ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Towards Randomized Algorithms and Models that We Can Trust: a Theoretical Perspective

Luca Oneto, Sandro Ridella, and Davide Anguita *

University of Genoa - Via Opera Pia 11a, 16145, Genova, Italy

Abstract. In the last decade it became increasingly apparent the inability of technical metrics to well characterize the behavior of intelligent systems. In fact, they are nowadays requested to meet also ethical requirements such as explainability, fairness, robustness, and privacy increasing our trust in their use in the wild. The final goal is to be able to develop a new generation of more responsible and trustworthy machine learning. In this paper, we focus our attention on randomized machine learning algorithms and models questioning, from a theoretical perspective, if it is possible to simultaneously optimize multiple metrics that are in tension between each other towards randomized machine learning algorithms that we can trust. For this purpose we will leverage the most recent advances coming from the statistical learning theory: distribution stability and differential privacy.

1 Introduction

In the last decade several methodological and technological breakthroughs in Machine Learning (ML) (e.g., Deep Learning and Large Language Models) have remarkably changed it [1, 2]. The consequence of this advancement invested research, industry, entertainment, and society at large making apparent the inability of the technical metrics [3–5] (e.g., accuracy, computational requirements, and non-regressivity) to well characterize the behavior of ML in the wild. The daily use of ML-based systems also calls for ethical requirements [6–9] (e.g., explainability, fairness, robustness, and privacy) to increase the trust in their use. In fact, many researchers have exposed many different reasons not to trust ML, especially its lack of ethical boundaries [10].

As a consequence, researchers started to study, theoretically and empirically, the ethical properties of ML proposing solution for building more explainable, fair, robust, and privacy aware model often focusing on a single or maximum two of these aspects [10, 11]. In this paper, we try to investigate what happens when multiple technical and ethical metrics, that are often in tension between each other, need to be simultaneously optimized. With this goal in mind we focus the attention on Randomized Model (RM) and Randomized Algorithm (RA) [12, 13] and we will leverage probably two of the powerful Statistical Learning Theory (SLT) [12, 13], i.e., Algorithmic Stability (AS) and Differential Privacy (DP), to show that it is possible to learn trustworthy RM and develop RA with consistency results, i.e., models that exhibit on the population technical and ethical

*This work is supported in part by ELSA – European Lighthouse on Secure and Safe AI funded by the European Union under grant agreement No. 101070617. This work is also been partially supported by Spoke 10 “Logistics and Freight” within the Italian PNRR National Centre for Sustainable Mobility (MOST), CUP I53C22000720001. This work is also partially supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

performance that are close to the one imposed on the training data [3]. Specifically in Section 2 we will formally define the notion of trustworthy RA and RM, in Section 3 we will derive consistency result on them, and finally Section 4 concludes the paper.

2 Randomized Algorithms and Models that We Can Trust

Let us consider the problem of learning, in a randomized way, a more trustworthy binary classifier[1]. More specifically we would like to build an algorithms and learn a model able to simultaneously exhibit good technical [3–5] (i.e., accuracy, computational requirements, and non-regressiveness) and ethical (i.e., explainability, fairness, robustness, and privacy) [11, 14] properties. We will focus our attention on RA and RM, namely algorithms that may return a different model when repeating the learning problem and models that may assign different labels to the same input if we repeat the labeling process [12, 13].

In this setting, let $\mathcal{D}{=}\{(x_1, s_1, y1), \cdots, (x_n, s_n, y_n)\}$ be a sequence of n samples drawn independently from an unknown probability distribution $\mathfrak{P}_{\mathcal{X}\times\mathcal{S}\times\mathcal{Y}}$ over $\mathcal{X}{\times}\mathcal{S}{\times}\mathcal{Y}$, where $\mathcal{Y}{=}\{\pm 1\}$ is the set of binary output labels, $\mathcal{S}{=}\{\pm 1\}$ represents group membership[2] (e.g., gender or ethnicity), and $\mathcal{X}$ is the input space. Let us consider a model $f{:}\mathcal{Z}{\rightarrow}\hat{\mathcal{Y}}{\subseteq}\mathbb{R}$ that exploits the information in the input space $\mathcal{Z}{=}\mathcal{X}$ (and possibly the group membership if allowed by the legislation $\mathcal{Z}{=}\mathcal{X}{\times}\mathcal{S}$ [15]) to label it.

From a technical perspective, f is chosen with the purpose to optimize a series of practical requirements.

In the simplest case, f is chosen inside a set of possible models $\mathcal{F}$ to minimize its error [3]

$$\mathtt{A}(f){=}\mathbb{E}_{z,y}\{\ell^{\mathtt{A}}(f, z, y)\}, \tag{1}$$

with $z \in \mathcal{Z}$ where $\ell^{\cdots}{:}\mathcal{F}{\times}\mathcal{Z}{\times}\mathcal{Y}{\rightarrow}[0, 1]$ is a loss function that measures the effectiveness of f in approximating $\mathbb{P}_{z,y}\{y|z\}$. This choice is performed by an algorithm $\mathscr{A}_{\mathcal{H}}$, characterized by a set of hyperparameters $\mathcal{H}$, based on $\mathcal{D}$: $\mathscr{A}_{\mathcal{H}}$ and $\mathcal{F}$ may be explicitly (e.g., linear models) but also implicitly (e.g., k-rule) related.

If we deal with RA $f{=}\mathscr{A}_{\mathcal{H}}(\mathcal{D})$, i.e., the model chosen by $\mathscr{A}_{\mathcal{H}}$ based on $\mathcal{D}$ may be different, this association is non-deterministic and then there is a $\mathfrak{P}_{\mathscr{A}_{\mathcal{H}}}$ over $\mathcal{F}$ given $\mathcal{D}$. Instead, if we deal with RM (Gibbs Classifiers), $\mathscr{A}_{\mathcal{H}}$ does not return a model but a probability distribution $\mathfrak{P}_f{=}\mathscr{A}_{\mathcal{H}}(\mathcal{D})$ over $\mathcal{F}$ such each time a label for an input $z \in \mathcal{Z}$ need to be labeled f is sampled from $\mathcal{F}$ according to $\mathfrak{P}_f$ and then $f(z)$ is computed possibly resulting in different label if the same sample is labeled multiple times.

Another technical metric, coming from the world of software engineering, is the non-regressivity [5]. In this case we suppose that the learning process is an iterative procedure of updates to the last version of the model. As a consequence what we have is the last model f^l that we want to update via another model f

[1] The presentation can be extended to the whole supervised learning setting, we do not report it here for simplicity.

[2] The presentation can be extended to multiple groups, we do not report it here for simplicity.

minimizing the following quantity

$$\mathtt{N}(f){=}\mathbb{E}_{z,y:\ell^{\cdots}(f^{l},z,y)\leq\varepsilon}\{\ell^{\cdots}(f,z,y)\}, \tag{2}$$

namely we do not want to introduce large error on sub-regions of $\mathcal{Z}$ where the last model f^{l} performed well. For example, by setting $\ell^{\cdots}(f,z,y){=}[yf(z)\leq 0]$, namely when using the Hard loss function, $\mathtt{N}(f)$ represent the Negative Flips [5].

A metric, which is harder to define in an unique way, is the sustainability (e.g., measured in terms of computing power and carbon emission) of the model training or forward phases [4]. This goal is achieved through mainly two approaches. The first one acts on the optimization process, i.e., the algorithm, which aims at finding the minimum of the interested metrics focusing on developing more efficient optimization algorithms [4]. The second approach focuses on $\mathcal{F}$ defining function space $\mathcal{F}^{\mathtt{S}}$, i.e., an algorithm, that prefers models with limited computational requirements, e.g., models that can be represented with fewer bits or models that exploit the minimum number of samples in $\mathcal{D}$ or or a minimum subspace of $\mathcal{X}$ [4, 16].

We could continue with other technical metrics of the models but we stop here since we are more interested in ethical metrics.

In fact, apart from the technical perspective there is also an ethical perspective that needs to be taken into account in order to trust f chosen by $\mathscr{A}_{\mathcal{H}}$. In fact, taking care of the technical requirements is not enough and we need to take care of the ethical requirements to enforce trust in or algorithms and models which need to be also, e.g., explainable, fair, and robust while preserving the privacy of the data of the individuals in $\mathcal{D}$ [11, 14]. Note also that each one of these requirements many different metrics can be exploited but general principles can be outlined.

Regarding explainability we can explain how model works (global explainability) or the individual predictions (local explainability) [6]. Regarding global explainability we need to act on the structure of $\mathcal{F}$ (and then $\mathscr{A}_{\mathcal{H}}$) selecting space of models belonging to certain classes that are more explainable (e.g., rule based or linear) $\mathcal{F}^{\mathtt{E}}$ [6]. Regarding local explainability most approaches (e.g., LIME or Grad-CAM) basically try to approximate f around z with an interpretable model, i.e., a model in $\mathcal{F}^{\mathtt{E}}$ [6]. More formally we would like our model to also minimize the following quantity

$$\mathtt{E}(f){=}\mathbb{E}_{z,y}\{\ell^{\mathtt{E}}(f,z,y)\}{=}\mathbb{E}_{z,y}\{\max_{\tilde{z}\in\mathcal{L}(z)}\min_{\tilde{f}\in\mathcal{F}^{\mathtt{E}}}|\ell^{\cdots}(f(\tilde{z}),y){-}\ell^{\cdots}(\tilde{f}(\tilde{z}),y)|\}, \tag{3}$$

namely, we would like the model f to be well approximable, locally in the sense of $\mathcal{L}(z)$, by an explainable model in $\mathcal{F}^{\mathtt{E}}$.

Regarding fairness, the underlying idea is quite simple: the model should not behave differently if applied to subgroups of the population [7]. More specifically what we want to minimize is the following quantity

$$\mathtt{F}(f){=}|\mathbb{E}_{z,y:\mathtt{G}_{-1}(z,y)}\{\ell^{\cdots}(f,z,\dot{y})\}{-}\mathbb{E}_{z,y:\mathtt{G}_{+1}(z,y)}\{\ell^{\cdots}(f,z,\dot{y})\}|, \tag{4}$$

where if we set $\ell^{\cdots}(f,z,y){=}[f(z)\leq 0]$, $\mathtt{G}_{\cdot}(z,y){=}[s{=}\cdot]$, and $\dot{y}{=}1$ we get as $\mathtt{F}(f)$ the Demographic Parity while if we set $\ell^{\cdots}(f,z,y){=}[f(z)\leq 0]$, $\mathtt{G}_{\cdot}(z,y){=}[s{=}\cdot,y{=}1]$, and $\dot{y}{=}y$ we get as $\mathtt{F}(f)$ the Equal Opportunity [7, 17].

Regarding robustness, what we want is that the labeling process should not be influenced by small, natural or adversarial, perturbations of the input, namely the model should not be induced into mistakes by slightly modifying the input data [8]. More formally we would like our model to also minimize the following quantity

$$\mathtt{R}(f){=}\mathbb{E}_{z,y}\{\ell^{\mathtt{R}}(f,z,y)\}{=}\mathbb{E}_{z,y}\{\max_{\tilde{z}\in\mathcal{P}(z)}\ell^{\cdots}(f(\tilde{z}),y)\}, \tag{5}$$

where $\mathcal{P}(z)$ are all the possible or admissible perturbations of z [8].

Finally, in order to preserve the privacy of the individual in $\mathcal{D}$ we have multiple options [9]. Surely using homomorphic encryption is able to maximize both privacy and utility but the associated computational overhead is often prohibitive [9]. Differential Privacy, instead, showed to be a good option to balance privacy and utility [12, 18]. RA are actually ϵ-DP if

$$\mathbb{P}\{\mathscr{A}_{\mathcal{H}}(\mathcal{D})\in\breve{\mathcal{F}}\}\leq \mathbb{e}^{\epsilon}\mathbb{P}\{\mathscr{A}_{\mathcal{H}}(\mathcal{D}\setminus(x_i,s_i,y_i))\in\breve{\mathcal{F}}\},\forall i\in\{1,\cdots,n\},\breve{\mathcal{F}}\subseteq\mathcal{F}. \tag{6}$$

namely the larger is ϵ the higher the ability to ensure privacy at the expense of utility [12].

3 Consistency Results

In this section we will show that it is possible to build a RA and a RM from data able to exhibit all the properties depicted in Section 2. More specifically we will show that for a particular choice of $\mathfrak{P}_{\mathscr{A}_{\mathcal{H}}}$ and $\mathfrak{P}_f$ the resulting model empirically exhibiting certain desired levels of technical and ethical metrics will also behave consistently on the population.

In our setting a good f is something that is able to simultaneously optimized the metric defined in Section 2, i.e., the technical ones $\mathtt{A}(f)$ and $\mathtt{N}(f)$ in $\mathcal{F}^{\mathtt{S}}$ and the ethical ones $\mathtt{E}(f)$, $\mathtt{F}(f)$, and $\mathtt{R}(f)$ in a $\mathcal{F}^{\mathtt{E}}$ exhibiting the DP property. In the non-randomized setting we then would like a function such that

$$f^*\colon \arg\min\nolimits_{\mathcal{F}^{\mathtt{S}}\cap\mathcal{F}^{\mathtt{E}}} \lambda^{\mathtt{A}}\mathtt{A}(f){+}\lambda^{\mathtt{N}}\mathtt{N}(f){+}\lambda^{\mathtt{E}}\mathtt{E}(f){+}\lambda^{\mathtt{F}}\mathtt{F}(f){+}\lambda^{\mathtt{R}}\mathtt{R}(f), \tag{7}$$

where $\lambda^{\mathtt{M}}{\in}[0,\infty)$ with $\mathtt{M}\in\{\mathtt{A},\mathtt{N},\mathtt{E},\mathtt{F},\mathtt{R}\}$ regulate the trade-off between the different metrics that are of-course in tension between each others [11, 14] and where f^* does not disclose information about the individuals in $\mathcal{D}$. The larger is $\lambda^{\mathtt{M}}$ the more important for us is the metric $\mathtt{M}$ and vice-versa. Since $\mathfrak{P}_{\mathcal{X}\times\mathcal{S}\times\mathcal{Y}}$ is unknowns we cannot find f^*, what we can find is its empirical estimator, namely

$$\hat{f}\colon \arg\min\nolimits_{\mathcal{F}^{\mathtt{S}}\cap\mathcal{F}^{\mathtt{E}}} \lambda^{\mathtt{A}}\hat{\mathtt{A}}(f){+}\lambda^{\mathtt{N}}\hat{\mathtt{N}}(f){+}\lambda^{\mathtt{E}}\hat{\mathtt{E}}(f){+}\lambda^{\mathtt{F}}\hat{\mathtt{F}}(f){+}\lambda^{\mathtt{R}}\hat{\mathtt{R}}(f), \tag{8}$$

where $\hat{\mathtt{A}}(f)$, $\hat{\mathtt{N}}(f)$, $\hat{\mathtt{E}}(f)$, $\hat{\mathtt{F}}(f)$, and $\hat{\mathtt{R}}(f)$ are the empirical conterparts of $\mathtt{A}(f)$, $\mathtt{N}(f)$, $\mathtt{E}(f)$, $\mathtt{F}(f)$, and $\mathtt{R}(f)$ respectively, i.e., the metrics computed using $\mathcal{D}$ instead of $\mathfrak{P}_{\mathcal{X}\times\mathcal{S}\times\mathcal{Y}}$. Inspired by this let us consider the following empirical distribution on $f{\in}\mathcal{F}^{\mathtt{S}}{\cap}\mathcal{F}^{\mathtt{E}}$

$$\mathtt{q}(f){=}\mathtt{Z}_{\mathtt{q}}\mathbb{e}^{-\gamma\left[\lambda^{\mathtt{A}}\hat{\mathtt{A}}(f)+\lambda^{\mathtt{N}}\hat{\mathtt{N}}(f)+\lambda^{\mathtt{E}}\hat{\mathtt{E}}(f)+\lambda^{\mathtt{F}}\hat{\mathtt{F}}(f)+\lambda^{\mathtt{R}}\hat{\mathtt{R}}(f)\right]}, \mathtt{Z}_{\mathtt{q}}^{-1}{=}\textstyle\int_{\mathcal{F}^{\mathtt{S}}\cap\mathcal{F}^{\mathtt{E}}}\mathtt{q}(f)df, \tag{9}$$

where $\gamma{\in}[0,\infty)$ and let us use it as $\mathfrak{P}_{\mathscr{A}_{\mathcal{H}}}$ for a RA or as $\mathfrak{P}_f$ for a RM. Basically $\mathtt{q}(f)$ weighs $\hat{f}$ and exponentially less than the other ones based on their distance,

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

in terms of cost function, from $\hat{f}$. The distribution of these weights is regulated by γ. The larger is γ the more weight is associated with the functions that we like, i.e., good both in terms of technical and ethical metrics. As a consequence, our desire would be to have γ as large as possible but this, as we will see soon, will not be allowed if we want to maintain consistency.

At this point we can prove our consistency results, i.e., that the empirical metrics $\hat{\mathtt{M}}$ are actually close to their value computed on the population $\mathtt{M}$ for a RA or RM based on the empirical distribution defined in Eq. (9). For this purpose, in the case of a randomized model, we can rely on the AS Theory [13, 17], a state-of-the-art SLT, to derive the following consistency result.

Theorem 1. *The metrics $\mathtt{M} \in \mathcal{M} = \{\mathtt{A}, \mathtt{N}, \mathtt{E}, \mathtt{F}, \mathtt{R}\}$ of a RM r which uses the empirical distribution defined in Eq.* (9) *can be bounded, for some positive universal constants c_1, c_2, and c_3, as follows*

$$\mathbb{P}\left\{|\mathtt{M}(r){-}\hat{\mathtt{M}}(r)|{\geq}{}^{c_1\gamma}/_n \textstyle\sum_{\mathtt{M}\in\mathcal{M}} \lambda^{\mathtt{M}}{+}c_2\gamma\sqrt{{}^{\ln{(c_3/\delta)}}/_n}\left(\textstyle\sum_{\mathtt{M}\in\mathcal{M}} \lambda^{\mathtt{M}}{+}1\right)\right\}{\leq}\delta. \tag{10}$$

The proof is not reported here because of space constraints but the idea, even if technically challenging, is simple: using the notion of Uniform Distribution Stability, and the related finite sample bounds, it is possible to bound the change in the distribution of Eq. (9) when changing one sample in $\mathcal{D}$[3] and then to bound the deviating between the empirical metrics and their value on the population [13, 17].

Theorem 1 allows us to state that, for a γ that does not increases faster than $O(\sqrt{n})$, the RM based on empirical distribution defined in Eq. (9) produces technically and ethically consistent prediction with the deception of the fact that this model does not respect the privacy of the individuals in $\mathcal{D}$. In order to address this point, let us consider the RA based on the empirical distribution defined in Eq. (9) and, relying on the DP Theory [12, 18], a state-of-the-art SLT, to derive the following consistency result.

Theorem 2. *Let us consider the RA which, given a dataset $\mathcal{D}$, selects a function $f{\in}\mathcal{F}^{\mathtt{S}}{\cap}\mathcal{F}^{\mathtt{E}}$ according to the empirical distribution defined in Eq.* (9)*. This algorithm is ϵ-DP with $\epsilon = {}^{\gamma}/_n \sum_{\mathtt{M}\in\mathcal{M}} c^{\mathtt{M}}\lambda^{\mathtt{M}}$ for some positive universal constants $c^{\mathtt{M}}$ and by setting $\gamma \leq {}^{\sqrt{c_1/n}}/_{(c_2 \sum_{\mathtt{M}\in\mathcal{M}} \lambda^{\mathtt{M}}/n)}$ for some positive universal constants c_1 and c_2, it is possible to bound the metrics $\mathtt{M} \in \mathcal{M} = \{\mathtt{A}, \mathtt{N}, \mathtt{E}, \mathtt{F}, \mathtt{R}\}$, for some positive universal constants $c_1^{\mathtt{M}}$ and $c_2^{\mathtt{M}}$, as follows*

$$\mathbb{P}\left\{|\mathtt{M}(r){-}\hat{\mathtt{M}}(r)|{\geq}\sqrt{{}^{c_1^{\mathtt{M}}\ln{(c_2^{\mathtt{M}}/\delta)}}/_n}\right\}{\leq}\delta. \tag{11}$$

Also in this case the proof is not reported here because of space constraints but the idea is the same behind Theorem 1 but, in this case, we exploited the bound bases on the finite sample empirical bound based on the DP Theory [12, 18].

Theorem 2 allows us to state that, for a γ that does not increases faster than $O(\sqrt{n})$, the RA based on empirical distribution defined in Eq. (9) produces technically and ethically consistent prediction in the sense described in Section 2: accurate, non-regressive, sustainable, explainable, fair, and robust preserving the privacy of the individuals in $\mathcal{D}$.

[3] This observation leaves room for the extension of this work to other, well behaving, metrics.

4 Conclusions

In this paper we presented a novel theoretical perspective to the problem of building RA or learning RM that try to simultaneously optimize multiple technical (e.g., accuracy, computational requirements, and non-regressivity) and ethical (e.g., explainability, fairness, robustness, and privacy) metrics, toward more trustworthy ML, deriving a series of consistency result where it is show that forcing, in a particular way, these properties on the training data ensures good performance also on the population.

This work is surely a fist step forward in the direction of building more holistic perspective to the problem of trustworthy ML but it shed some light on an important problem that is becoming every day more and more urgent as many ML breakthroughs are reaching society at large increasing the cancers on the potential unwanted ethical impact of the these technologies.

References

[1] D. Silver, A. Huang, C. J. Maddison, A. Guez, and Others. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[2] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

[4] A. Van Wynsberghe. Sustainable ai: Ai for sustainability and the sustainability of ai. *AI and Ethics*, 1(3):213–218, 2021.

[5] S. Yan, Y. Xiong, K. Kundu, S. Yang, S. Deng, M. Wang, W. Xia, and S. Soatto. Positive-congruent training: Towards regression-free model updates. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[6] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, and Others. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[7] D. Pessach and E. Shmueli. A review on fairness in machine learning. *ACM Computing Surveys*, 55(3):1–44, 2022.

[8] B. Biggio and F. Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

[9] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys*, 54(2):1–36, 2021.

[10] B. Li, P. Qi, B. Liu, S. Di, and Others. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023.

[11] L. Oneto, N. Navarin, B. Biggio, F. Errica, and Others. Towards learning trustworthily, automatically, and with guarantees on graphs: An overview. *Neurocomputing*, 493:217–243, 2022.

[12] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[13] A. Elisseeff, T. Evgeniou, M. Pontil, and L. P. Kaelbing. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(1), 2005.

[14] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi. Trustworthy artificial intelligence: a review. *ACM Computing Surveys*, 55(2):1–38, 2022.

[15] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on fairness, accountability and transparency*, 2018.

[16] L. Oneto, A. Ghio, S. Ridella, and D. Anguita. Learning resource-aware classifiers for mobile devices: from regularization to energy efficiency. *Neurocomputing*, 169:225–235, 2015.

[17] L. Oneto, M. Donini, M. Pontil, and J. Shawe-Taylor. Randomized learning and generalization of fair and private classifiers: From pac-bayes to stability and differential privacy. *Neurocomputing*, 416:231–243, 2020.

[18] L. Oneto, S. Ridella, and D. Anguita. Differential privacy and generalization: Sharper bounds with applications. *Pattern Recognition Letters*, 89:31–38, 2017.