018) 1–15.

[38] C. Ellervik, J. Vaught, Preanalytical variables affecting the integrity of human biospecimens in biobanking, Clinical chemistry 61 (2015) 914–934.

[39] A. Carobene, E. Guerra, M. Locatelli, F. Ceriotti, S. Sandberg, P. Fernandez-Calle, A. Coşkun, A. K. Aarsand, Providing correct estimates of biological variation—not an easy task. the example of s100-β protein and neuron-specific enolase, Clinical Chemistry 64 (2018) 1537–1539.

[40] R. Haeckel, A. Carobene, W. Wosniok, Problems with estimating reference change values (critical differences), Clinica Chimica Acta 523 (2021) 437–440.

[41] A. K. Aarsand, A. H. Kristoffersen, S. Sandberg, B. Støve, A. Coşkun, P. Fernandez-Calle, J. Díaz-Garzón, E. Guerra, F. Ceriotti, N. Jonker, The european biological variation study (EuBIVAS): Biological variation data for coagulation markers estimated by a bayesian model, Clinical Chemistry 67 (2021) 1259–1270.

[42] T. Røraas, P. H. Petersen, S. Sandberg, Confidence intervals and power calculations for within-person biological variation: effect of analytical imprecision, number of replicates, number of samples, and number of individuals, Clinical chemistry 58 (2012) 1306–1313.

[43] A. Carobene, M. Strollo, N. Jonker, G. Barla, W. A. Bartlett, S. Sandberg, M. S. Sylte, T. Røraas, U. Ø. Sølvik, P. Fernandez-Calle, Sample collections from healthy volunteers for biological variation estimates' update: a new

project undertaken by the working group on biological variation established by the european federation of clinical chemistry and laboratory medicine, Clinical Chemistry and Laboratory Medicine (CCLM) 54 (2016) 1599–1608.

[44] A. K. Aarsand, P. Fernandez-Calle, C. Webster, A. Coskun, E. Gonzales-Lao, J. Diaz-Garzon, N. Jonker, J. Minchinela, M. Simon, F. Braga, The EFLM biological variation database, 2020. URL: `https://biologicalvariation.eu/`.

[45] S. Sandberg, A. Carobene, A. K. Aarsand, Biological variation–eight years after the 1st strategic conference of EFLM, Clinical Chemistry and Laboratory Medicine (CCLM) (2022).

[46] T. Herlau, M. N. Schmidt, M. Mørup, Completely random measures for modelling block-structured sparse networks, Advances in Neural Information Processing Systems 29 (2016).

[47] O. Kallenberg, Random measures, Theory and applications, Springer, 2017.

[48] A. Coskun, F. Braga, A. Carobene, X. T. Ganduxe, A. K. Aarsand, P. Fernández-Calle, J. D.-G. Marco, W. Bartlett, N. Jonker, B. Aslan, et al., Systematic review and meta-analysis of within-subject and between-subject biological variation estimates of 20 haematological parameters, Clinical Chemistry and Laboratory Medicine (CCLM) 58 (2020) 25–32.

[49] S. Buoro, A. Carobene, M. Seghezzi, B. Manenti, A. Pacioni, F. Ceriotti, C. Ottomano, G. Lippi, Short-and medium-term biological variation estimates of leukocytes extended to differential count and morphology-structural parameters (cell population data) in blood samples obtained from healthy people, Clinica Chimica Acta 473 (2017) 147–156.

[50] S. Buoro, M. Seghezzi, B. Manenti, A. Pacioni, A. Carobene, F. Ceriotti, C. Ottomano, G. Lippi, Biological variation of platelet parameters determined by the Sysmex XN hematology analyzer, Clinica Chimica Acta 470 (2017) 125–132.

[51] S. Buoro, A. Carobene, M. Seghezzi, B. Manenti, P. Dominoni, A. Pacioni, F. Ceriotti, C. Ottomano, G. Lippi, Short-and medium-term biological variation estimates of red blood cell and reticulocyte parameters in healthy subjects, Clinical Chemistry and Laboratory Medicine (CCLM) 56 (2018) 954–963.

[52] K. Zhou, Y. Sun, L. Li, Z. Zang, J. Wang, J. Li, J. Liang, F. Zhang, Q. Zhang, W. Ge, et al., Eleven routine clinical features predict covid-19 severity uncovered by machine learning of longitudinal measurements, Computational and structural biotechnology journal 19 (2021) 3640–3649.

[53] F. Cabitza, A. Campagner, F. Soares, L. G. de Guadiana-Romualdo, F. Challa, A. Sulejmani, M. Seghezzi, A. Carobene, The importance of

being external. methodological insights for the external validation of machine learning models in medicine, Computer Methods and Programs in Biomedicine 208 (2021) 106288.

[54] L. Famiglini, A. Campagner, A. Carobene, F. Cabitza, A robust and parsimonious machine learning method to predict icu admission of covid-19 patients, Medical & Biological Engineering & Computing (2022) 1–13.

[55] S. A. Fayaz, M. Zaman, S. Kaul, M. A. Butt, Is deep learning on tabular data enough? an assessment, International Journal of Advanced Computer Science and Applications 13 (2022).

[56] R. Shwartz-Ziv, A. Armon, Tabular data: Deep learning is not all you need, Information Fusion 81 (2022) 84–90.

[57] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on tabular data?, arXiv preprint arXiv:2207.08815 (2022).

[58] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, G. Kasneci, Deep neural networks and tabular data: A survey, IEEE Transactions on Neural Networks and Learning Systems (2022).

[59] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. Williamson, F. Mahmood, Synthetic data in machine learning for medicine and healthcare, Nature Biomedical Engineering 5 (2021) 493–497.

[60] T. Denœux, D. Dubois, H. Prade, Representations of uncertainty in artificial intelligence: Probability and possibility, in: A Guided Tour of Artificial Intelligence Research, Springer, 2020, pp. 69–117.

[61] K. Zheng, P. C. Fung, X. Zhou, K-nearest neighbor search for fuzzy objects, in: Proceedings of the 2010 ACM SIGMOD international conference on Management of data, 2010, pp. 699–710.

[62] K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, Kernel mean embedding of distributions: A review and beyond, Foundations and Trends® in Machine Learning 10 (2017) 1–141.

[63] A. Seveso, A. Campagner, D. Ciucci, et al., Ordinal labels in machine learning: a user-centered approach to improve data validity in medical settings, BMC Medical Informatics and Decision Making 20 (2020) 1–14.

[64] D. Dubois, H. Prade, S. Sandri, On possibility/probability transformations, in: Fuzzy logic, Springer, 1993, pp. 103–112.

[65] A. Grønlund, L. Kamma, K. Green Larsen, Margins are insufficient for explaining gradient boosting, Advances in Neural Information Processing Systems 33 (2020) 1902–1912.

[66] S. Hanneke, A. Kontorovich, Stable sample compression schemes: New applications and an optimal svm margin bound, in: Algorithmic Learning Theory, PMLR, 2021, pp. 697–721.

[67] M. Hildebrandt, The issue of proxies and choice architectures. why eu law matters for recommender systems, Frontiers in Artificial Intelligence (2022) 73.

[68] P. Boyeau, J. Hong, A. Gayoso, M. Jordan, E. Azizi, N. Yosef, Deep generative modeling for quantifying sample-level heterogeneity in single-cell omics, bioRxiv (2022) 2022–10.

[69] N. Thams, M. Oberst, D. Sontag, Evaluating robustness to dataset shift via parametric robustness sets, arXiv preprint arXiv:2205.15947 (2022).

[70] A. Campagner, Learnability in "learning from fuzzy labels", in: 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2021, pp. 1–6.

[71] G. Ma, F. Liu, G. Zhang, J. Lu, Learning from imprecise observations: An estimation error bound based on fuzzy random variables, in: 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2021, pp. 1–8.

[72] H. Hotelling, The generalization of student's ratio, in: Breakthroughs in statistics, Springer, 1992, pp. 54–65.

[73] R. Arratia, L. Gordon, Tutorial on large deviations for the binomial distribution, Bulletin of mathematical biology 51 (1989) 125–131.

Appendix A. Appendix A: Analysis of the WSF Algorithm

Pseudo-code for the WSF algorithm is reported in Algorithm 2, in the main text. As described in the main text, the computational complexity of WSF is $O(n\,d\,|S|\log(|S|))$ where d is the dimensionality of the input space.

In regard to the generalization error of WSF w.r.t. the data generating random measure, for each base model h_{it}, let

$$L_S(h_{it}) = \sum_{(x,y)\in S} \mathbb{E}_{(x^0,y)\ \ is_{poss}(x,y)} \left[\mathbb{1}_{h(x^0)\neq y}\right]$$

and $L_D(h_{it}) = \mathbb{E}_{S\ \ D^m} L_S(h)$, where D is the intensity measure describe in Section 3.1. Assume further, that for all $h \in H$, with probability larger than $1-\delta$ if $(x-x^0)^{\top}(x-x^0) \leq T^2_{d,|S|}{}_{d}(1-\delta)$ it holds that $h(x) = h(x^0)$, where T is Hotelling's T-squared distribution [72]. Intuitively, this latter condition can be understood as a strong form of *stability* [29] for models in H: if two instantiations likely come from the same distribution due to IV, then with high probability they will be classified in the same way by each $h \in H$. Then, letting

V_{it} be the out-of-bag sample for model h_{it}, by Hoeffding's inequality and above assumptions it follows that, with probability $1 - \delta$, it holds that

$$L_D(h_{it}) - L_{V_{it}}(h_{it}) + \sqrt{\frac{\log(2|V_{it}|/\delta)}{2|V_{it}|}}. \tag{A.1}$$

Thus, the expected error of each model h_{it} in the WSF ensemble is close (with high probability) to the respective out-of-bag-sample estimate, as long as the dataset size is big enough. The per-base model error estimate calculated above can also be directly used to provide an estimate for the expected error of the WSF model. Let $\rho = \sum_{it} L_{V_{it}}(h_{it}) + \sqrt{\frac{\log(2|V_{it}|/\delta)}{2|V_{it}|}} - \frac{1}{2}$. Intuitively, ρ represents an upper bound on the joint probability of error the base models, which is simply obtained by an application of the union bound [29]. Then, assuming the h_{it} err independently of each other, and noting that WSF errs on an instance x iff at least $ens/2$ base models err, with probability greater than $1 - \frac{\delta}{ens}$ the generalization error of WSF can be upper bounded through an application of Chernoff's bound for binomial distributions [73] by $e^{-ens\,KL(\frac{1}{2}||\rho)}$, where $KL(a||b) = a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b}$ is the Kullback-Leibler divergence. Thus, intuitively, the expected error of WSF decreases rapidly with the number of ensembled base models, as long as their total error is small and they are independent: this theoretical results, thus, explains the good robustness to IV exhibited by WSF on the considered task.