Machine Learning (2024) 113:2701–2731
https://doi.org/10.1007/s10994-024-06515-0

Fair and green hyperparameter optimization via multi-objective and multiple information source Bayesian optimization

Antonio Candelieri[1] · Andrea Ponti[1,2] · Francesco Archetti[3]

Received: 20 August 2022 / Revised: 12 December 2023 / Accepted: 18 January 2024 /
Published online: 28 February 2024

Abstract

It has been recently remarked that focusing only on accuracy in searching for optimal Machine Learning models amplifies biases contained in the data, leading to unfair predictions and decision supports. Recently, multi-objective hyperparameter optimization has been proposed to search for Machine Learning models which offer equally Pareto-efficient trade-offs between accuracy and fairness. Although these approaches proved to be more versatile than fairness-aware Machine Learning algorithms—which instead optimize accuracy constrained to some threshold on fairness—their carbon footprint could be dramatic, due to the large amount of energy required in the case of large datasets. We propose an approach named FanG-HPO: fair and green hyperparameter optimization (HPO), based on both multi-objective and multiple information source Bayesian optimization. FanG-HPO uses subsets of the large dataset to obtain cheap approximations (aka information sources) of both accuracy and fairness, and multi-objective Bayesian optimization to efficiently identify Pareto-efficient (accurate and fair) Machine Learning models. Experiments consider four benchmark (fairness) datasets and four Machine Learning algorithms, and provide an assessment of FanG-HPO against both fairness-aware Machine Learning approaches and two state-of-the-art Bayesian optimization tools addressing multi-objective and energy-aware optimization.

Keywords Fair Machine Learning · Green Machine Learning · Multi-objective optimization · Multiple information source Bayesian optimization

Editor: Rita P. Ribeiro.

Andrea Ponti and Francesco Archetti have contributed equally to this work.

Extended author information available on the last page of the article

1 Introduction

1.1 Rationale and motivations

A low misclassification/prediction error is not the only performance metric of interest in searching for the most suitable Machine Learning (ML) model to use in a successful decision support application. Additional metrics like *fairness*, *interpretability*, and *privacy* have been increasingly becoming important during last years. This paper focuses on *fairness*, a desired property of the decision support provided by a ML model: it must not be "biased" towards a specific person or groups of individuals (Barocas et al., 2017; Buolamwini & Gebru, 2018; Pessach & Shmueli, 2022). The topic is known as *FairML* (Mehrabi et al., 2021), with approaches organized into three different families: (i) *post-processing* to modify a pre-trained model to increase the fairness of its outcomes, (ii) *in-processing* to enforce fairness constraints during training, and (iii) *pre-processing* to modify the data representation and then apply standard ML algorithms (Friedler et al., 2019; Hort et al., 2022). These approaches regard the design of *fairness-aware* (or *fair-by-design*) ML algorithms, but they suffer from one or more of the following drawbacks (Perrone et al., 2021): the intervention performed to deal with biases is (*i*) specific to the model class (e.g., linear models only), (ii) limited to a specific definition of fairness, (iii) limited to a single, binary sensitive feature, (iv) requires access to sensitive feature information at prediction time, and (v) results in a randomized classifier that may generate different prediction for the same input at different times.

These considerations have recently led to a new strategy: instead of designing fairness-aware ML algorithm, FairML can be addressed as the hyperparameter optimization (HPO) of a ML algorithm, by also considering fairness. Two mechanisms have been recently proposed to include fairness into HPO: (*i*) **constrained optimization**—aimed at minimizing the misclassification error while satisfying some fairness constraint (Perrone et al., 2020, 2021) – and (*ii*) **multi-objective optimization**—aimed at minimizing, simultaneously, misclassification error and some *unfairness* metric (Schmucker et al., 2020).

More recently, the topic has been named *Fairness-aware AutoML* (Weerts et al., 2023), where automated Machine Learning (AutoML) generically refers to automatizing the design of a ML pipeline (aka workflow), in which HPO of ML algorithms is a specific task (Hutter et al., 2019; He et al., 2021).

It is important to remark that fairness-constrained HPO could not be viable in many real-life settings, because a suitable threshold on fairness could be difficult to be established a-priori. According to this consideration, and to recent trends in the research field (Nguyen et al., 2023), we have decided to focus on the multi-objective HPO approach.

In addition to fairness, this paper also addresses the issue of *energy-efficiency* of HPO. Nowadays, it has become crucial to consider the dual role of artificial intelligence (AI) and ML in the climate crisis. On the one hand, they can support more sustainable and low-emission decisions, from design to management of critical systems such as smart energy-grids, transportation, healthcare and water utilities, and they can also provide accurate climate change predictions. On the other hand, AI and ML are themselves *energivorous* and, consequently, significant CO_2 emitters, leading to the concept of Red-AI (Dhar, 2020). Prevalence of Red-AI is also quantified in Schwartz et al. (2020), reporting that the total cost of producing accurate ML models increases linearly with (i) the cost of executing the model on a single example, (ii) the size of the training dataset and (iii) the number of HPO experiments, which controls how many times the model is trained on the dataset.

Astonishing results are also reported in Strubell et al. (2019) and Hao (2019), which analysed the training process of many natural language processing (NLP) models to estimate the energy cost in kilowatts required. When these figures are converted into approximate carbon emissions it comes out that the carbon footprint of training a single large NLP model is equal to the amount of CO_2 emitted by 125 round-trip flights between New York and Beijing or, equivalently, five American average cars in their lifetimes, including their manufacturing processes.

Consequently, the research community has been focusing on the Green-AI topic and also starting to propose novel approaches to make HPO—and more generally AutoML—"greener", for instance by using smaller portions of the available databases/datasets, as proposed in the seminal work of Swersky et al. (2013) up to the most recent ones, such as in Klein et al. (2017), Candelieri et al. and (2021). Another possibility consists into early discarding unpromising hyperparameter configurations to save and re-allocate computational resources (aka, *successive halving*, first proposed in Jamieson & Talwalkar 2016). A well known example is Hyperband (Li et al., 2017). A wider overview about Green AutoML is given in Tornede et al. (2023), along with an indication of future research directions.

Bayesian optimization (BO) is a sample-efficient, sequential, model-based, global optimization method, well-suited for optimizing black-box, expensive, and multi-extremal objective functions (Frazier, 2018; Archetti & Candelieri, 2019; Garnett, 2023). Thanks to its sample-efficiency, BO is the core component of most of the current AutoML solutions, both open-source and commercial. BO has been recently extended to also deal with multiple objectives (Hernández-Lobato et al., 2016; Paria et al., 2020), as well as multiple information sources which can queried under different costs (Ghoreishi & Allaire, 2019; Belakaria et al., 2020a; Candelieri et al., 2021; Candelieri & Archetti, 2021; Khatamsaz et al., 2020). A special case is when information sources can be organized hierarchically depending on their quality of approximation (aka *fidelity*), leading to the so-called multi-fidelity optimization, originally proposed in Kennedy & O'Hagan (2000). Both in multiple information source and multi-fidelity optimization energy efficiency is achieved by suitably using the less expensive (as well as low-fidelity) sources to keep low the cumulative query cost, that is a proxy of energy consumption and, consequently, CO_2 emissions.

1.2 Contributions

The main contributions of our paper are:

1. A comparative analysis between fairness-aware ML and Fairness-aware AutoML (specifically, HPO) algorithms, on a set of four relevant benchmark (fairness) datasets.
2. A new fair and green hyperparameter optimization algorithm, namely **FanG-HPO**, based on both multi-objective and multiple information source Bayesian Optimizations to simultaneously address Fairness-aware and Green AutoML.
3. A computational assessment of FanG-HPO against other two state-of-the-art BO suites enabling both multi-objective and energy efficient HPO, specifically:
    - **autogluon-FairBO** (Schmucker et al., 2020), in which HPO is addressed as a bi-objective optimization task (i.e, simultaneously minimizing misclassification error and unfairness on 10 fold cross validation) and using *successive halving* to reduce energy consumption, and consequently CO_2 emissions.
    - **BoTorch-MOMF** (multi-objective and multi-fidelity) (Irshad et al., 2021), implementing a generic multi-objective and multi-fidelity BO framework. We have used

it to target HPO as a bi-objective task and to address energy efficiency by selectively using the entire dataset (high-fidelity source) or a portion if it (low-fidelity source) to compute the misclassification error and the unfairness metric of every hyperparameter configuration, on 10-fold cross validation. BoTorch-MOMF has been released only quite recently, within the BoTorch suite.

The rest of the paper is organized as follows: Sect. 2 provides the main background on multi-objective and multiple information source optimization, along with the definition of the unfairness metric adopted in this study. In Sect. 3 the FanG-HPO approach is detailed. Section 4 describes the experimental setting and Sect. 5 reports the results. Finally, Sect. 6 provides conclusions and perspectives.

1.3 Related works

As far as fairness in ML is concerned, recent reviews are given in Barocas et al. (2017), Pessach & Shmueli (2022) and Weerts et al. (2023). Specifically, (Weerts et al., 2023) aims at raising awareness among AutoML researchers and developers about limitations of Fairness-aware AutoML, while remarking the potential of AutoML as a tool enabling the research on fairness in ML.

With respect to the energy efficiency, a recent review is given in Tornede et al. (2023), providing important hints about quantifying sustainability of AutoML systems, along with an overview and taxonomy of currently available energy-efficient AutoML systems.

Our paper represents a contribution to the effort of providing AutoML practitioners with tools enabling the development of more *societal*—both fair and green—decision support systems based on ML.

Although BO has been extended to deal with multiple objectives (Svenson & Santner, 2016; Feliot et al., 2017; Yang et al., 2019; Iqbal et al., 2020; Daulton et al., 2020) as well as multiple fidelities and multiple information sources (Lam et al., 2015; Poloczek et al., 2017; Ghoreishi & Allaire, 2019; Candelieri & Archetti, 2021, 2021a; Ariafar et al., 2021), there is a still lack of solutions jointly addressing the two problems. On the other hand, the research interest on this specific challenge has been quickly increased, especially because its applicability to many other real-life problems than fair and green ML, as demonstrated by very recent works (Sun et al., 2022; Irshad et al., 2021).

The only two available research tools—at the authors' knowledge—are **autogluon-FairBO** (Schmucker et al., 2020), combining multi-objective and multi-fidelity optimization by building upon Hyperband (Li et al., 2017), and **BoTorch-MOMF** (Irshad et al., 2021). Although an implementation of autogluon-FairBO is still under review to be included into the autogluon[1] suite, the code is freely available[2] and it has been used to implement the first competitor of FanG-HPO. On the contrary, BoTorch-MOMF is already integrated and freely available in the BoTorch suite,[3] and is considered in this paper as the second competitor of FanG-HPO.

Finally, with respect to fairness-aware ML algorithms, relevant works are Komiyama et al. (2018), Zafar et al. (2019) and Scutari et al. (2021).

[1] https://auto.gluon.ai/stable/index.html.

[2] https://github.com/awslabs/autogluon/tree/0.3.1.

[3] https://botorch.org/tutorials/Multi_objective_multi_fidelity_BO.

2 Background

This paper addresses FairML as a multi-objective problem and, at the same time, uses multiple information sources (i.e., a small portion of the large target dataset) to improve energy-efficiency and, consequently, reduce CO_2 emissions. Here, we briefly summarize the basic background about multi-objective optimization, fairness metrics, and multiple information source optimization.

2.1 Multi-objective optimization

Multi-objective optimization (MO) concerns solving problems with more than one objective function to be optimized simultaneously, that is:

$$\min_{\mathbf{x}\in\Omega} \mathbf{f}(\mathbf{x}) \tag{1}$$

where Ω is the *search space*, typically box-bounded in $\mathbb{R}^d$, and $\mathbf{f} : \Omega \to \mathbb{R}^M$ is the vector-valued function of the multiple objectives. In MO, due to the conflicting nature of the objectives, it does not exist a unique solution $\mathbf{x}^* \in \Omega$ to the problem (1). The final aim is to identify a set of equally efficient trade-offs among the objectives. This set of efficient trade-offs can be depicted within the space spanned by the M conflicting objectives, allowing for drawing the so-called **Pareto front** (aka *frontier* or *boundary*). The associated set of solutions—into the search space Ω - is instead known as **Pareto set**. An example is shown in Appendix A.1. Formally, the Pareto set consists of only **dominant** (aka *not-dominated*) **solutions**, where a solution $\mathbf{x}$ is said to dominate another solution $\mathbf{x}'$ if their objectives, respectively $\mathbf{f}(\mathbf{x}) = \left(f_1(\mathbf{x}), ..., f_M(\mathbf{x})\right)$ and $\mathbf{f}(\mathbf{x}') = \left(f_1(\mathbf{x}'), ..., f_M(\mathbf{x}')\right)$, satisfy the following two conditions:

$$f_m(\mathbf{x}) \leq f_m(\mathbf{x}') \ \forall \ m \in \{1, ..., M\} \tag{2}$$

$$\exists \ j \in \{1, ..., M\} \ : f_j(\mathbf{x}) < f_j(\mathbf{x}') \tag{3}$$

Equation (2) means that $\mathbf{x}$ is not worse than $\mathbf{x}'$ in all the objectives, and Eq. (3) means that $\mathbf{x}$ is strictly better than $\mathbf{x}'$ in at least an objective. The Pareto dominance symbol, $\prec$, is used to synthesize (2–3): $\mathbf{f}(\mathbf{x}) \prec \mathbf{f}(\mathbf{x}')$.

If the objectives are black-box, their values can only be known point-wise by querying $\mathbf{f}(\mathbf{x})$ at specific locations. Given all the queries performed so far, the set of **non-dominated solutions** (respectively, **outcomes**) is the current approximation of the **Pareto set** (respectively, **front**). If the objectives are also expensive to evaluate, in terms of time or resources, then (1) must be solved efficiently, meaning that a good Pareto front/set approximation has to be found within a limited number of queries. Thus, sample-efficiency of BO was the driver of its successful extension to the MO setting (i.e., MOBO), mainly along three different strategies:

- *Scalarization* which maps the vector of all objectives into a scalar parametrized function whose optimizer, computed by a single objective method, can span, as the parameters vary, the whole Pareto set (Paria et al., 2020; Zhang & Golovin, 2020).

The key drawback of scalarization is that it does not consider the geometry of the Pareto front approximation.
- Maximization of some index related to the quality of the Pareto front approximation. A common choice is the *dominated hypervolume indicator* (HV), that is the volume of the region dominated by a Pareto front approximation.
- *Information theoretic based*, which aims at reducing the uncertainty/entropy about the Pareto front (Belakaria et al., 2019; Suzuki et al., 2020; Belakaria et al., 2020b), recently also considering the multi-fidelity setting (Belakaria et al., 2020a).

The BO methods considered in this paper rely on the first or the second strategy.

It is important to remark that, while in “vanilla” BO a *probabilistic surrogate model* is used to approximate the black-box objective function, almost all the MOBO approaches in literature adopt a probabilistic surrogate model for each one of the objectives, assuming independence among them. This is a reasonable assumption, because in MO the objectives should be competing and uncorrelated. Exactly as in BO, every new query contributes to better approximate the objective function—which is vector-valued in MOBO – through the update of the probabilistic surrogate model. The second key component of BO is the *acquisition function*, which deals with the well-known exploration-exploitation dilemma. All the “vanilla” BO acquisition functions can be used in the case of scalarization— because the multi-objective problem is mapped into a single-objective one—on the contrary, expected hypervolume improvement (EHVI) is an acquisition function specifically designed for vector-valued MOBO, basically extending the idea underlying the well-known Expected Improvement (EI) to the multi-objective setting.

In this paper we consider HPO of a classification model, with two different objectives to minimize: the misclassification error (MCE) and the *unfairness* metric known as differential statistical parity (DSP), which is detailed in the next section. Both the objectives are computed through stratified 10-fold cross validation (10FCV), so they are black-box, expensive, multi-extremal, and possibly noisy (depending on the specific ML algorithm to be optimized or the cross-validation procedure).

2.2 Differential statistical parity as unfairness metric

There is not a unique definition—and consequently metric—of fairness (Verma et al., 2018). Instead, different alternatives have been proposed depending on application domains and specific use cases. In this paper we consider the DSP—which has been also recently considered in Schmucker et al. (2020). We refer to the standard framework where F_L denotes the true labels for the target feature, F_{Sens} is the sensitive feature, and $\widehat{F}_L$ denotes the predicted labels. *Statistical parity* (SP) requires that positive predictions are unaffected by the value of the sensitive feature, independently of the actual label:

$$P\left(\widehat{F}_L = 1 | F_{Sens} = 0\right) = P\left(\widehat{F}_L = 1 | F_{Sens} = 1\right)$$

Finally, the absolute value of the difference between the two terms is the DSP, that is a measure of the violation of the above condition and, consequently, a measure of unfairness.

$$DSP = \left| P\left(\widehat{F}_L = 1 | F_{Sens} = 0\right) - P\left(\widehat{F}_L = 1 | F_{Sens} = 1\right) \right|$$

2.3 Multiple information source optimization

Multiple information source optimization (MISO) aims at searching for the global optimum of a black-box, expensive and multi-extremal function, namely the *ground-truth*, given the possibility to also query less expensive *information sources* which are its approximations. The final goal is to find an optimal solution for the *ground-truth* while satisfying some constraint on the query cost accumulated along the search process, by effectively and efficiently using the cheap information sources. MISO has been defined for single-objective problems: differently from multi-objective, here the subscript is used to denote a specific information source, where $f_1(\mathbf{x})$ is the ground-truth and $f_s(\mathbf{x})$, with $s \in \{2, ..., S\}$, are the cheap information sources.

The MISO problem can be formulated as:

$$\mathbf{x}^* = \underset{\mathbf{x}\in\Omega}{\arg\min} f_1(\mathbf{x}) \tag{4}$$

$$\text{subject to:} \sum_{(s,\mathbf{x})\in Z^{1:n}} c_s \leq C_{max} \tag{5}$$

where $Z^{1:n} = \left\{ \left(s^{(i)}, \mathbf{x}^{(i)} \right) \right\}_{i=1:n}$ denotes the set of source-location pairs sequentially queried, c_s is the cost for querying $f_s(\mathbf{x})$, and C_{max} is the maximum query cost that can be cumulated along the optimization process.

BO has been also successfully extended to deal with MISO problems, where each information source is individually modelled through a probabilistic surrogate model—usually a Gaussian process (GP)—fitted on the queries performed on that source. Then, all the individual models are combined into a single one, which is used to drive the choice of the next promising source-location pair to query, such as in Ghoreishi & Allaire (2019) and Candelieri & Archetti (2021).

3 FanG-HPO

The proposed fair and green HPO approach aims at solving the problems (4-5), but with the scalar objective function replaced by a vector-valued one, that is:

$$\mathbf{x}^* = \underset{\mathbf{x}\in\Omega}{\arg\min}\ \mathbf{f_1}(\mathbf{x}) \tag{6}$$

$$\text{subject to:} \sum_{(s,\mathbf{x})\in Z^{1:n}} c_s \leq C_{max} \tag{7}$$

where $\mathbf{f_1}$ is the vector-valued ground-truth, while all the other cheaper information sources are $\mathbf{f_s}$, with $s \in \{2, ..., S\}$.

3.1 Modelling objectives and information sources

In FanG-HPO, both objectives and information sources are modelled independently via GP regression (Williams & Rasmussen, 2006; Gramacy, 2020). A GP is a probabilistic regression model whose predictive mean, $\mu(\mathbf{x})$, and uncertainty, $\sigma(\mathbf{x})$, are conditioned on previous observations. A brief introduction to GP regression is provided in the Appendix A.2.

Thus, at a generic iteration, FanG-HPO learns $S \times M$ GP models, leading to the following set of predictive means and uncertainty functions:

$$\left\{ \mu_{sm}(\mathbf{x}), \sigma_{sm}(\mathbf{x}) \right\}_{\substack{s=1\,:\,S,\\ m=1\,:\,M}}$$

Then, a single GP model is fitted, for each objective, by combining the GPs which individually model that specific objective on every information source. In FanG-HPO this operation is performed by using the Augmented Gaussian Process (AGP) methodology recently proposed in Candelieri & Archetti (2021). More precisely, a set of indices identifying "*reliable*" observations from cheaper sources is computed for each pair *source - objective*, where "*reliable*" means they are not too discrepant with respect to the ground-truth:

$$\begin{aligned}\mathcal{I}_{sm} = &\left\{ i \,:\, \left|\mu_{1m}(\mathbf{x}) - \mu_{sm}\left(\mathbf{x}^{(i)}\right)\right| \le \alpha\sigma_{1m}\left(\mathbf{x}^{(i)}\right), \mathbf{x}^{(i)} \in \mathbf{X}_s \right\},\\ &\forall\, s \neq 1, \forall\, m \in \{1, \ldots, M\}\end{aligned} \tag{8}$$

where α is a technical parameter to tune reliability of the observations from cheap information sources. In Candelieri & Archetti (2021) the suggested value is $\alpha = 1$.

Then, the observations on the ground-truth are "augmented" with those identified by $\mathcal{I}_{sm}$, separately for each objective $m \in \{1, \ldots, M\}$:

$$\widehat{\mathbf{X}}_m \leftarrow \mathbf{X}_1 \cup \left\{ \mathbf{x}^{(i)} \in \mathbf{X}_s \,:\, i \in \mathcal{I}_{sm}, \forall s \neq 1 \right\} \tag{9}$$

$$\widehat{\mathbf{Y}}_m \leftarrow \mathbf{Y}_{1[m]} \cup \left\{ y^{(i)} \in \mathbf{Y}_{s[m]} \,:\, i \in \mathcal{I}_{sm}, \forall s \neq 1 \right\} \tag{10}$$

where $\mathbf{X}_s$ are the locations queried on source s and $\widehat{\mathbf{Y}}_{[m]}$ are the values observed for the objective m and associated to the set $\widehat{\mathbf{X}}_m$ (i.e., the symbol $[m]$ is the operator selecting only the column m of the $n_s \times M$ matrix $\mathbf{Y}_{sm}$, with n_s the number of queries performed on the source s).

Finally, FanG-HPO fits M independent AGPs, with predictive means and uncertainty respectively denoted with $\widehat{\mu}_m(\mathbf{x})$ and $\widehat{\sigma}_m(\mathbf{x})$, and both conditioned to $\left\{ \widehat{\mathbf{X}}_m, \widehat{\mathbf{Y}}_m \right\}$.

3.2 Deriving the next source-location to query

The next *source-location* pair to query, namely $(s', \mathbf{x}')$, is derived by solving a multi-objective problem whose objectives are approximated by the AGPs obtained as previously described. Having M independent AGPs is in line with recent results in literature: as reported in Zhan et al. (2017) considering a separate GP modelling each objective independently makes easy the implementation of multi-objective optimization approaches, while using dependent GP models—such as multi output GPs—does not provide any relevant benefit against independent GPs (Svenson & Santner, 2016).

More precisely, $(s', \mathbf{x}')$ is obtained according to the following two-steps procedure:

1. **Selecting $\mathbf{x}'$**. First, the location $\mathbf{x}'$ is selected depending on the well-known expected hypervolume improvement (EHVI). Hypervolume Improvement (HVI) is defined as the relative increase in the hypervolume indicator, when an outcome $\mathbf{y}$, associated to a solution $\mathbf{x}$, is added to the current Pareto front approximation. In BO, the HVI is a random variable because $\mathbf{y}$ is a (set of) random variable itself, and this leads to the EHVI.

$$\mathbf{x}' = \arg\max_{\mathbf{x}\in\Omega} \text{EHVI}(\mathbf{x}, \mathcal{P}, \mathbf{r}) \tag{11}$$

where $\mathcal{P}$ is the current approximated Pareto front and $\mathbf{r}$ is the **reference point**. In this paper $\mathbf{r}$ is the worst point, with both MCE and DSP equal to 1. Although a closed formula for the EHVI exists (Feliot et al., 2017), it is expensive to calculate. In FanG-HPO the fast calculation proposed in (Zhao et al., 2018) is used, that is an extension, to the EHVI computation, of the Walking Fish Group (WFG) technique (While et al., 2011), one of the fastest algorithms for calculating the hypervolume of a Pareto front approximation.

2. **Selecting s'**. Then, the information source s' is selected according to both its query cost and its discrepancy with respect to the ground-truth at $\mathbf{x}'$, with respect to all the objectives, that is:

$$s' = \arg\min_{s\in\{1\ldots,S\}} c_s \cdot \sum_{m=1}^{M} \left|\mu_{1m}(\mathbf{x}') - \mu_{sm}(\mathbf{x}')\right| \tag{12}$$

Contrary to other recent approaches which propose to query the ground-truth on a regular basis, such as in Khatamsaz et al. (2020), at each iteration FanG-HPO adaptively chooses among all the sources, including the ground-truth. However, just to ensure a sufficient quality of the approximation provided by the AGPs, before solving (12) FanG-HPO checks if the number of augmenting observations coming from cheap sources is larger than those from the ground-truth: in that case $s' = 1$ is selected, instead of solving (12).

4 Experimental setting

4.1 Datasets and Machine Learning algorithms

To select a suitable set of benchmark datasets on fairness, we have based our choice on the paper (Le Quy et al., 2022) which provides a detailed overview and analysis of real-world tabular datasets frequently used in fairML. Specifically, Bayesian networks were used to model and analyse the relationship between protected attributes and target class, for each considered dataset. We limited our selection to four binary classification datasets, resulting difficult in terms of achievable accuracy and statistical parity, according to the results reported in table 15 of Le Quy et al. (2022). It is important to anticipate that our results on the four selected datasets are homogeneous, suggesting that was useless to extend the analysis to other—easier—datasets. [i.e., a more extended experimental campaign is out of the scope of this paper. For a wider set of experiments on effectiveness of Fainess-aware ML one can refer to very recent studies, such as Nguyen et al. (2023)].

More specifically, the four selected datasets are known with the names: ADULT, COMPAS, GERMAN CREDIT, and LAW SCHOOL ADMISSIONS. They are taken from the R package `fairml`,[4] which also provides implementations of a set of fairness-aware algorithms. The same datasets are also available on the well-known UCI Repository,[5] but the versions available in the R package could be slightly different due to some basic pre-processing operations. It is important to clarify that these operations are related to common data pre-processing (e.g., identifiers removal) and not to any fairness-oriented pre-processing technique. As follows we report, for each dataset, some details relevant for our experiments. A brief description of the four datasets is reported in the following.

- **ADULT**—the aim of the associated classification task is to predict whether personal income exceeds 50K$ per year, using the U.S. 1994 Census data. The dataset consists of 30,162 instances and 14 features (among them, two are sensitive: "gender" and "race").
- **COMPAS**—data refers to criminal offenders screened in Florida (US) during 2013–2014. The aim of the associated classification task is to predict the recidivism of crime in two years. The dataset consists of 5855 instances and 16 features (among them, two are sensitive: "gender" and "race").
- **GERMAN CREDIT**—data refers to credit scoring and the aim of the associated classification task is to predict defaults on consumer loans in the German market. The dataset consists of 1000 instances and 21 features (among them, one is sensitive: "gender").
- **LAW SCHOOL ADMISSIONS**—data refers to a survey among students attending law school in the U.S. in 1991. Although the original task associated to this dataset is a regression task (i.e., predicting the Undergraduate Grade Point Average), we have decided to consider a different target feature, specifically the one assessing whether the student has passed the bar exam on the first try. Thus, the classification task we consider in our experiments is to predict this outcome. The dataset consists of 20,800 instances and 11 features (among them, two are sensitive: "gender" and "race").

Before performing HPO, all the datasets have been (further) pre-processed by applying one-hot-encoding on all the nominal features, increasing the final number of features (including the target feature) to: 52 for ADULT, 20 for COMPAS, 47 for GERMAN CREDIT, 51 for LAW SCHOOL ADMISSIONS. Pre-processing has also increased the number of sensitive features for some dataset. As better detailed in Sect. "Availability of data and material", all the pre-processed datasets, as they have been used in this study, are available for replicability, along with the code.

With respect to the HPO of the ML algorithms, we have selected four completely different ML algorithms: multi-layer perceptron (MLP), random forest (RF), eXtreme Gradient Boosting (XGB), and support vector machine (SVM) with an RBF kernel. The number of hyperparameters to be optimized is respectively: 10 for MLP, 2 for RF, 7 for XGB, and 2 for SVM. The details about their associated search spaces are reported in the following four Tables 1, 2, 3, and 4. For MLP and XGB we have used the same search spaces defined in Schmucker et al. (2020), while RF and SVM are defined by us, because they were not considered in the quoted study.

[4] https://www.rdocumentation.org/packages/fairml/versions/0.6.1/topics/fairml-package.

[5] https://archive.ics.uci.edu/ml/index.php.

Table 1 sklearn MLP's search space

Hyerparameter	Type	Domain	Scaling
n_Layers	Integer	{1,2,3,4}	Linear
Layer_1	Integer	{2,...,32}	Linear
Layer_2	Integer	{2,...,32}	Linear
Layer_3	Integer	{2,...,32}	Linear
Layer_4	Integer	{2,...,32}	Linear
Alpha	Real	$[10^{-6},10^{-1}]$	Log_{10}
Learning_rate_init	Real	$[10^{-6},10^{-1}]$	Log_{10}
Beta_1	Real	[0.001,0.99]	Log_{10}
Beta_2	Real	[0.001,0.99]	Log_{10}
Tol	Real	$[10^{-5},10^{-2}]$	Log_{10}

Table 2 sklearn RF's search space. The range of the hyperparameter MAX_FEATURES depends on the dataset: $|F|$ denotes the number of features, excluded the target one

Hyerparameter	Type	Domain	Scaling
n_Estimators	Integer	{100,...,1000}	Linear
Max_features	Integer	$\{2, ..., \|F\|\}$	Linear

Table 3 sklearn XGBoost's search space

Hyerparameter	Type	Domain	Scaling
n_Estimators	Integer	{1,...,256}	Linear
Learning_rate	Real	[0.01,1.0]	Log_{10}
Gamma	Real	[0.0,0.1]	Linear
Reg_alpha	Real	$[10^{-3},10^{3}]$	Log_{10}
Reg_alpha	Real	$[10^{-3},10^{3}]$	Log_{10}
Subsample	Real	[0.01,1.0]	Linear
Max_depth	Integer	{1,2,...,16}	Linear

Table 4 sklearn SVM's search space

Hyerparameter	Type	Domain	Scaling
C	Real	[0.0001,10,000]	$\log_{10}$
Gamma	Real	[0.0001,10,000]	$\log_{10}$

4.2 Compared methods

We have compared our approach, **FanG-HPO**, against two state-of-the-art BO suites enabling both multi-objective and energy efficient optimization, specifically **autogluon-FairBO** and **BoTorch-MOMF**.

Bi-objective optimization (i.e., simultaneous minimization of 10FCV MCE and 10FCV DSP), is addressed via scalarization in autogluon-FairBO and via EHVI maximization in both FanG-HPO and BoTorch-MOMF.

Another important difference regards the strategies adopted to deal with energy-efficiency. According to the taxonomy in Tornede et al. (2023):

autogluon-FairBO belongs to the family of *"early discarding of unpromising candidates"* methods and is based on Hyperband (Li et al., 2017).

BoTorch-MOMF and **FanG-HPO** belong to the family of the *"multi-fidelity performance measurements"* methods, but with a significant methodological difference. BoTorch-MOMF uses a multi-output Gaussian process (GP) to model three objectives: not only 10FCV MCE and 10FCV DSP, but also the query costs associated to the sources. Moreover, the fidelity is also included as an additional decision variable (i.e., it is treated just like a hyperparameter of the ML to be optimized). On the contrary, Fang-HPO uses an independent Augmented Gaussian Process (AGP) for each objective, fitted by merging observations on the different information sources. This difference is even more important in terms of acquisition function: although both BoTorch-MOMF and FanG-HPO use EHVI, the first penalizes the associated value depending on the cost of the source (the higher the fidelity the higher the cost) while the second adopts the two steps mechanism explained in Sect. 3.2.

Another important comparison performed in our study is between the three HPO methods, all together, and two well known FairML algorithms, both available in the R package `fairml`. The two algorithms are named `zlrm` and `fgrrm` (Scutari et al., 2021). This comparison is important to evaluate, in terms of fairness, the effectiveness of Fairness-aware AutoML with respect to fairness-aware (aka fairness-by-design) ML algorithms.

4.3 Performance metrics

As previously mentioned, in this paper HPO is aimed at simultaneously minimize 10FCV MCE and 10FCV DSP), separately for every pair *ML algorithm - dataset*. It is important to remark that, given a specific dataset, DSP is computed for every sensitive feature, according to what previously reported in Sect. 2.2. To obtain a single value—instead of a vector—we have decided to consider DSP= $\max_{i=1:n_{Sens}} \{DSP_i\}$, where n_{Sens} is the number of sensitive features and DSP_i is DSP value associated to the i-th sensitive feature. Basically, we minimize the worst DSP over all the sensitive features.

Being respectively a multi-fidelity and a multiple information source optimization approach, it is quite simple to define sources and their query costs for BoTorch-MOMF and FanG-HPO. Specifically, 10FCV MCE and 10FCV DSP are computed on the high fidelity / expensive source (i.e., the ground-truth) when the associated hyperparameter configuration is evaluated on the entire dataset, otherwise they are computed on the low fidelity / cheap source if the hyperparameter configuration is evaluated on a stratified sample (i.e., 50%) of the original dataset. Just for the sake of simplicity, we can assume that the *nominal* query cost of the two information sources are, respectively, $c_1 = 1$ and $c_2 = 0.5$. It is important to remark that nominal query cost is not a direct proxy of energy consumption and CO_2 emissions, but it drives energy-efficient choices in the two approaches.

Being based on Hyperband, autogluon-FairBO adopts successive halving on the validation folds. In brief, if a configuration of the hyperparameters is not promising according to the results iteratively collected on the folds, it is discarded and the 10 fold cross validation procedure is early stopped. As a consequence, we cannot define in advance the cost of evaluating a hyperparameter configuration in autogluon-FairBO. Thus, we cannot define information sources and their query costs a-priori: if the 10FCV procedure terminates with success, then we consider that the query has been performed on the ground-truth, and we use $c_1 = 1$, otherwise we consider $c_2 = n_f/10$, with n_f the number of folds analysed before halving.

As far as energy consumption and CO_2 emissions are concerned, we decided to consider *runtime* (i.e., query time) as a suitable proxy, as better detailed and motivated in Sec. 5.3.

Finally, it is important to remark that the query cost is not applicable in the case of the fairness-aware algorithms (i.e., HPO is not performed on them because they have not hyperparameters to be optimized).

4.4 Experimental protocol

Autogluon-FairBO does not allow to specify a maximum cumulative query cost as a termination criterion. The only option is to provide a maximum number of queries, that is a maximum number of hyperparameter configurations to evaluate. To have a fair comparison against the other two approaches, namely BoTorch-MOMF and FanG-HPO, we have decided to implement the following experimental protocol (for all the ML algorithms and datasets):

1. Executing **autogluon-FairBO** with a limit of 200 queries;
2. Computing the cost of each query as $n_f/10$, with n_f the number of folds considered before halving (if any);
3. Computing the resulting overall query cost accumulated by autogluon-FairBO on that run (call it *budget*);
4. Selecting the first $2(d + 1)$ hyperparameter configurations queried by autogluon-FairBO and divide them, randomly, in two sets of size $d + 1$ each, with d the number of hyperparameters to optimize.
5. Running **BoTorch-MOMF** by initializing the multi-output GPs with the two sets mentioned above (i.e., it is important to recall that fidelity is treated as an additional hyperparameter to be optimized and the associated query cost is part of the acquisition function). The previously computed *budget* is set as a threshold for the cumulative query cost (i.e., termination criterion).
6. Running **FanG-HPO** by initializing the two AGPs with the two sets of observations mentioned above. The same termination criterion of BoTorch-MOMF is used also for FanG-HPO.

To mitigate the randomness of the initialization in autogluon-FairBO, ten independent runs have been performed for each pair *ML algorithm - dataset*. The experimental protocol has been applied for each one of the independent runs. Analogously, five independent runs have been performed for `zlrm` and `fgrrm`, separately. A constraint on the unfairness must be provided for these two algorithms: we set this value to 0.1 (for each sensitive feature).

All the experiments have been performed on a Microsoft Azure virtual machine, Standard D16ds v5 (16 vcpus, 64 GiB memory) Ubuntu 18.04 LTS.

5 Results

In this section we summarize the most relevant results of our study. Every result is first stated and then commented, to make more easy-to-read this section. Moreover, results are organized into three subsections:

- the first is related to a comparison between FairML and Fairness-aware AutoML algorithms,
- the second is related to the cost-effectiveness of the three BO-based approach considered in the paper,
- and finally the third subsection illustrates the ecological profiles of the three BO-based methods.

5.1 Fairness related results

As a first step we have performed a comparison, in terms of Pareto optimality, between Fairness-aware ML and Fairness-aware AutoML algorithms. To achieve this, for every pair *dataset - ML algorithm*, we have selected the dominant hyperparameter configurations ***among all those generated by the three BO-based approaches over all the 10 independent runs***. It is important to remark that only hyperparameter configurations evaluated on the entire datasets are considered in this operation. We call the resulting approximated Pareto front ***"super Pareto front"***. Figure 1 depicts the super Pareto fronts along with the MCE–DSP trade-offs provided by the Fairness-aware ML algorithms. In the following, the main results derived from it.

Result 1. *Fairness-aware AutoML (Pareto) dominates Fairness-aware ML algorithms.*

With respect to the four datasets, it always exists at least one super Pareto front dominating the MCE–DSP trade-offs of the two Fairness-aware ML algorithms. This result is also in line with what recently reported in (Cruz & Hardt, 2023) about post-processing of Fairness-aware ML algorithms.

Result 2. *Bi-objective HPO of RF leads to super Pareto fronts smaller than those of HPO of other ML algorithms, in terms of both HV and number of Pareto optimal hyperparameter configurations.*

More specifically, the super Pareto front associated to bi-objective HPO of RF is quite limited in terms of 10FCV DSP, for all the four datasets considered. Thus, although accurate, the final RF models obtained via HPO are not so fair.

Result 3. *Overall, the bi-objective HPO of XGB has led to the best results.*

The super Pareto front associated to XGB is always larger than the others in terms of both HV and number of Pareto optimal hyperparameter configurations.

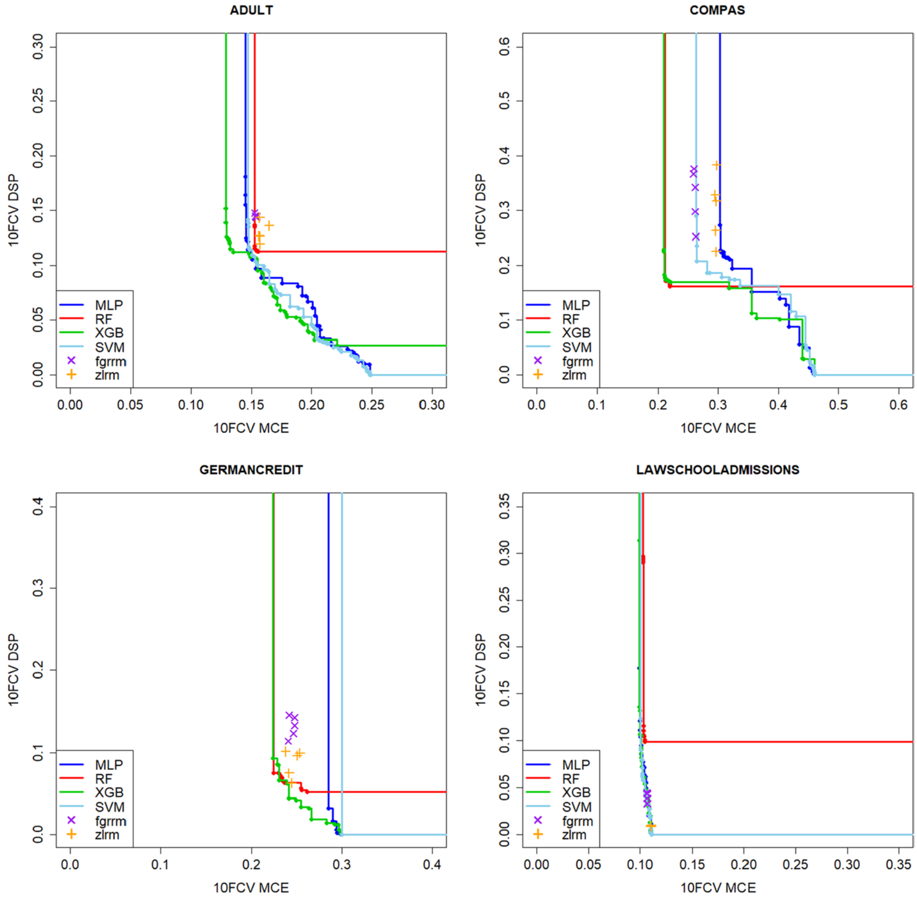

Fig. 1 Comparison, on each dataset, between MCE–DSP trade-offs provided by Fairness-aware ML algorithms and *super Pareto fronts* obtained through HPO of four ML learning algorithms. Super Pareto fronts refers to all the dominant hyperparameter configurations identified by the three BO-based approaches all together (i.e., autogluon-FairBO, BoTorch-MOMF, and FanG-HPO), over 10 independent runs

5.2 Cost-effectiveness of fairness-aware HPO methods

When comparing AutoML systems, it is not correct to just consider the final performances at the end of their run. Instead, one has to look at the performance curves, usually given by the best observed value of a metric with respect to the number of queries performed or the cumulative query cost.

In this section we report the curves of the best HV (of the approximated Pareto front) with respect to the cumulative query cost: one curve for every BO-based approach and run, and separately for each pair *ML algorithm - dataset*. These curves allow us to compare the three BO-based approaches in terms of their cost-effectiveness.

It is important to remark that, in this case, we do not deal with the *super Pareto front*: instead, we consider the HV of approximated Pareto front consisting of the dominant hyperparameter configurations – evaluated on the entire dataset, only (i.e., queries on the groud-truth)—at increasing values of the cumulative query cost. It is also important to

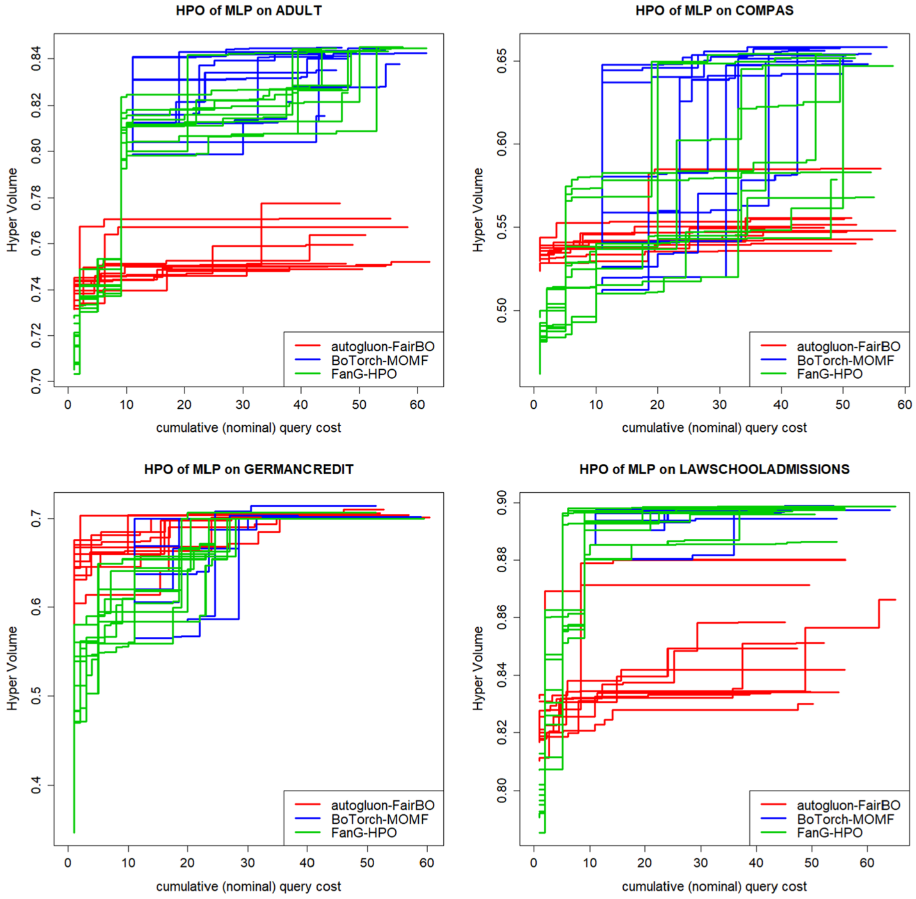

Fig. 2 Cost-effectiveness of the three BO-based approaches for bi-objective HPO of a MLP classifier over 10 independent runs, separately for the four datasets

recall that, in this subsection, we refer to the *nominal* query costs, associated to the two different sources and defined in Sect. 4.3.

Result 4. *HPO of MLP on the four datasets: successive halving (i.e., autogluon-FairBO) is less cost-effective than multi-fidelity and multiple information source BO (i.e., BoTorch-MOMF and FanG-HPO).*

Although autogluon-FairBO starts from higher initial values of HV (due to the different type of initialization of the approaches, as described in the experimental protocol), both BoTorch-MOMF and FanG-HPO are able to overcome it within a small cumulative nominal cost (Fig. 2).

Result 5. *The "pathological" behaviour of RF (previously reported in Result 2) affects all the three BO-based methods.*

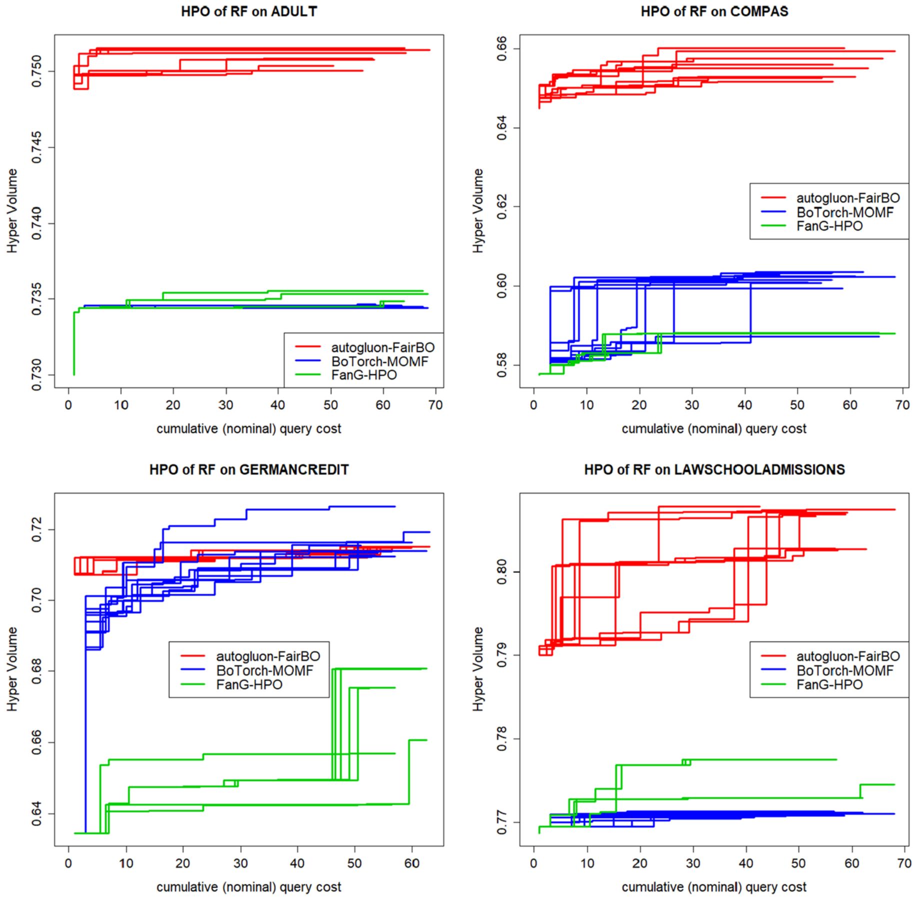

Fig. 3 Cost-effectiveness of the three BO-based approaches for bi-objective HPO of a RF classifier over 10 independent runs, separately for the four datasets

As depicted in Fig. 3, although BoTorch-MOMF and FanG-HPO are both able to improve with respect to their initial HV values, they could not close the gap with autogluon-FairBO (except for BoTorch-MOMF on the GERMANCREDIT dataset). Moreover, on average, the improvement with respect to the initial HV value is not so relevant, for all the approaches. The main motivation underlying these two issues is the degenerate shape of the associated approximated Pareto front for RF (as shown in the previous subsection).

Result 6. *HPO of XGB on the four datasets: successive halving (i.e., autogluon-FairBO) is less cost-effective than multi-fidelity and multiple information source BO (i.e., BoTorch-MOMF and FanG-HPO).*

Basically, this result is aligned with what previously obtained on MLP: the multi-fidelity and the multiple information source BO approaches (i.e., BoTorch-MOMF and FanG-HPO) are more cost-effective than the successive halving mechanism of autogluon-FairBO (Fig. 4).

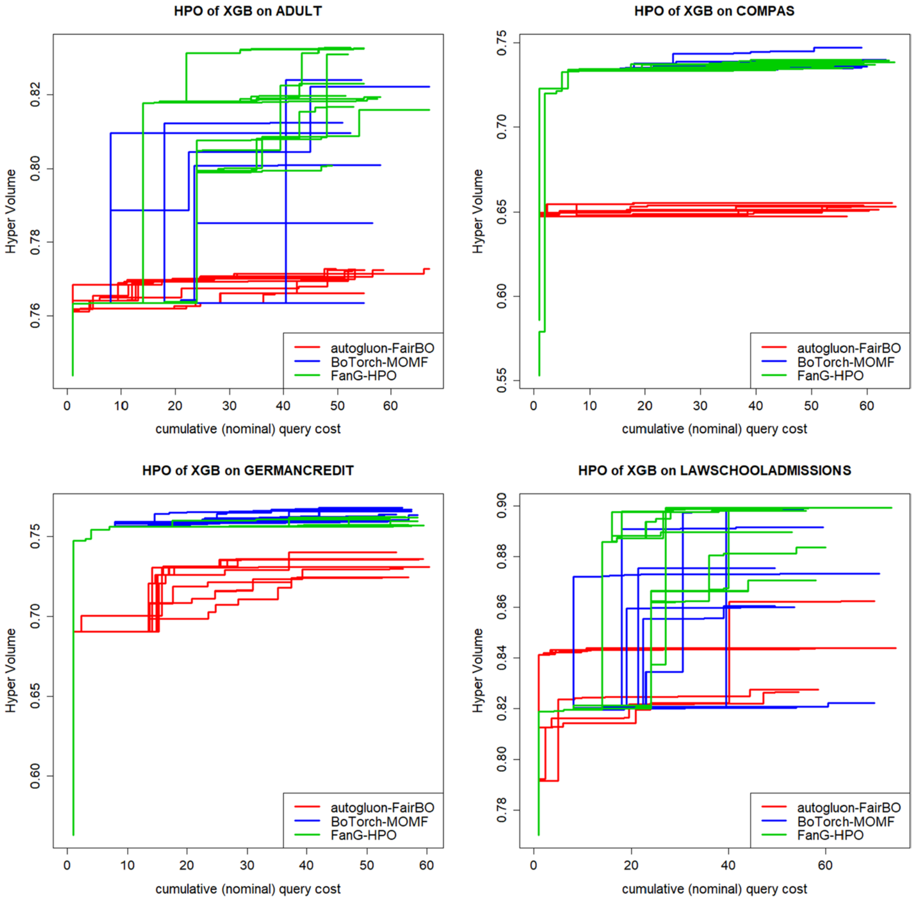

Fig. 4 Cost-effectiveness of the three BO-based approaches for bi-objective HPO of a XGB classifier over 10 independent runs, separately for the four datasets

Result 7. *HPO of SVM on the four datasets: FanG-HPO is the most cost-effective approach.*

Although there is not a clear winner between autogluon-FairBO and BoTorch-MOMF, FanG-HPO is always among the most cost-effective approach, on all the four datasets considered (Fig. 5).

5.3 Ecological performance profiles

Computing the cost-effectiveness in terms of cumulative nominal query cost does not provide a direct quantification of the energy consumption and, consequently, the carbon footprint of the three BO-based approaches.

As reported in Tornede et al.(2023), although *runtime* is a poor measure of energy efficiency—basically because it is hardware-dependent—it can be straightforward

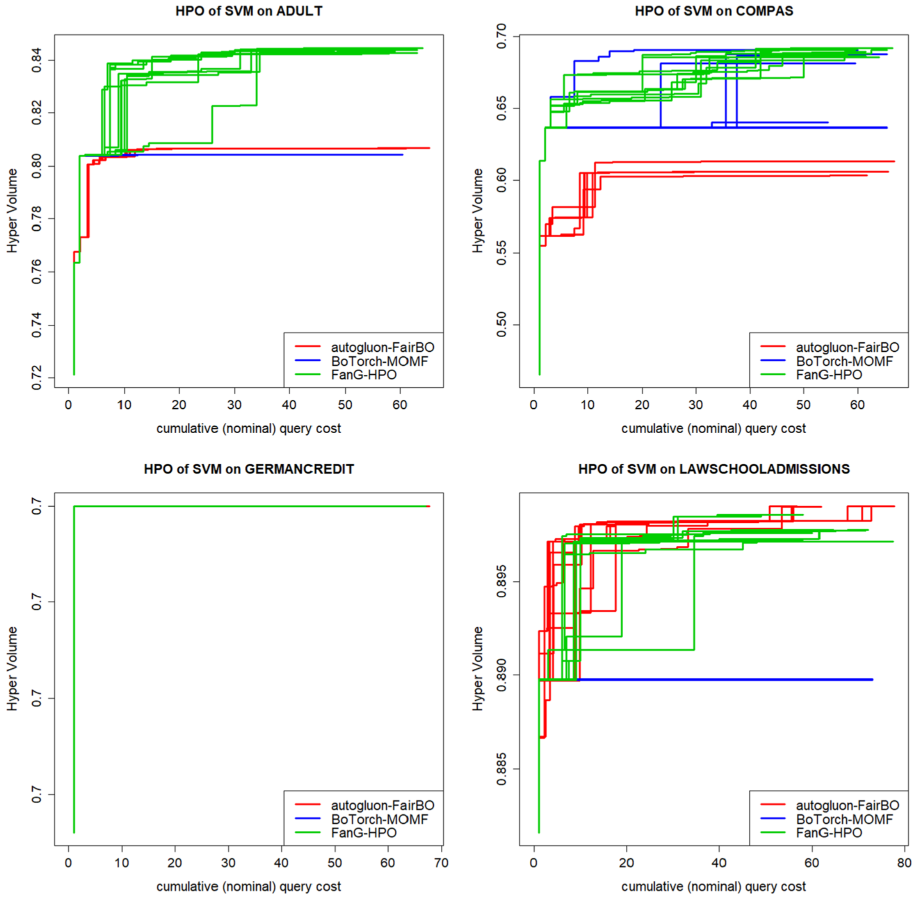

Fig. 5 Cost-effectiveness of the three BO-based approaches for bi-objective HPO of a SVM classifier over 10 independent runs, separately for the four datasets

measured on most hardware, contrary to other measures. Moreover, it is a quite practical proxy of the environmental impact whenever any additional information, such as the energy consumption of used hardware (per time unit) and the composition of the energy mix, is not available. Runtime also depends on other factors, such as the degree of parallelism and heterogeneity of the execution environment, but when the same hardware is used for running a set of competing approaches—just like in our case—it can be considered a good proxy of the CO $_2$ footprint of each competitor, at the location and time it was executed.

According to these considerations, we have decided to redraw the previous cost-effectiveness curves in terms of *cumulative query time* (i.e., runtime), instead of cumulative nominal query cost. It is also important to remark that, to guarantee a fair comparison, we have only considered the runtime required to evaluate hyperparameter configurations (i.e., query time), ignoring the computational time of the approaches themselves (which is in any case negligible with respect to the query time). When runtime is considered instead of the cumulative nominal query cost, the curves are named *ecological performance profiles*

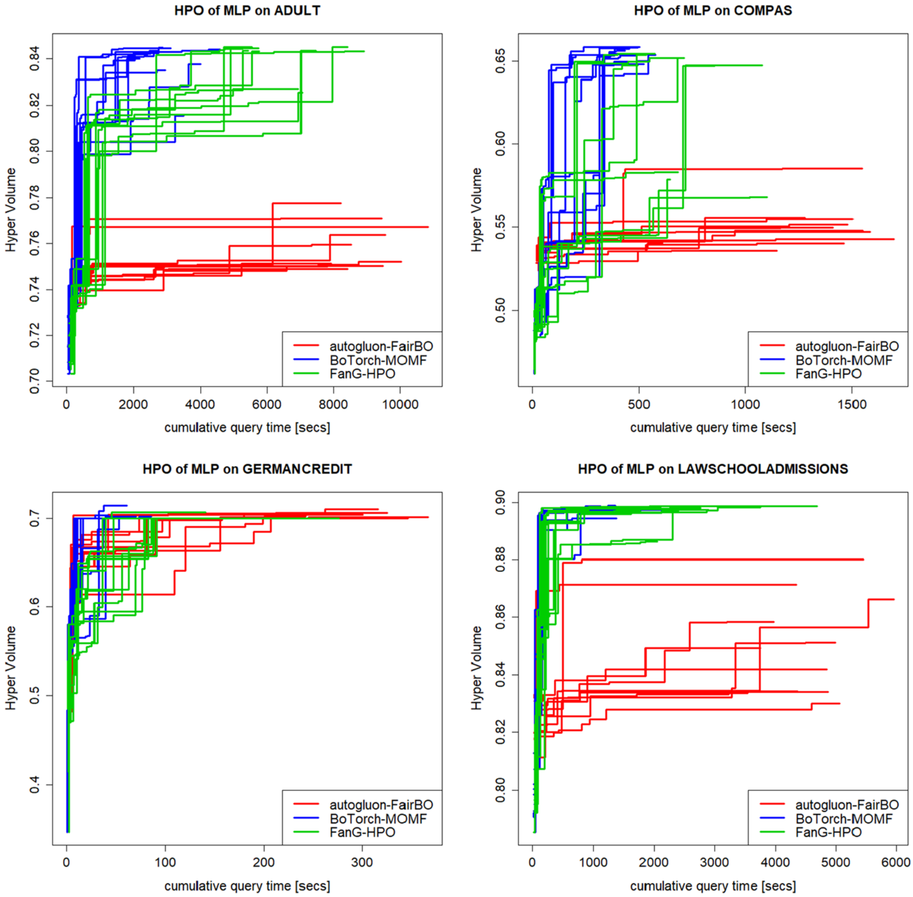

Fig. 6 HPO of MLP on the four datasets: ecological performance profiles of the three BO-based approaches

(Tornede et al., 2023). They are reported in the following Figs. 6, 7, 8 and 9, one for each ML algorithm.

Result 8. *Overall, successive halving (i.e., autogluon-FairBO) resulted less "ecological" than multi-fidelity and multiple information source bi-objective HPO (i.e., BoTorch-MOMF and FanG-HPO, respectively).*

As it can be noticed in Figs. 6, 7, 8 and 9, the cumulative query time of BoTorch-MOMF and FanG-HPO are significantly lower than autogluon-FairBO's one. Only in the case of HPO RF – whose pathological behavior has been already commented – the cumulative query time of BoTorch-MOMF is larger than the autogluon-FairBO's one.

Result 9. *XGB is the ML algorithm on which performing bi-objective HPO yields the best results; FanG-HPO is the most effective and green method for HPO.*

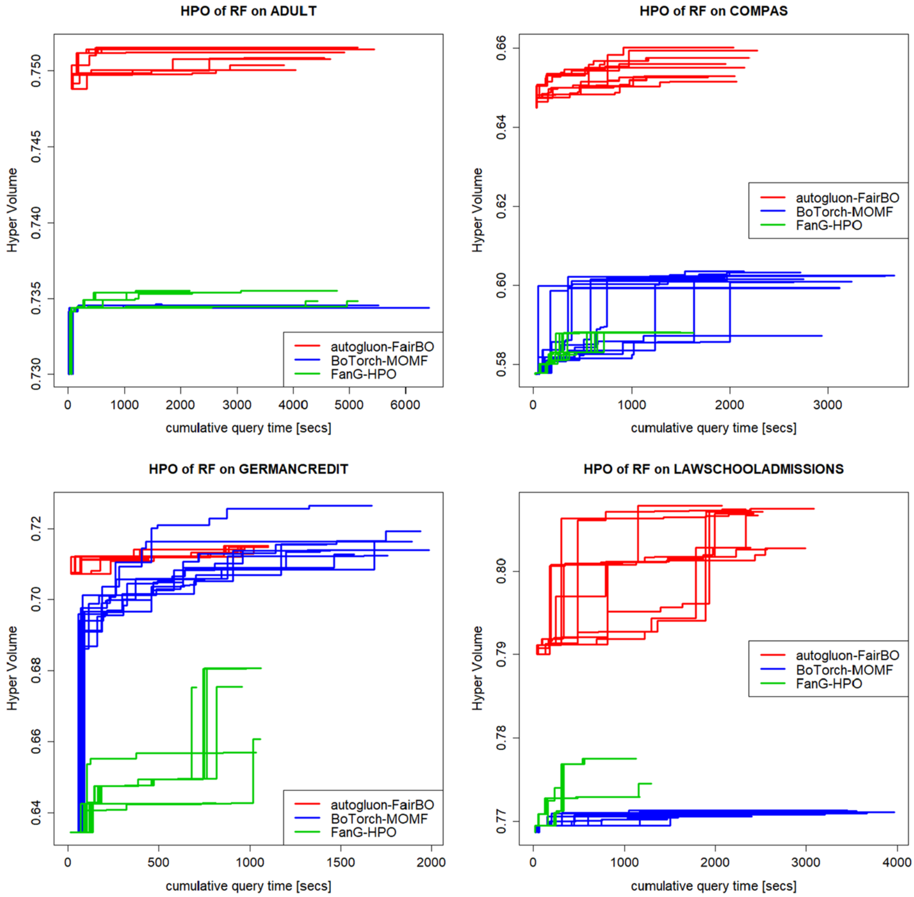

Fig. 7 HPO of RF on the four datasets: ecological performance profiles of the three BO-based approaches

From a ML developer and practitioner, this is one of the most valuable result from our research. As widely proven, XGB is usually among the most performing ML algorithms, on a number of datasets. Moreover, performing bi-objective HPO on it (i.e., accuracy and fairness) leads to the the richest set of Pareto optimal models. Finally, performing HPO of XGB through FanG-HPO will result into the best ecological performance profile.

5.4 Additional results and considerations

Finally, we have investigated how much frequently every BO-based approach queries the high-fidelity / expensive information source to deal with energy efficiency. As follows, the most relevant results.

Result 10. *On average, FanG-HPO is the approach that more frequently uses the expensive source.*

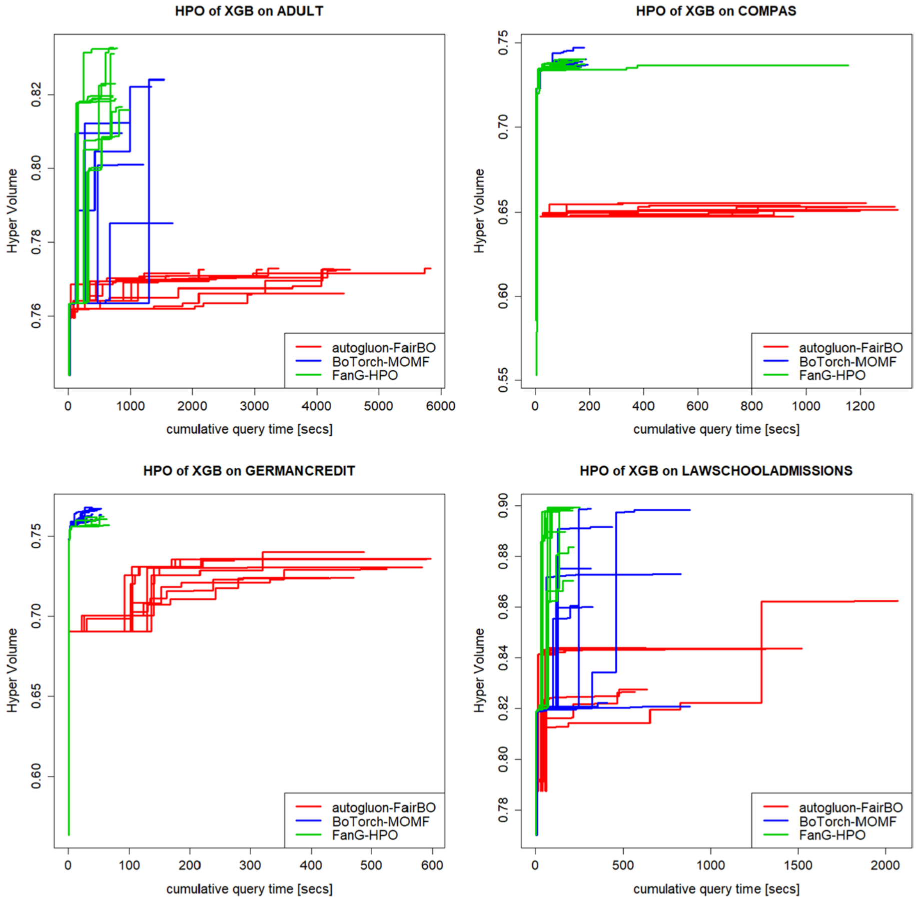

Fig. 8 HPO of XGB on the four datasets: ecological performance profiles of the three BO-based approaches

From Table 5 it is easy to notice that FanG-HPO queries the expensive information source (i.e., the ground truth) significantly more frequently than the other approaches. This is quite obvious with respect to autogluon-FairBO: indeed, halving occurs quite frequently and, even if it happens at 9 out of 10 folds, the associated query is not on counted as “on the ground truth”. Then, except for HPO on SVM, BoTorch-MOMF shows a quite constant behaviour, that is querying the high-fidelity source around 40% of the times.

It is important to remark that, although FanG-HPO performs more queries on the expensive source, its cumulative query time (i.e., runtime) is not so higher than the BoTorch-MOMF’s one; actually, it is even smaller in many cases, given the same termination criterion. This means that FanG-HPO is “clever” in using the different sources, as clearly stated in the next—and last—result of this paper.

Result 11. *On average, FanG-HPO is the approach providing the largest sets of Pareto optimal models.*

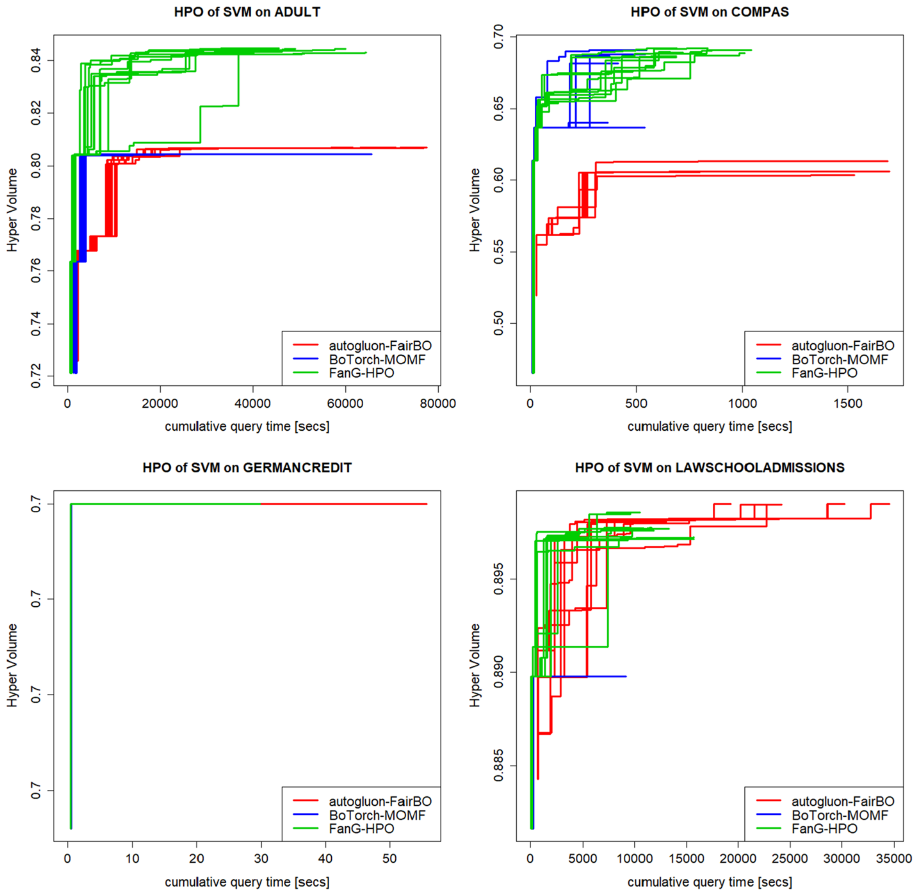

Fig. 9 HPO of SVM on the four datasets: ecological performance profiles of the three BO-based approaches

Table 6 reports the percentages of Pareto optimal hyperparameter configurations with respect to the total number of configurations evaluated—averaged on 10 independent runs.

6 Conclusions and perspectives

Although it is not the main goal, this paper empirically proves the highest effectiveness of Fairness-aware AutoML approaches (specifically bi-objective HPO) with respect to fairness-aware ML algorithms (a more recent and extended experimental campaign is offered by Nguyen et al. (2023)).

From a ML developer/practitioner's point of view, one of the most practical result is that XGB is the best algorithm to address FairML via HPO. On the four datasets considered, the Pareto optimal hyperpamater configurations obtained for XGB dominate, almost completely, those of the other ML algorithms as well as those from two well-known fairness-aware ML algorithms.

Table 5 Percentage of queries on the ground-truth (i.e., hyperparameter configurations evaluated on the entire dataset): mean and standard deviation on 10 independent runs

ML algorithm	Dataset	autogluon-FairBO	BoTorch-MOMF	FanG-HPO
MLP	ADULT	3.64% (0.90%)	40.47% (3.86%)	60.24% (6.68%)
	COMPAS	3.97% (0.36%)	40.37% (4.05%)	55.33% (8.72%)
	GERMAN CREDIT	3.74% (0.37%)	41.20% (4.04%)	61.36% (7.28%)
	LAW SCHOOL ADMISSIONS	4.36% (0.48%)	41.52% (3.82%)	58.12% (5.34%)
RF	ADULT	4.91% (0.97%)	41.30% (4.28%)	91.25% (1.88%)
	COMPAS	4.55% (0.66%)	45.67% (6.58%)	92.69% (0.81%)
	GERMAN CREDIT	4.09% (0.33%)	40.43% (4.45%)	84.18% (4.50%)
	LAW SCHOOL ADMISSIONS	4.41% (0.94%)	43.12% (4.85%)	90.17% (1.76%)
XGB	ADULT	3.73% (0.82%)	44.88% (6.30%)	84.09% (3.71%)
	COMPAS	4.22% (0.66%)	41.55% (5.56%)	73.28% (10.62%)
	GERMAN CREDIT	3.84% (0.24%)	38.28% (6.61%)	53.73% (9.78%)
	LAW SCHOOL ADMISSIONS	4.85% (0.74%)	41.60% (5.25%)	66.38% (14.78%)
SVM	ADULT	5.67% (0.13%)	8.05% (6.77%)	68.06% (13.65%)
	COMPAS	4.78% (0.35%)	16.23% (7.41%)	53.50% (6.38%)
	GERMAN CREDIT	4.67% (1.30%)	8.33% (3.52%)	47.87% (1.69%)
	LAW SCHOOL ADMISSIONS	5.28% (1.03%)	11.12% (7.45%)	52.14% (7.54%)

Table 6 Percentage of Pareto optimal hyperparameter configurations: mean and standard deviation on 10 independent runs

ML algorithm	Dataset	autogluon-FairBO	BoTorch-MOMF	FanG-HPO
MLP	ADULT	0.81% (0.26%)	12.02% (2.97%)	**16.02%** (3.66%)
	COMPAS	0.90% (0.24%)	10.37% (2.04%)	**12.96%** (3.30%)
	GERMAN CREDIT	0.36% (0.15%)	**1.97%** (0.86%)	1.76% (0.56%)
	LAW SCHOOL ADMISSIONS	1.21% (0.35%)	10.63% (2.15%)	**13.90%** (1.92%)
RF	ADULT	1.00% (0.24%)	2.96% (0.75%)	**4.16%** (0.95%)
	COMPAS	0.88% (0.37%)	**7.39%** (2.27%)	5.72% (1.40%)
	GERMAN CREDIT	0.60% (0.09%)	4.62% (1.69%)	**7.60%** (2.35%)
	LAW SCHOOL ADMISSIONS	0.75% (0.18%)	**7.41%** (1.41%)	3.82% (2.72%)
XGB	ADULT	1.14% (0.23%)	8.06% (1.09%)	**19.07%** (3.34%)
	COMPAS	0.55% (0.18%)	7.77% (1.61%)	**11.22%** (1.72%)
	GERMAN CREDIT	0.76% (0.11%)	8.99% (2.08%)	**9.30%** (2.21%)
	LAW SCHOOL ADMISSIONS	1.01% (0.33%)	7.61% (1.51%)	**14.98%** (2.58%)
SVM	ADULT	2.64% (0.16%)	4.46% (1.46%)	**19.67%** (2.75%)
	COMPAS	0.81% (0.07%)	2.45% (0.82%)	**7.17%** (2.75%)
	GERMAN CREDIT	0.18% (0.02%)	0.98% (0.13%)	**1.35%** (0.17%)
	LAW SCHOOL ADMISSIONS	1.56% (0.18%)	0.90% (0.17%)	**7.74%** (2.36%)

The highest (mean) values are in bold

As far as Fairness-aware AutoML (i.e., bi-objective HPO) is concerned, the two approaches belonging to the so-called family of *multi-fidelity performance* *measurements* methods Tornede et al. (2023), that is BoTorch-MOMF and FanG-HPO (proposed in this paper), offer better *ecological performance profiles* than successive halving (i.e., autogluon-FairBO). Specifically, defining the information sources—as done in BoTorch-MOMF and FanG-HPO—instead of using successive halving, improves the ability to learn from different information sources and to take advantage of that over the sequential optimization process. This consideration is also in line with other results recently reported in Candelieri et al. (2021), where single-objective HPO, based on multiple information source optimization, resulted more effective and efficient than FABOLAS Klein et al. (2017) which instead addresses multi-fidelity by including the size of the dataset as a further hyperparameter to optimize.

Although they belong to the same family, the underlying methodological background of BoTorch-MOMF and FanG-HPO is significantly different (as clarified in Sect. 4.2). This is at the basis of the better performances provided, on average, by FanG-HPO. Currently, one of the most important practical advantages is that ML developers can choose among BoTorch-MOMF and FanG-HPO according to their programming skills, specifically coding in Python or R. On the other hand, we are currently working on porting our code in Python for evaluating its integration into the BoTorch platform.

It is important to remark that, whenever a real-life decision support system has to be developed, the choice of the most appropriate metrics, both in terms of error and fairness, must be carefully chosen depending on the specific problem/dataset. In this study we have just selected misclassification error and DSP for all the experiments in order to have a homogeneous experimental setting, even if they could be not the most appropriate choices for each one of the dataset. On the other hand, it is also interesting to remark that all the three approaches are *agnostic* to the underlying semantic of the adopted metrics, so they show similar behaviours also in the case that different metrics could be used.

Future works are going to investigate the possibility to also consider *cost-aware*—aka *location-dependent* or *frugal*—optimization, recently proposed in Lee et al. (2020), Candelieri & Archetti (2021a), Luong et al. (2021) and Wu et al. (2021), where nominal sources' query costs are not fixed but depends on the hyperparameters configuration to evaluate and can be learned along the optimization process. Moreover, validating FanG-HPO on more realistic datasets and a larger set of ML algorithms will allow us and other research groups to further extend and improve it.

More in detail, the github repository contains:

- The four pre-processed (fairness) datasets.
- All the basic R scripts implementing AGP, EHVI, and the core functionalities of FanG-HPO.
- One R script for each pair *dataset - ML algorithm* to run FanG-HPO starting from the associated autogluon-FairBO run.
- Python scripts (called from R scripts) for training MLP, RF, XGB, and SVM classifiers, given a specific configuration of their hyperparameters.
- All the results obtained with FanG-HPO: https://drive.google.com/drive/folders/14o7FbZAwUWfJn2QHn0foqRdHtVCob1tx?usp=sharing

Moreover, from the following Google Drive folders it is possible to download:

- Results of the autogluon-FairBO runs, used as a starting points for the associated FanG-HPO and BoTorch-MOMF runs: https://drive.google.com/drive/folders/1qxSU2iuyvf1BZFfkDyrYPLSueyFPc3J3?usp=sharing
- Results obtained with BoTorch-MOMF: https://drive.google.com/drive/folders/1060Zv3YvKLc8401FPwVd9VSVg_P-N57w?usp=sharing
- The autogluon-FairBO based HPO code implemented and used to run experiments: https://drive.google.com/drive/folders/1-2PYP6uS-r8Oe70ZwSxplJr6kaWPdDCM?usp=sharing (for installation and configuration, please use the official documentation of autogluon).
- The BoTorch-MOMF based HPO code implemented and used to run experiments: https://github.com/andreaponti5/FanG-HPO-MOMF.git (for installation and configuration, please use the official documentation of BoTorch).
- all the detailed results of all the FanG-HPO runs, along with results from the `zlrml` and `fgrrm` algorithms: https://drive.google.com/drive/folders/1H_l7yYqIC6Zy2e7LRLznQKe2GZ2sgPwe?usp=sharing

Appendix A

A.1 Multi-objective optimization

Figure 10 summarizes the main concepts of Pareto analysis in the multi-objective setting. For the sake of visualization, a two objectives problem, with a two dimensional search space Ω, is considered. On the left hand side the search space and the (unknown) Pareto

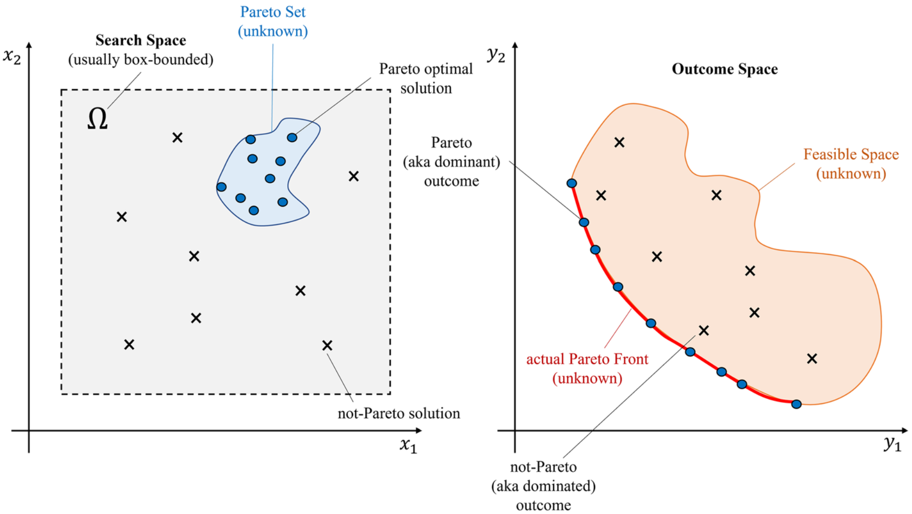

Fig. 10 On the left: box-bounded search space Ω, (unknown) Pareto set, Pareto and not-Pareto solutions. On the right: (unknown) feasible space within the outcome space (i.e., consisting of all the outcomes associated to solutions in the search space), Pareto (aka dominant) outcomes, not-Pareto (aka dominated outcomes), and actual (unknown) Pareto front

set are depicted; on the right hand side, the outcome space spanned by the two objectives is illustrated, along with the (unknown) feasible space (containing the outcomes associated to all the possible solutions in the search space) and the (unknown) actual Pareto front.

A.2 Gaussian process regression

A Gaussian process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution, and it is completely specified by its mean function, $\mu(\mathbf{x})$, and covariance function, $k(\mathbf{x},\mathbf{x}')$. A GP is denoted with $\mathcal{GP}\left(\mu(\mathbf{x}),k(\mathbf{x},\mathbf{x}')\right)$ (Williams & Rasmussen, 2006; Gramacy, 2020). Importantly, these two scalar-valued functions can be conditioned on a set of available observations, leading to a *probabilistic regression model* which can be used to make predictions at any location $\mathbf{x}$, according to the so-called GP's *predictive* (aka *posterior*) mean and standard deviation. While the first represents the predicted value, the second represents the associated predictive uncertainty.

Consider to have performed n queries, then denote with $\mathbf{X}^{1:n}=\left\{\mathbf{x}^{(i)}\right\}_{i=1:n}$ the set of queried locations and with $\mathbf{Y}^{1:n}=\left\{y^{(i)}\right\}_{i=1:n}$ the associated observed outcomes, possibly noisy (i.e., $y^{(i)}=f\left(\mathbf{x}^{(i)}\right)+\varepsilon^{(i)}$, where $\varepsilon^{(i)}$ is assumed to be a zero-mean Gaussian noise, $\varepsilon^{(i)}\sim\mathcal{N}\left(0,\sigma_\varepsilon^2\right),\forall\, i\in\{1,\dots,n\}$). Then, the GP's predictive mean and variance, conditioned to the n performed queries, are respectively computed as follows:

$$\mu(\mathbf{x})=\mathbf{k}\left(\mathbf{x},\mathbf{X}^{1:n}\right)\left[\mathbf{K}+\sigma_\varepsilon^2\mathbf{I}\right]^{-1}\mathbf{y} \tag{13}$$

$$\sigma^2(\mathbf{x})=k\left(\mathbf{x},\mathbf{x}\right)-\mathbf{k}\left(\mathbf{x},\mathbf{X}^{1:n}\right)\left[\mathbf{K}+\sigma_\varepsilon^2\mathbf{I}\right]^{-1}\mathbf{k}\left(\mathbf{X}^{1:n},\mathbf{x}\right) \tag{14}$$

where $\mathbf{K}$ is an $n\times n$ matrix whose entries are $k_{ij}=k\left(\mathbf{x}^{(i)},\mathbf{x}^{(j)}\right)$, and $\mathbf{k}(\mathbf{x},\mathbf{X}^{1:n})$ is the n-dimensional row vector $\left(k(\mathbf{x},\mathbf{x}^{(1)}),\dots,k(\mathbf{x},\mathbf{x}^{(n)})\right)$. Just for completeness, $\mathbf{k}\left(\mathbf{X}^{1:n},\mathbf{x}\right)$ is the n-dimensional column vector $\mathbf{k}\left(\mathbf{x},\mathbf{X}^{1:n}\right)^\top$. The GP's predictive uncertainty is the posterior standard deviation, that is $\sigma(\mathbf{x})=\sqrt{\sigma(\mathbf{x})^2}$.

Before conditioning a GP to a set of observations, two priors must be provided, relatively to the mean and the covariance functions. Usually, the first is set to zero: this is not a limitation because the posterior mean will be not confined to this value. However, it is possible to incorporate explicit basis functions for expressing prior information on the function to approximated by the GP model (Williams & Rasmussen, 2006). On the contrary, the covariance function is chosen among a set of possible *kernel functions*, such as Squared Exponential, Power Exponential, and Matérn kernels, offering different modelling options with respect to structural properties of the function to be approximated, especially *smoothness* (Gramacy, 2020; Archetti & Candelieri, 2019; Frazier, 2018; Williams & Rasmussen, 2006).

Acknowledgements We greatly acknowledge the DEMS Data Science Lab, Department of Economics Management and Statistics (DEMS), University of Milano-Bicocca, for supporting this work by providing computational resources.

Author contributions All the authors equally contributed to the design of the approach and the writing of the manuscript. AC and AP equally contributed to the code development and experiments execution. All the authors contributed to the analysis of the results and the final conclusions.

Funding Open access funding provided by Università degli Studi di Milano - Bicocca within the CRUI-CARE Agreement. The study is not part of the activities of any co-founded project.

Availability of data and material Pre-processed datasets, as used in the study, are publicly available at:https://github.com/acandelieri/FanG-HPO_jml

Declarations

Conflicts of interest Authors declare no conflicts of interests / competing interests.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Code availability FanG-HPO was developed in R and integrates, through the `reticulate` R package, Python code (i.e., sklearn modules for ML algorithms: MLP, RF, XGB, and SVM). Code is publicly available at the following github repository: https://github.com/acandelieri/FanG-HPO_jml.

References

Archetti, F., & Candelieri, A. (2019). *Bayesian optimization and data science*. Springer.

Ariafar, S., Mariet, Z., Brooks, D.H., Dy, J.G. & Snoek, J. (2021). Faster & more reliable tuning of neural networks: Bayesian optimization with importance sampling. In *AISTATS* (pp. 3961–3969).

Barocas, S., Hardt, M. & Narayanan, A. (2017). Fairness in machine learning. NIPS tutorial.

Belakaria, S. & Deshwal, A. (2019). Max-value entropy search for multi-objective Bayesian optimization. In *International conference on neural information processing systems (NeurIPS)*.

Belakaria, S., Deshwal, A., & Doppa, J. R. (2020). Multi-fidelity multi-objective Bayesian optimization: an output space entropy search approach. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*, 10035–10043.

Belakaria, S., Deshwal, A., Jayakodi, N. K., & Doppa, J. R. (2020). Uncertainty-aware search framework for multi-objective Bayesian optimization. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*, 10044–10052.

Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91). PMLR.

Candelieri, A. & Archetti, F. (2021a). Miso-wildcosts: Multi information source optimization with location dependent costs. arXiv preprint arXiv:2102.04951.

Candelieri, A., & Archetti, F. (2021). Sparsifying to optimize over multiple information sources: an augmented gaussian process based algorithm. *Structural and Multidisciplinary Optimization, 64*, 1–17.

Candelieri, A., Perego, R., & Archetti, F. (2021). Green machine learning via augmented gaussian processes and multi-information source optimization. *Soft Computing, 25*, 1–13.

Cruz, A.F., & Hardt, M. (2023). Unprocessing seven years of algorithmic fairness. arXiv preprint arXiv:2306.07261.

Daulton, S., Balandat, M., & Bakshy, E. (2020). Differentiable expected hypervolume improvement for parallel multi-objective Bayesian optimization. *Advances in Neural Information Processing Systems, 33*, 9851–9864.

Dhar, P. (2020). The carbon impact of artificial intelligence. *Nature Machine Intelligence, 2*(8), 423–425.

Feliot, P., Bect, J., & Vazquez, E. (2017). A Bayesian approach to constrained single-and multi-objective optimization. *Journal of Global Optimization, 67*(1–2), 97–133.

Frazier, P.I. (2018). Bayesian optimization, recent advances in optimization and modeling of contemporary problems (pp. 255–278). INFORMS.

Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Hamilton, S.C.E.P. & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 329–338).

Garnett, R. (2023). *Bayesian optimization*. Cambridge University Press.

Ghoreishi, S. F., & Allaire, D. (2019). Multi-information source constrained Bayesian optimization. *Structural and Multidisciplinary Optimization, 59*(3), 977–991.

Gramacy, R. B. (2020). *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*. Chapman and Hall/CRC.

Hao, K. (2019). *Training a single AI model can emit as much carbon as five cars in their lifetimes*. MIT Technology Review.

He, X., Zhao, K., & Chu, X. (2021). Automl: A survey of the state-of-the-art. *Knowledge-Based Systems, 212*, 106622.

Hernández-Lobato, D., Hernandez-Lobato, J., Shah, A. & Adams, R. (2016). Predictive entropy search for multi-objective Bayesian optimization. In *International conference on machine learning* (pp. 1492–1501). PMLR.

Hort, M., Chen, Z., Zhang, J.M., Sarro, F. & Harman, M. (2022). Bias mitigation for machine learning classifiers: A comprehensive survey. arXiv preprint arXiv:2207.07068.

Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). *Automated machine learning: Methods, systems, challenges*. Springer Nature.

Iqbal, M.S., Su, J., Kotthoff, L. & Jamshidi, P. (2020). Flexibo: Cost-aware multi-objective optimization of deep neural networks. arXiv preprint arXiv:2001.06588.

Irshad, F., Karsch, S. & Döpp, A. (2021). Expected hypervolume improvement for simultaneous multi-objective and multi-fidelity optimization. arXiv preprint arXiv:2112.13901.

Jamieson, K. & Talwalkar, A. (2016). Non-stochastic best arm identification and hyperparameter optimization. In *Artificial intelligence and statistics* (pp. 240–248). PMLR.

Kennedy, M. C., & O'Hagan, A. (2000). Predicting the output from a complex computer code when fast approximations are available. *Biometrika, 87*(1), 1–13.

Khatamsaz, D., Peddareddygari, L., Friedman, S. & Allaire, D.L. (2020). Efficient multi-information source multiobjective Bayesian optimization. In *AIAA Scitech 2020 Forum* (pp. 2127).

Klein, A., S. Falkner, S. Bartels, P. Hennig, & F. Hutter 2017. Fast Bayesian optimization of machine learning hyperparameters on large datasets. In Artificial Intelligence and Statistics (pp. 528–536). PMLR.

Komiyama, J., Takeda, A., Honda, J. & Shimao, H. (2018) Nonconvex optimization for regression with fairness constraints. In International conference on machine learning (pp. 2737–2746). PMLR.

Lam, R., D. Allaire, & K.E. Willcox 2015. Multifidelity optimization using statistical surrogate modeling for non-hierarchical information sources. In *56th AIAA/ASCE/AHS/ASC structures, structural dynamics, and materials conference* (pp. 0143).

Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., & Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12*(3), e1452.

Lee, E.H., Perrone, V., Archambeau, C. & Seeger, M. (2020). Cost-aware Bayesian optimization. arXiv preprint arXiv:2003.10870 .

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research, 18*(1), 6765–6816.

Luong, P., Nguyen, D., Gupta, S., Rana, S., & Venkatesh, S. (2021). Adaptive cost-aware Bayesian optimization. *Knowledge-Based Systems, 232*, 107481.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR), 54*(6), 1–35.

Nguyen, G., Biswas, S. & Rajan, H. (2023). Fix fairness, don't ruin accuracy: Performance aware fairness repair using AutoML. arXiv preprint arXiv:2306.09297 .

Paria, B., Kandasamy, K. & Póczos, B. (2020). A flexible framework for multi-objective bayesian optimization using random scalarizations. In *Uncertainty in Artificial Intelligence* (pp. 766–776). PMLR.

Perrone, V., Donini, M., Kenthapadi, K. & Archambeau, C. (2020). Bayesian optimization with fairness constraints. In *International conference on machine learning (automated machine learning workshop).*

Perrone, V., Donini, M., Zafar, M.B., Schmucker, R., Kenthapadi, K. & Archambeau, C. (2021). Fair bayesian optimization. In *Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society* (pp. 854–863).

Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys (CSUR), 55*(3), 1–44.

Poloczek, M., Wang, J. & Frazier, P. (2017). Multi-information source optimization. *Advances in Neural Information Processing Systems* 30 .

Schmucker, R., Donini, M., Perrone, V., Zafar, M.B. & Archambeau, C. (2020) Multi-objective multi-fidelity hyperparameter optimization with application to fairness. In *NeurIPS Workshop on Meta-Learning* (Vol. 2).

Schwartz, R., Dodge, J., Smith, N., & Etzioni, O. (2020). Green AI. *Communications of the ACM, 63*(12), 54–63.

Scutari, M., Panero, F. & Proissl, M. (2021). Achieving fairness with a simple ridge penalty. arXiv preprint arXiv:2105.13817.

Strubell, E., Ganesh, A. & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243 .

Sun, Q., Chen, T., Liu, S., Chen, J., Yu, H., & Yu, B. (2022). Correlated multi-objective multi-fidelity optimization for HLS directives design. *ACM Transactions on Design Automation of Electronic Systems (TODAES), 27*(4), 1–27.

Suzuki, S., S. Takeno, T. Tamura, K. Shitara, & M. Karasuyama 2020. Multi-objective Bayesian optimization using pareto-frontier entropy. In *International conference on machine learning* (pp. 9279–9288). PMLR.

Svenson, J., & Santner, T. (2016). Multiobjective optimization of expensive-to-evaluate deterministic computer simulator models. *Computational Statistics & Data Analysis, 94*, 250–264.

Swersky, K., Snoek, J., & Adams, R. P. (2013). Multi-task Bayesian optimization. *Advances in Neural Information Processing Systems, 26*, 2004–2012.

Tornede, T., Tornede, A., Hanselle, J., Mohr, F., Wever, M., & Hüllermeier, E. (2023). Towards green automated machine learning: Status quo and future directions. *Journal of Artificial Intelligence Research, 77*, 427–457.

Verma, S. & Rubin, J. (2018). Fairness definitions explained. In *2018 IEEE ACM international workshop on software fairness (FairWare)*, pp. 1–7. IEEE.

Weerts, H., Pfisterer, F., Feurer, M., Eggensperger, K., Bergman, E., Awad, N., Vanschoren, J., Pechenizkiy, M., Bischl, B. & Hutter, F. (2023). Can fairness be automated? Guidelines and opportunities for fairness-aware AutoML. arXiv preprint arXiv:2303.08485 .

While, L., Bradstreet, L., & Barone, L. (2011). A fast way of calculating exact hypervolumes. *IEEE Transactions on Evolutionary Computation, 16*(1), 86–95.

Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning* (Vol. 2). MIT Press.

Wu, Q., Wang, C., & Huang, S. (2021). Frugal optimization for cost-related hyperparameters. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*, 10347–10354.

Yang, K., Emmerich, M., Deutz, A., & Bäck, T. (2019). Multi-objective Bayesian global optimization using expected hypervolume improvement gradient. *Swarm and Evolutionary Computation, 44*, 945–956.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2019). Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research, 20*(1), 2737–2778.

Zhan, D., Cheng, Y., & Liu, J. (2017). Expected improvement matrix-based infill criteria for expensive multiobjective optimization. *IEEE Transactions on Evolutionary Computation, 21*(6), 956–975.

Zhang, R. & Golovin, D. (2020). Random hypervolume scalarizations for provable multi-objective black box optimization. In *International conference on machine learning* (pp. 11096–11105). PMLR.

Zhao, G., Arroyave, R. & Qian, X. (2018). Fast exact computation of expected hypervolume improvement. arXiv preprint arXiv:1812.07692

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Antonio Candelieri[1] · **Andrea Ponti**[1,2] · **Francesco Archetti**[3]

✉ Antonio Candelieri
antonio.candelieri@unimib.it

Andrea Ponti
andrea.ponti@unimib.it

Francesco Archetti
francesco.archetti@unimib.it

[1] Department of Economics, Management, and Statistics, Univesrity of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milan, Italy

[2] OAKS SRL, Milan, Italy

[3] Department of Computer Science, Systems, and Communication, University of Milano-Bicocca, Viale Sarca, 336, 20126 Milan, Italy