



AutoML in heavily constrained applications

Felix Neutatz¹ · Marius Lindauer² · Ziawasch Abedjan²

Received: 30 January 2023 / Revised: 16 August 2023 / Accepted: 10 October 2023 / Published online: 17 November 2023
© The Author(s) 2023

Abstract

Optimizing a machine learning pipeline for a task at hand requires careful configuration of various hyperparameters, typically supported by an AutoML system that optimizes the hyperparameters for the given training dataset. Yet, depending on the AutoML system's own second-order meta-configuration, the performance of the AutoML process can vary significantly. Current AutoML systems cannot automatically adapt their own configuration to a specific use case. Further, they cannot compile user-defined application constraints on the effectiveness and efficiency of the pipeline and its generation. In this paper, we propose CAML, which uses meta-learning to automatically adapt its own AutoML parameters, such as the search strategy, the validation strategy, and the search space, for a task at hand. The dynamic AutoML strategy of CAML takes user-defined constraints into account and obtains constraint-satisfying pipelines with high predictive performance.

Keywords AutoML · Constraints · Meta-Learning

1 Introduction

Recently, there has been intensive research on automated machine learning (AutoML) to facilitate the design of machine learning (ML) pipelines [4, 5, 18, 24, 26, 29, 47, 53–55, 59, 61]. Existing work entails hyperparameter optimization, neural architecture search, and the generation of end-to-end ML pipelines, consisting of data preprocessing, feature engineering, model selection, and postprocessing.

1.1 AutoML with constraints

In practice, AutoML can be subject to two kinds of constraints: *ML application* and *Search* constraints. *ML application* constraints impose restrictions, such as limits on training/inference time and ML pipeline size, or additional quality criteria, such as adversarial robustness or differential privacy, on the final ML pipeline. The ML application

constraints on resource consumption are particularly relevant in systems that work with dynamic data and rely on fast response time [36, 52]. *Search* constraints impose restrictions on the AutoML search process itself, such as limiting the search time, main memory usage, or parallelism.

Depending on the real-world setting and its commanding constraints, users have to configure the AutoML system differently to achieve the optimal result within a limited search time budget. With emerging applications in the realm of edge computing and real-time analysis, further constraints need to be considered. Autonomous driving relies on real-time video analysis [13] and to keep up with a sufficiently high frame rate, the model has to follow tight inference time constraints. As ML models have become successful, they have also gained traction on smaller devices, such as smartphones, requiring them to reduce their memory footprints and to predict fast. For streaming use cases, it might be important to continuously retrain to adapt to concept drift over time [11]. For fast-changing environments, such as fraud detection for high-frequency transactions, the models are subject to demanding training time constraints. Further, streaming ML requires constraints on millisecond latency and high throughput [21, 42]. There are also concerns regarding population-based quality criteria. For example, Schelter et al. [49] showed that mean-value imputation introduces bias and should be omitted from the ML hyperparameter search space if the application is subject to fairness constraints.

✉ Felix Neutatz
f.neutatz@tu-berlin.de

Marius Lindauer
m.lindauer@ai.uni-hannover.de

Ziawasch Abedjan
abedjan@lbs.uni-hannover.de

¹ TU Berlin, Berlin, Germany

² Leibniz Universität Hannover, Hanover, Germany

AutoML systems have several *AutoML parameters*, such as those defining the search space, the search strategy, e.g., different variants of Bayesian optimization and evolutionary algorithms, the validation strategy, e.g., hold-out and cross-validation, and the sampling strategy, which strongly influence the search process. We call an arbitrary initialization of these parameters an *AutoML configuration*. The *default AutoML configuration* is the initialization of each AutoML parameter with its default value and typically enables the entire search space for ML hyperparameters.

Generally, there is no single AutoML configuration that always yields a model with high predictive performance on all kinds of datasets and in particular subject to any of the aforementioned constraints. Typically, expert knowledge is required to configure and adapt an AutoML system to such settings.

1.2 Adapting AutoML configurations

We envision an AutoML system that automatically adapts to a user-specified *ML task*, i.e., not only to the dataset but also taking into account user-defined *ML application constraints* and *search constraints*, to achieve the best overall anytime performance. We call this new paradigm *constraint-driven AutoML*, where the data scientists and domain experts who know the constraints of the ML applications upfront, e.g., resource restrictions for IoT devices or legal restrictions, only need to specify the constraints but do not need to manually adjust the space of pipeline designs. We note that AutoML addresses two groups of users: non-domain experts seeking low- or no-code solutions, and ML experts seeking support in their day-to-day business. We rather address the latter user group with knowledge of the task-specific constraints.

State-of-the-art AutoML systems [14, 17, 43] do not support ML application constraints out of the box, and they do not adapt the search process to user-specified search constraints. Both adaptations are in fact non-trivial because AutoML systems have many of their own parameters, such as those defining the search space, the search strategy, and the validation strategy. For instance, if the user specifies a search time of five minutes, the well-known AutoML system Auto-Sklearn [17, 18] will consider the same ML hyperparameter search space as if it had a whole week, although only a very small fraction of the ML hyperparameter space can be covered. Theoretically, users could modify the AutoML system parameters to reduce the search space. Still, even for experts, it is difficult to estimate which part of the ML hyperparameter space to consider or which sample size suffices for a given task. Similar to ML hyperparameter sensitivity to the dataset at hand, AutoML's anytime performance strongly depends on its own parameters and their optimal setting depends on the ML task. An intuitive approach would be to frame the problem as a multi-objective optimization task to explore ML

pipelines across all constraint dimensions. However, even if we consider it a multi-objective optimization problem, it is still unclear how to select the AutoML parameters to search efficiently.

We propose an efficient solution for constraint-driven AutoML by leveraging meta-learning, which so far has only been applied to a few subproblems in our setting. For instance, Auto-Sklearn2 [17, 18] leverages meta-learning to warm-start Bayesian optimization (BO). Specifically, it searches for the best set of ML hyperparameters on all datasets in a repository. For a new dataset, it compares the dataset with all datasets in the repository and applies BO with an initial portfolio of ML hyperparameter configurations of the most similar dataset to accelerate the search. Additionally, it learns which validation strategy and initial portfolio are beneficial for which dataset. However, Auto-Sklearn2's meta-learning approach cannot support constraints because one would need to independently train the meta-learning for each possible set of constraint settings, which is infeasible. Further, their approach only supports predicting discrete strategy decisions using pairwise meta-modeling, i.e., a meta-model predicts the better out of two possible AutoML strategies. This approach cannot handle continuous AutoML parameters, and even covering all possible combinations of sampling strategy, validation, and search space strategy is typically infeasible.

Another meta-learning approach is to learn a surrogate model that learns offline whether a given ML pipeline can satisfy specified constraints. However, this approach does not adapt the AutoML parameters [37]. For instance, it is not possible to adapt the validation strategy based on the specified constraints.

To remedy the aforementioned limitations and to enable all degrees of freedom in constraint-driven AutoML, we addressed three major challenges:

1. **Huge meta-learning space.** The combined space of AutoML parameters, constraints, and datasets is huge. We need to draw meta-training instances from this huge space to enable the meta-training. To prune the ML hyperparameter space, we have to consider the trade-off between search runtime and predictive performance. If we prune too much of the ML hyperparameter space, the optimization might not find ML pipelines with high predictive performance. If we prune too little, the search might be inefficient. To estimate which AutoML configurations will be successful, it is critical to consider the dataset and user-specified constraints.
2. **Meta-training labels.** To predict an AutoML configuration for a given task, a meta-model has to be trained on similar tasks. Choosing the right meta-training examples and an appropriate prediction target is a problem we intend to solve.

3. **Nondeterministic AutoML.** AutoML is a nondeterministic and stochastic process. Across multiple runs, the same AutoML configuration might lead to significantly different outcomes because both the AutoML optimizer (e.g., Bayesian optimization) and ML model training are stochastic. So, if we naively train a meta-model on such a noisy signal, the meta-model might be inaccurate.

1.3 Contributions

To address these challenges, we propose a new constraint-driven AutoML system, CAML, which dynamically configures its AutoML parameters by taking into account the user-specified ML task (i.e., dataset and constraints). Learning from previous *AutoML runs* (i.e., dataset, constraints, AutoML configuration), CAML generates AutoML configurations and estimates which of them are promising for a new ML task. To this end, we make the following contributions:

1. We propose alternating sampling as a training data generation strategy—a combination of active learning, Bayesian optimization, and meta-learning. It is parallelized and efficiently explores the huge search space of datasets, AutoML configurations, and constraints to learn a meta-model that estimates the success of AutoML configurations and accelerates the search process.
2. To instantaneously extract the most promising AutoML configurations from the meta-model at runtime, we propose offline AutoML configuration mining that provides CAML with a large pool of promising AutoML configurations. As the meta-model can rank 100k configurations in less than a second, this pool allows for fast AutoML configuration retrieval.
3. To ensure high adaptability for a wide set of constraint settings, we implemented CAML in a way to allow the user to configure whether or not it optimizes any ML hyperparameter. It also supports ML application constraints based on metrics, such as training/inference time, ML pipeline size, and equal opportunity [22]—a fairness metric.
4. We report extensive experiments with CAML and compare it to state-of-the-art AutoML systems. We provide our implementation, datasets, and evaluation framework in our repository [39].

Main Findings. Our study lets us draw the following conclusions:

1. CAML does not only outperform the default AutoML configuration but also state-of-the-art systems, such as TPOT [43], AutoGluon [14], and Auto-Sklearn2 [17], in constrained settings.
2. CAML outperforms hand-tailored constraint-specific AutoML solutions, such as Auto-Sklearn 2 [17]. Man-

ually adapting AutoML system configurations to diverse constraints or even combinations of multiple constraints is nearly impossible due to unforeseeable side effects. Therefore, solutions, such as CAML, are required.

3. CAML is the first step toward our vision of constraint-driven AutoML. This way, we can cover multiple diverse constraints and add/remove additional ones without AutoML systems expertise.

2 Three-step problem

The three-step problem represents the search for the optimal setting of three parameters as described in Fig. 1: the AutoML parameters, ML hyperparameters, and model parameters.

Before we formalize the problem of constraint-driven AutoML, we formally define the problem of finding optimal model parameters for a given supervised machine learning model and the AutoML problem of finding the optimal algorithm and ML hyperparameters, e.g., selecting a data encoding, feature preprocessor, and classification model, and all their corresponding hyperparameters.

2.1 Supervised ML problem

The supervised ML problem is to find the parameters θ for a predictive model f by minimizing the loss \mathcal{L} of mapping $f : x_i \mapsto \hat{y}_i$ for a given training dataset $D_{\text{train}} = \{(x_0, y_0), \dots, (x_n, y_n)\}$.

$$\theta^* \in \arg \min_{\theta \in \Theta} \sum_{(x_i, y_i) \in D_{\text{train}}} \mathcal{L}_{\text{train}}(y_i, f(x_i; \theta)). \quad (1)$$

In practice, the problem is often more complex since the loss might be regularized to achieve better generalization performance, and stochastic optimizers might lead to different model parameters returned by the learning process.

2.2 The AutoML problem

The combined algorithm selection problem and hyperparameter optimization problem of AutoML [56] is to determine the predictive pipeline $a \in A$ and its corresponding hyperparameters $\lambda \in \Lambda$, inducing a model $f^{(a, \lambda; D_{\text{train}})}(\cdot; \hat{\theta})$ with some approximated model parameters $\hat{\theta}$, that achieve the lowest loss on the validation set D_{valid} . Formally:

$$\arg \min_{a \in A, \lambda \in \Lambda} \sum_{(x_i, y_i) \in D_{\text{val}}} \mathcal{L}_{\text{val}}(y_i, f^{(a, \lambda; D_{\text{train}})}(x_i; \hat{\theta})). \quad (2)$$

We note that the training loss $\mathcal{L}_{\text{train}}$ (e.g., cross-entropy) does not have to be the same as the validation loss \mathcal{L}_{val} (e.g., balanced accuracy). Since the ML model training can already

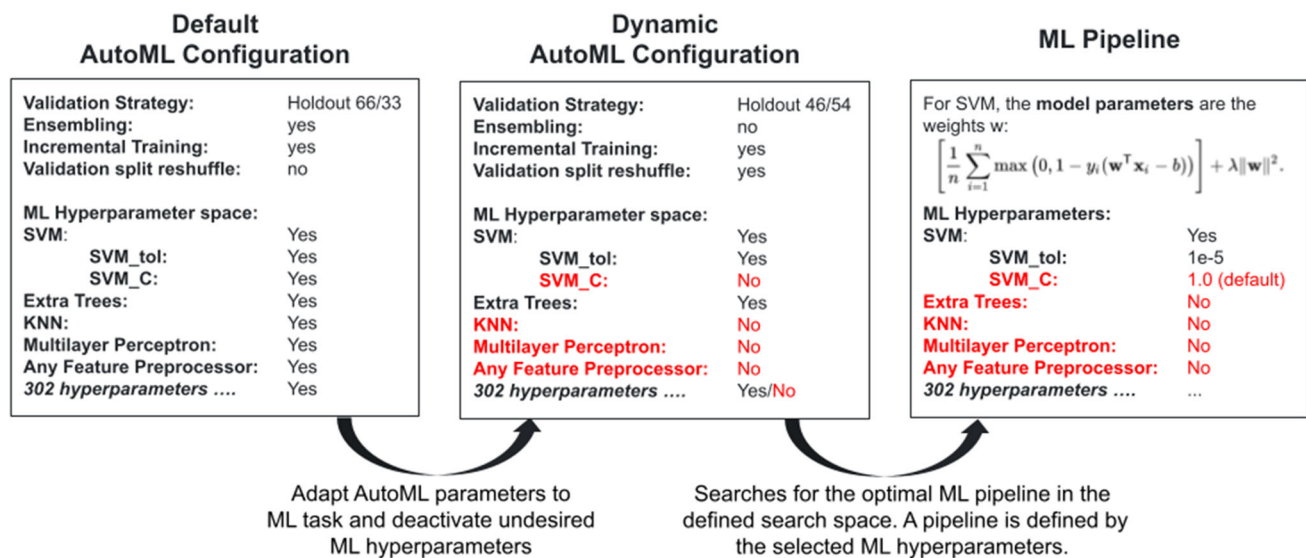


Fig. 1 Example: Adapting the AutoML Parameters for constraint-driven AutoML

take some time (e.g., training a DNN), AutoML has to be very efficient in evaluating different configurations from $A \times \Lambda$.

2.3 Constrained-driven AutoML problem

The problem that we address in this paper is to find the parameters ω of a given AutoML system to efficiently find an ML pipeline that adheres to all user-specified constraints and achieves the highest predictive performance for a specified ML task. Formally,

$$\max_{\omega \in \Omega} m(\omega) \text{ s.t. } \forall c_i \leq t_i, i \in [0, n] \quad (3)$$

where ω is a vector representing an AutoML system's own configuration; $m(\omega)$ is the average validation loss of the final ML model \hat{f} returned by the AutoML system; c_i are the constraints, and t_i are the user-specified constraint thresholds, i.e., search time ≤ 5 min or ML pipeline size ≤ 1 MB. For constraints, we distinguish between search constraints and ML application constraints. Search constraints concern the AutoML search process, such as search time, search main memory, and evaluation time, and ML application constraints concern the final ML pipeline, such as training/inference time, and fairness.

Although optimizers with implicit learning of these unknown constraints can be used, we hypothesize that zero-shot adjusting of the AutoML system's own parameters (including the configuration space $A \times \Lambda$) will address this problem efficiently.

Choosing the AutoML configuration based on a specified dataset and constraints is challenging because both the solution space (possible AutoML configurations) as well as the task space (possible datasets and constraint thresholds)

are huge. Any change in any of these components might affect the final predictive performance. The non-determinism of both ML and AutoML further aggravates these challenges.

Figure 1 illustrates how constraint-driven AutoML impacts the configurations. Instead of using the default AutoML configuration, our system automatically adapts its AutoML parameters to the user-specified *ML task*, which is defined by the dataset and constraints at hand. In this example, several classification methods are excluded as they are expected not to meet the specific constraint (marked red in the Dynamic AutoML Configuration). Then, the dynamically configured AutoML system searches for ML pipelines based on the remaining ML hyperparameter search space. Finally, the *model parameters* are fit to the dataset, e.g., SVM tunes the weights w . Previously disabled hyperparameters either remain disabled if irrelevant or are set to default if required. For example, the dynamic AutoML configuration excluded the regularization parameter of the SVM model. However, as the final pipeline uses SVM, it will simply use the default parameter here.

3 Constraint-driven AutoML

To meta-learn AutoML's own parameters ω , we propose CAML, as illustrated in Fig. 2. Given a user-specified dataset, search constraints, and ML application constraints, CAML decides which AutoML configuration—namely, which ML hyperparameter space, search strategy, and validation strategy—a given AutoML system should search to yield an ML pipeline with high predictive performance. The workflow consists of an offline and an online phase.

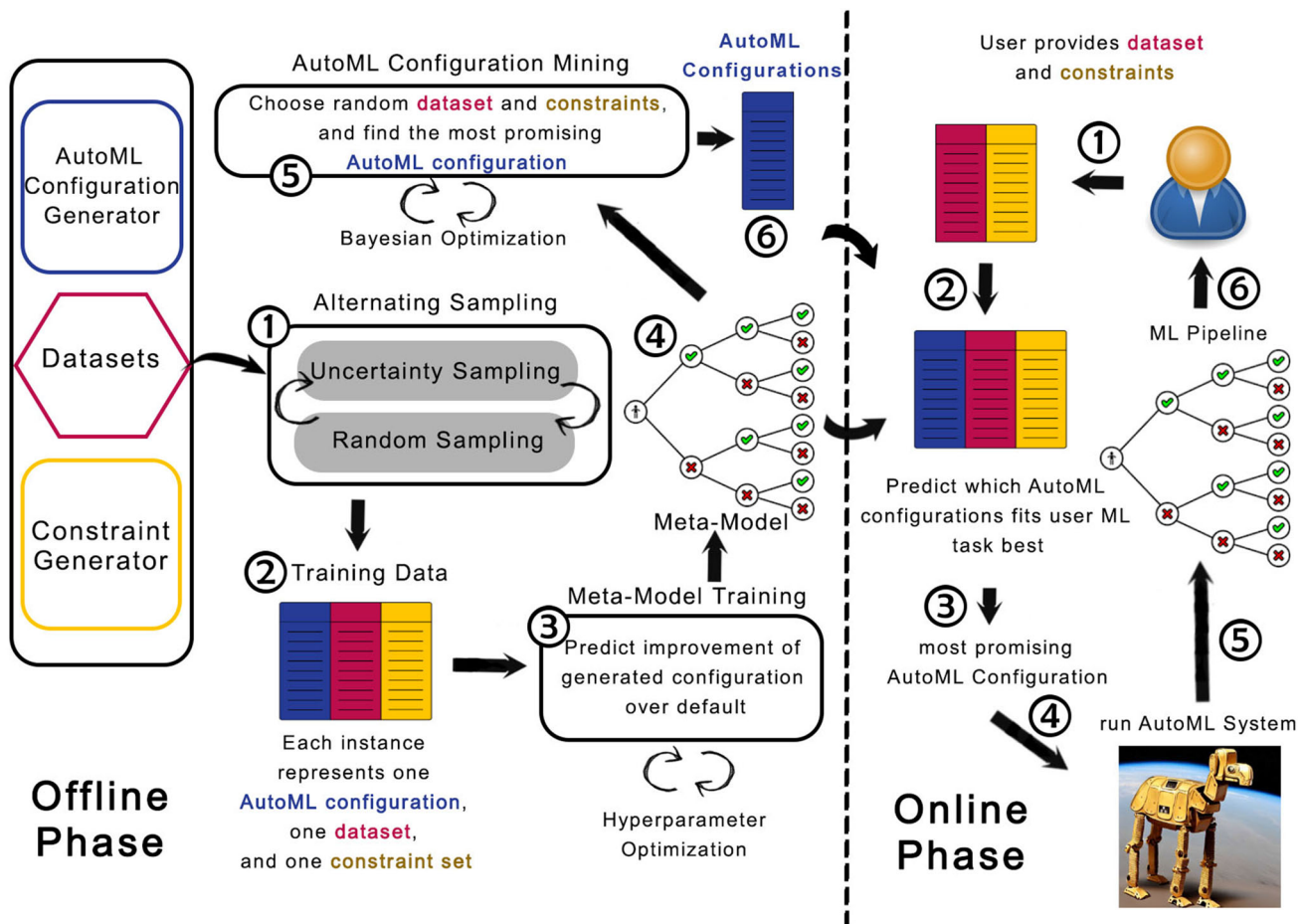


Fig. 2 System architecture of CAML

The **offline phase** consists of three main steps: training data generation, meta-model training, and AutoML configuration mining. As input, the AutoML system engineer has to provide the AutoML space and the constraint space via generators. In this paper, we benchmark training/inference time, pipeline size, and equal opportunity constraints. The engineer can extend this constraint set depending on the ML application's needs. Further, CAML requires a repository of datasets. Meta-learning performs better if the user-provided datasets are similar to the datasets that are present in the repository. So there are two possible approaches to create the repository. Inside organizations, one could resort to the own history of datasets that were used in prior data science pipelines. Other than that one should create the repository with datasets that differ in dimensions, such as the number of instances, features, classes, missing values, and feature types. There are already public repositories that to some degree fulfill this diverse requirements. Following prior studies, we collected the datasets for our benchmark repository from platforms, such as OpenML [58], UCI ML Repository [27], Kaggle [50], and HuggingFace [10]. CAML leverages

an alternating strategy ① of random and uncertainty sampling to both explore and exploit the huge space of AutoML configurations, datasets, and constraints. Based on the resulting training data ②, CAML learns and optimizes the meta-model using cross-validation while ensuring cross-dataset generalization ③.

Ideally, the meta-training would consider the best AutoML configuration for each ML task. However, identifying the best configuration for an ML task is nearly impossible as it would require testing the huge space of configurations per ML task. As this goal is computationally infeasible, we relax our original problem formulation from Sect. 2 as follows:

$$\max_{\omega \in \Omega} P(m(\omega) > m(\omega_{\text{default}})) \text{ s.t. } \forall c_i \leq t_i, i \in [0, n] \quad (4)$$

To ensure a robust meta-learning approach, our intuition is to identify the AutoML configuration that is most likely more effective than the default AutoML configuration. Thus, we train a meta-learning model that predicts whether a configuration that is different from the default AutoML configuration will result in better performance for a given task.

In the final step of the offline phase, CAML leverages the meta-model ④ to search for the estimated optimal AutoML configuration for a random dataset and random constraints ⑤. CAML leverages BO to address this search problem. The result of this step is a large pool of promising AutoML configurations ⑥ for a diverse set of use cases.

In the **online phase**, the user specifies the dataset and the constraints ①. To prepare them for the meta-model training, we encode both the dataset and the constraints in the meta-feature representation (see Sect. 3.1.4) and combine them with the mined AutoML configurations ②. Then, CAML leverages the meta-model to predict which of the mined AutoML configurations fits the user-specified dataset and constraints best ③. Then, CAML equips the AutoML system with the resulting AutoML configuration ④ and executes it ⑤ with the specified search constraints. Finally, the AutoML system returns an ML pipeline that satisfies all ML application constraints ⑥.

3.1 Training data for meta-learning

We propose active meta-learning—an approach to efficiently apply meta-learning in a scenario where the corresponding training data, both instances and labels, do not exist and need to be generated; A meta-training instance comprises a combination of a dataset, an AutoML configuration, and constraints. The label of such a training instance should specify how fitting or successful generated AutoML parameters are. The meta-model should learn from a set of such training instances whether a generated configuration leads to better performance than the default AutoML configuration.

To train such a meta-model, we have to answer the following questions: How do we generate the labels? How can we effectively gather training data? How do we encode an AutoML run as meta-features?

3.1.1 Meta-target label

To learn which AutoML configurations are promising, we need a meta-training dataset with prediction labels for previous AutoML runs. We need to define what *success* means for a given AutoML run. We cannot simply choose the predictive performance as a label for an AutoML run, because the performance lives on different scales depending on the ML task at hand. Some ML tasks are harder to solve because some constraints are very restrictive. For instance, the constraint “ML pipeline size $\leq 5\text{KB}$ ” is more restrictive than “ML pipeline size $\leq 500\text{MB}$ ”, leading to different optimally achievable prediction performance values. Therefore, we have to find a metric that considers the entire context of an ML task as an anchor point. To provide such an anchor

point, we run the AutoML system with default configuration as a baseline during meta-learning. The default AutoML configuration uses the full ML hyperparameter search space and the default AutoML parameters, such as hold-out validation with 33% validation data. Now, our learning task is to predict whether a generated AutoML configuration yields higher predictive performance than the default AutoML configuration for the same task. This proxy metric is independent of the performance scales and the constraint hardness. To account for the nondeterministic behavior of AutoML, we run the AutoML system several times (ten times in our experiments) for both the generated configuration and the default configuration. Then, we obtain the fraction of cases where the default AutoML configuration was outperformed. We note that this might not lead to the optimum as defined in Eq. 3, but ensures a robust choice of an AutoML configuration, avoiding performance degradation caused by non-determinism. To avoid unnecessary computation for unsatisfiable settings in the meta-training, we first evaluate the given AutoML configuration. If all ten runs yield no ML pipeline that satisfies the specified constraints, we do not need to evaluate the default AutoML configuration anymore.

The meta-model for active learning is a random forest regression model that predicts the fraction of runs that the given AutoML configuration outperformed the default configuration. As shown before [56], random forest is a well-suited model for handling large complex and structured hyperparameter spaces, see Sect. 3.1.4.

Algorithm 1 Training data generation

Input: AutoML system A , Datasets D , Constraint Space C , AutoML parameter space Ω , Random iterations K , Sampling time t .

Output: X, Y , groups.

```

1:  $X \leftarrow \emptyset$ 
2:  $Y \leftarrow \emptyset$ 
3: groups  $\leftarrow \emptyset$ 
4: for  $i = 0$  to  $K$  do                                     ▷ cold start
5:    $d, c, \omega \leftarrow \text{random\_sample}(D, C, \Omega)$ 
6:    $X \leftarrow X \cup \{\text{encode}(d, c, \omega)\}$ 
7:    $Y \leftarrow Y \cup \{A(d, c, \omega)\}$                                ▷ Running CAML
8: while  $t$  not elapsed do                                     ▷ alternating sampling
9:   if  $\text{rand}() \geq 0.5$  then
10:    meta_model.fit( $X, Y$ )
11:     $d, c, \omega \leftarrow \arg \max_{d \in D, c \in C, \omega \in \Omega} \sigma(\text{meta\_model.predict}(\text{encode}(d, c, \omega)))$ 
12:  else
13:     $d, c, \omega \leftarrow \text{random\_sample}(D, C, \Omega)$ 
14:     $X \leftarrow X \cup \{\text{encode}(d, c, \omega)\}$ 
15:     $Y \leftarrow Y \cup \{A(d, c, \omega)\}$                                ▷ Running CAML
16:    groups  $\leftarrow \text{groups} \cup d$ 
17: return  $X, Y$ , groups.

```

3.1.2 Alternating sampling

To efficiently explore the space of AutoML configurations, datasets, and constraints, we leverage active learning, specifically uncertainty sampling [51]. Similar to the approach presented by Yu et al. [62] that reduces labeling effort for standard ML classification tasks, our system chooses and generates those meta-training instances that the meta-model is most uncertain about. However, if we only sample ML tasks around the decision boundary of whether a given AutoML configuration outperforms the default configuration, we might miss configurations that outperform the default configuration by large margins. While we *exploit* the space with uncertainty sampling, we additionally *explore* it with random sampling in an alternating fashion.

Algorithm 1 describes the training data generation process. Sampling requires a repository of datasets, an AutoML system, a constraint space, and a space of AutoML parameters. To start active learning, we need initial training instances that yield the first meta-model. CAML chooses these first instances randomly (Lines 4–7). In particular, CAML randomly chooses the dataset d , the constraints c , and the AutoML configuration ω (Line 5). Then, those components are encoded as meta-features and added to the meta-training set (Line 6). The corresponding AutoML run is executed and compared with the default configuration to obtain the corresponding label (Line 7). Then, the alternating sampling process starts (Line 8). The system chooses uniformly at random whether to apply random or uncertainty sampling. Uncertainty sampling picks the most uncertain instance among all given instances. To find uncertain instances in this huge search space (combinations of datasets, AutoML configurations, and constraints), we leverage BO, which learns a surrogate model to predict which AutoML parameters yield high predictive performance and samples only promising instances by trading off exploration and exploitation. In Line 11, BO identifies the combination of (d, c, ω) that leads to the highest standard deviation across all trees of the random forest meta-model. We repeat this two-step loop until the time limit has been reached.

3.1.3 Parallelization and optimizations

To speed up the presented sequential algorithm, we parallelize it asynchronously. Each worker always accesses the latest training instances. Once a new meta-training instance and a corresponding label are available, the meta-training data is locked briefly to add the new instance. We found that the more common approach [64] to predict the label for a newly sampled instance with the current meta-model and adding both to the meta-training data does not work well for our scenario. Our label is only predicted and is thus only an approximation of the ground truth. If the label is not correct,

the search could fall into the wrong direction. Therefore, our approach only adds a new instance once the label is confirmed. To avoid the same instances being evaluated in parallel, we start each nondeterministic BO run with different seeds. As the search space is huge, it is highly unlikely that similar instances will be sampled during the same period.

3.1.4 Meta-feature representation

To estimate whether an AutoML configuration yields higher predictive performance than the default AutoML configuration, the meta-model has to know the dataset, the AutoML parameters, and the constraints. We encode each of these components in a meta-feature vector.

Dataset Features For encoding datasets into meta-feature vectors, multiple approaches have been proposed [7, 18, 57]. We leverage the 32 meta-features proposed by Feurer et al. [18], such as the class entropy, the number of features, classes, and instances.

Constraint Features All constraints, such as inference time $\leq 0.001s$, can be represented by the corresponding threshold. If the user does not specify the constraint, we set the maximum possible default value. Extending the set of constraints is always possible. The safest strategy is to train the meta-model from scratch. However, one can also leverage the assumption that the missing constraint was simply set to default. Thus, all previous training instances can be appended with the default value for the new constraint and new instances with novel thresholds for the constraint can be generated for new instances. This way, we can continue meta-training asynchronously without the need of starting from scratch. The same reasoning applies to extending the search space of the AutoML parameters. However, this only works, if one does not change the underlying AutoML system that we compare to, e.g., if one uses the state-of-the-art AutoML system as a comparison, one can leverage the assumption that the missing component was simply not chosen. This way, we can continue meta-training without the need of starting from scratch.

AutoML Configuration Features To encode an AutoML configuration, we distinguish numeric parameters and categorical ones. Numeric AutoML parameters, such as the choice of the validation fraction, are simply added to the meta-feature vector. We encode the ML hyperparameters as binary values. The AutoML system either optimizes each ML hyperparameter (*True*) or uses its corresponding default value (*False*). For instance, the AutoML system can optimize the number of neighbors for K nearest neighbors or use its default $K = 5$.

We follow the well-known assumption that the ML hyperparameter space has a tree structure where each node represents an ML hyperparameter [3, 56] and each edge represents the dependency on its parent ML hyperparameter.

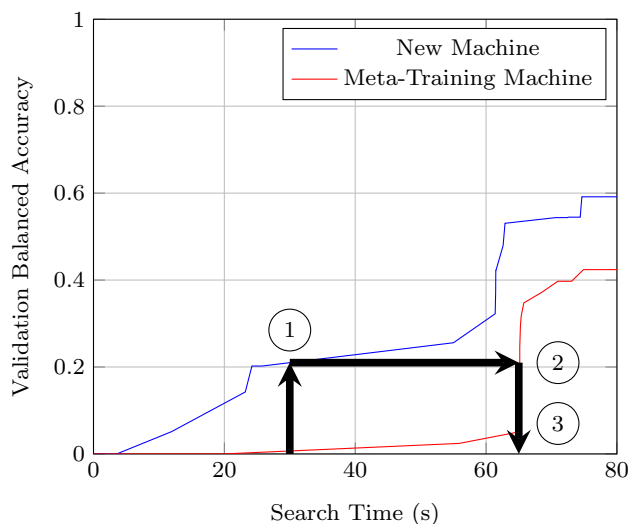


Fig. 3 Mapping search time from one environment to another environment

Figure 5 shows a branch of this tree. We describe the details of how we structure this tree in Sect. 3.4. If we do not optimize an ML hyperparameter higher up in the tree, we will not optimize any of its descendant ML hyperparameters either. For instance, if we remove the K -nearest-neighbor classifier from the choice of possible classifiers, we also do not need to optimize the number of neighbors k . We refer the reader to our repository [40] for the complete tree space that we leverage.

The aforementioned set of meta-features assumes uniform hardware specifications at training and deployment time which cannot always be guaranteed. If the hardware of meta-learning training is different from the hardware where CAML is deployed, one can apply calibration strategies that were proposed for database query optimization cost models [20]. For instance, one could run a lightweight benchmark to understand the hardware performance difference and obtain corresponding scaling functions.

We propose a simple calibration method to implement this idea. In particular, one can execute an AutoML system and keep track of the highest predictive performance on the validation set for one or multiple datasets that generally benefit from longer search times and use a performance mapping to calibrate the search time. One could implement the same idea using the test set but this would require additional computation and usually the test data is not accessible to the AutoML system during runtime.

During an offline step, the static CAML is applied on the selected datasets on both machines, the source and the target environment, and records the improvement of validation accuracy over time. These benchmarks lead to two graphs, as shown in Fig. 3. During the online process for a new dataset, one can now specify a desired search time

on the target machine, which will be internally mapped to a search time that achieves the same validation accuracy on the source machine. In Fig. 3, we marked 30s on the target machine and the graph visualizes how it maps to a different search time based on the equality of the validation accuracy. CAML searches for the search time where the meta-training machine reached this validation accuracy and uses the adjusted search time to configure the AutoML parameters for the new machine. Note that it still runs only 30s on the target machine but sets the configuration space based on the adjusted search time. The advantage of this calibration method is that it works for any hardware setup without the requirement of obtaining hardware meta-information. To improve the reliability of the calibration, one should conduct multiple runs and average the results. The approach will be costly if the targeted search times are rather high. However, we argue that in these cases calibration is not necessary as the search time is long enough. This is also validated by our experiments discussed in Sect. 4.7.

3.2 Meta-model training

Once the meta-data sampling is finished, CAML trains the final meta-model. The straightforward approach would be to use the same model that was trained for uncertainty sampling. However, this model is suboptimal because it might be overfitted to certain datasets that are more frequently sampled than others due to their uncertainty estimation. Further, we do not optimize the model hyperparameters during uncertainty sampling as it would significantly slow down the training data generation. For these reasons, we apply hyperparameter optimization on the meta-model after sampling has finished with 10-fold grouped cross-validation avoiding that training instances with the same dataset do not appear in both training and test folds.

To achieve optimal performance, we train two meta-models, one for AutoML configuration mining and one to rank the large pool of mined AutoML configurations.

For AutoML configuration mining, we use the same objective as for the surrogate model for uncertainty sampling (see Sect. 3.1.1): we predict the fraction of runs that the given AutoML configuration outperformed the default one (regression). For ranking the mined AutoML configurations, we predict whether the given AutoML configuration outperforms the default one at least once (classification). The regression meta-model contains more information than the classification meta-model because it estimates how much better the given AutoML configuration is compared to the default one whereas the classification model estimates only whether the AutoML configuration is better than the default one. However, as the regression task is much harder than the classification task, the regression meta-model is more likely to make mistakes and therefore more unstable. Yet, as we

describe in Sect. 3.3, we query the regression meta-model many times, avoid local optima/mistakes, and converge over time to a well-performing AutoML configuration. In turn, we only query the ranking meta-model once. Therefore, we need to make sure that it makes no mistakes and is as conservative as possible. This way, we ensure that the highest ranked AutoML configuration is robust—meaning it outperforms at least the default configuration.

3.3 AutoML configuration mining

Given an ML task and a generated configuration, the trained regression meta-model can predict whether the generated configuration will be more effective than the default configuration or not. The question is how we can leverage this regression meta-model to find the best AutoML configuration for a new dataset and user-specified constraints. To use the trained regression meta-model, we need a set of generated candidate configurations for each of which we can carry out the inference. Here, we are looking for the AutoML configuration that yields the best predictive performance for a given dataset and given constraints.

The simplest approach would be to generate a large number of random configuration candidates and let the regression meta-model predict which of these configurations has the highest likelihood of success. The disadvantage of this approach is that many of the randomly generated configurations will perform poorly and we cannot generate all possible configurations. The advantage of this approach is that the generation of these random configurations can be performed in the offline phase. During the online phase, we would only apply inference. The cost of inference is minimal, e.g., predicting one million configurations takes around 1 s.

Instead of random sampling, we could also apply BO. We could maximize the estimated likelihood that a generated configuration outperforms the default configuration, and freeze all meta-features for the user-specified dataset and constraints:

$$\hat{\omega} \leftarrow \arg \max_{\omega \in \Omega} \text{meta_model.predict}(\text{encode}(d, c, \omega)) \quad (5)$$

The advantage of BO is that it would adjust the configuration to the specified dataset and constraints. The disadvantage of BO is that it is slow. For instance, performing 1000 iterations would take more than 700 s. Waiting for more than 10 min before we even start the AutoML system is not viable—especially if the user is interested in fast development cycles.

We propose a hybrid approach that combines the strengths of both probing strategies. In the offline phase, we randomly sample a dataset and constraints—similar to Line 5. But instead of randomly sampling a configuration ω , we lever-

age BO to find the most promising configuration for this randomly generated ML task with the help of the regression meta-model. This way, we generate a large number of promising random configurations offline. In the online phase, we let the classification meta-model choose which of these promising random configurations fits the specified dataset and constraints best. Then, CAML sets up the actual AutoML system with this configuration and executes it.

3.4 AutoML parameters

Adapting AutoML parameters is only meaningful if there is a wide range of parameters that are in fact adaptable. In contrast to Auto-Sklearn and AutoGluon, we implemented CAML to not only provide access to the common user-adjustable AutoML parameters, such as whether to use ensembling, incremental training, or which validation strategy, but also to allow external adjustment of every single ML hyperparameter in the search space. This way, it can be dynamically decided whether those parameters should be optimized or not, as shown in Fig. 1.

We extend the ML hyperparameter space of Auto-Sklearn [18] additionally supporting oversampling strategies random oversampling, SMOTE [9], and ADASYN [23] to address class imbalance. Further, we added support for one-vs-rest classification to improve multi-class classification. We refer the reader to our repository [40] for the complete tree space that we leverage. We structure the ML hyperparameter space in a tree [40], as proposed in Auto-Weka [56]. Figure 5 represents a slice of the leveraged tree space. The first level of the tree contains all main components of the ML pipeline: categorical encoding, imputation, scaling, classifier, feature preprocessor, augmentation, sampling, and class weighting. Below this level, each component can be implemented by various strategies and each strategy has its own hyperparameters. This way, the ML hyperparameter space naturally builds up a tree. The hierarchical organization of the ML hyperparameter space is essential to allow the meta-model to prune a large part of the ML hyperparameter space as early as possible. This way, the AutoML system will not optimize the child ML hyperparameters if their parent ML hyperparameter is not optimized. Instead, the system will use their default value. For instance, by providing a hierarchical structure, we allow the meta-model to realize that no preprocessing transformation will be beneficial for a specific setting, instead of deciding for every single preprocessor and all its corresponding hyperparameters whether to optimize it or not.

3.5 Constraints

In constraint-driven AutoML, the user can define constraints, which might concern the AutoML process or the ML application, as shown in Fig. 4.

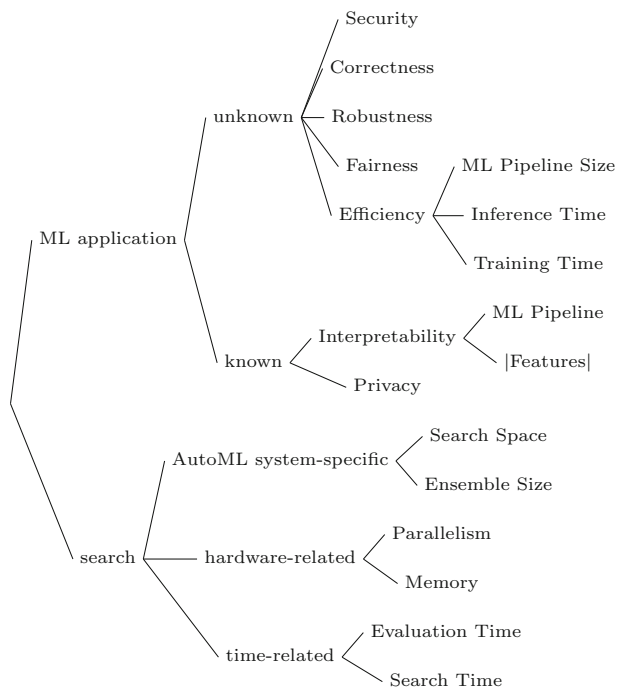


Fig. 4 AutoML constraints

Search constraints limit time-related, hardware-related, or system-specific aspects of the AutoML process. Time-related search constraints limit the search time or the evaluation time. Hardware-related search constraints are limits on the memory or parallelism. System-specific search constraints are limits on the size of ensembles or the search space.

The most important search constraint limits the *search time*. This search constraint is mandatory for each AutoML run and therefore it represents the class of search constraints well. For fast development cycles, data scientists will limit the search time to less than an hour to quickly experiment with the pipeline.

ML application constraints restrict the ML pipelines with regard to different quality dimensions. Zhang et al. [63] described 7 quality dimensions that can serve as constraints: correctness, robustness, security, privacy, efficiency, fairness, and interpretability. These constraints can be categorized along two dimensions.

Gelbart et al. [19] differentiate between unknown and known constraints as also illustrated in our constraint taxonomy. *Known constraints* are those constraints that can be checked before training and evaluating a model. For instance, knowing that an ϵ -differentially private implementation of classifiers [8] is used apriori ensures that privacy constraints are satisfied. Another example of known constraints is a restriction with respect to the ML pipeline components or the number of features to improve the interpretability of the resulting ML pipeline. In contrast, *unknown constraints* refer to those that can only be checked once the model is trained

and evaluated. For instance, most efficiency constraints have this property.

Generally, our approach can integrate any known constraint easily by adding an if statement at the beginning of the objective function. For our experiments, we focus on unknown constraints.

The second dimension along which one can differentiate constraints refers to their dependence on the ML pipeline and/or the data. For our experiments, we focus on constraints that significantly depend on the pipeline and not so much on the dataset. To incorporate more dataset-dependent constraints, such as fairness one would need to use more dataset-specific meta-features in the meta-model.

All in all, among the seven quality dimensions proposed by Zhang et al. [63], we focus on correctness, efficiency, and fairness. In particular, we always maximize correctness, i.e., the predictive performance. Further, we choose three well-established efficiency constraints *training time*, *inference time*, and *ML pipeline size*¹, and equal opportunity [22] which is a fairness measure. All four are *unknown* constraints and depend on the ML pipeline.

The relevance of the three efficiency constraints is particularly high in edge computing and streaming scenarios. In streaming scenarios, reducing inference time is vital to ensure continuous real-time predictions. As the data is evolving, the model requires constant retraining. In continuous training scenarios, enforcing training time limits plays a significant role. The same constraint type is relevant for federated learning [32], where users continue training on their own devices. Finally, to apply ML on IoT devices or smartphones, it is important to limit memory consumption.

3.6 Extending the list of constraints

First, one has to define the user-defined function that describes the constraint. The process depends on whether we want to create an ML application or search constraint.

For ML application constraints, one has to implement the following template:

```
def constraint(pipeline, training_time, X_train,
              y_train, X_val, y_val, threshold, constraint_
              specific): True/False
```

This function takes the trained pipeline and its training time, the split data, the constraint threshold, and constraint-specific parameters, such as the sensitive attribute for fairness. The output of this function is whether the given ML pipeline passes the constraint or not.

After implementing the user-defined function, one has to add a new feature to the meta-data representation and con-

¹ For some ML models, such as random forest and KNN, the model size is data dependent.

tinue meta-training. To account for the new meta-data feature, first one has to retrain the meta-model. With the retrained meta-model, one can continue alternating sampling including the new constraint. Finally, one has to generate additional configurations that also cover the new constraint as described in Sect. 3.3.

For search constraints, one has to additionally implement an initialize function that starts the measuring at the beginning of search and another function that checks whether the search constraint is still satisfied.

3.7 Constrained optimization

So far we know how to train the meta-learning approach and how to retrieve an adapted AutoML configuration dynamically. Now, we explain how CAML optimizes the ML hyperparameters under constraints. Previous systems by default consider the predictive performance as the objective function, which is not sufficient and requires adjustment. Furthermore, aspects such as ensembling have to be adjusted as we need to make sure that only constraint-satisfying models are ensembled and that the final ensemble also satisfies the constraints.

To support ML application constraints we formulate the objective function as follows for CAML:

$$\max \left(-1 \cdot \left(\sum_{i=1}^n \Delta c_i \right) + \left(\left[\sum_{i=1}^n \Delta c_i == 0 \right] \cdot BA \right) \right),$$

where Δc_i is the distance to satisfying the i th constraint and BA is balanced accuracy. This objective ensures to satisfy the constraints first and then optimizes the balanced accuracy. This way, the user can set thresholds for any of the supported constraints through an API. As the BO framework to maximize this objective, we choose Optuna [1], which leverages the tree-structured Parzen estimator (TPE) as the surrogate model. TPE is well-suited for our tree-structured ML search space.

To enable model ensembling in CAML, we integrate the greedy ensembling strategy proposed by Caruana et al. [6]. The strategy iteratively adds the model that maximizes ensemble validation predictive performance as long as all constraints are satisfied.

To enable hyperparameter optimization on large data, we implement incremental training similar to successive halving [31]. First, we train a model on a small sample containing 10 instances per class. Then, we double the training set size and train the model again. We continue this approach until either the evaluation time is over or the ML hyperparameter configuration is pruned because it performed worse than the median configuration of the history. Further, for constraint metrics that monotonically increase with the training set size, such

as the training time or ML pipeline size, we stop the configuration evaluation as early as possible to avoid unnecessary computation. As incremental training might result in a large number of ML hyperparameter evaluations, the danger of overfitting increases. Lévesque proposes to reshuffle the validation split after each evaluation to avoid overfitting [30]. Therefore, we implemented this option in CAML as well and expose it as an AutoML parameter.

4 Experiments

Our experiments aim to answer the following questions:

1. How does our dynamically configured AutoML system compare to state-of-the-art AutoML systems?
2. How does dynamic AutoML system configuration perform when one or multiple ML application constraints have been defined?
3. Is alternating sampling more efficient than random sampling for generating the meta-learning training data?
4. How does the number of mined AutoML configurations affect the predictive performance of CAML?
5. How does a changing the hardware environment affect the predictive performance of CAML?
6. How does the number of constraints affect meta-training?

4.1 Setup

We evaluate our approach on the same dataset split as used by Feurer et al. [17]: 39 meta-test datasets and 207 meta-train datasets. To extend our framework for fairness constraints, we add 17 fairness-related datasets provided by Neutatz et al. [41] to the meta-train datasets because common datasets do not annotate sensitive attributes that are required to measure fairness. As test datasets for fairness, we use the five fairness datasets that Ding et al. proposed to benchmark fair ML systems [12]. As a prediction accuracy metric, we leverage balanced accuracy that can handle binary, multi-class, and unbalanced classification problems. To compare the performance across datasets, we report the average and the standard deviation across datasets by repeatedly random sampling one result out of ten runs with different seeds with replacement. This approach ensures that we report the uncertainty induced by our system and not the different hardness scales of the datasets. Similarly, we test significance using the Mann–Whitney U rank test with $\alpha = 0.05$ between repeatedly random sampled averages. We mark a number with a star (*) if it passes this test. Note that in some cases the rounded average is very similar or the same, but one approach still passes the significance test to be better than the other approach. In these cases, we bold the results of the approach that passes the significance test.

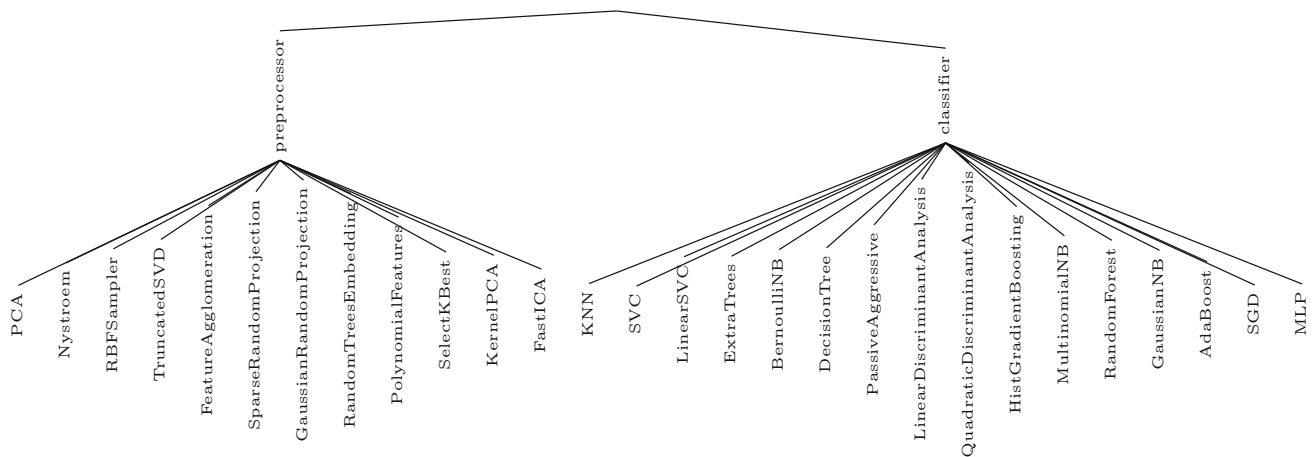


Fig. 5 Slice of the tree space that we use in our implementation

Due to our limited resources, we sample the meta-training dataset for two weeks, which amounts to 6,915 meta-training instances in total. Further, we mine AutoML configurations for two weeks using BO for 2,000 iterations, which amounts to 11,911 AutoML configurations. As AutoML parameter space, CAML chooses (i) the hold-out fraction, which affects both the size of training and the validation set, (ii) whether to use model ensembling, (iii) whether to use incremental training, (iv) whether to reshuffle the validation split, and (v) the whole adjustable ML hyperparameter space with 302 ML hyperparameters. Note that we do consider the time required for ensembling for the search time as it can be run in parallel to the model search as performed for Auto-Sklearn2 [17]. We ran the experiments on Ubuntu 16.04 machines with $28 \times$ Intel(R) Xeon(R) Gold 6132 CPU @ 2.60GHz cores and 264 GB memory.

Baselines. We compare our system with the state-of-the-art AutoML systems:

1. *TPOT* (0.11.5) is a genetic programming-based AutoML system that optimizes feature preprocessors and ML models [43].
2. *AutoGluon* (0.3.2) is an AutoML system that focuses on model ensembling and stacking [14].
3. *Auto-Sklearn2* (0.14.0) [17] is the latest version of the well-known AutoML system Auto-Sklearn1 [18] that leverages BO, meta-learning, and model ensembling to find the Sklearn [45] ML pipelines that achieve high predictive performance. Further, we extended the system to support the constraints for pipeline size, inference/training time, and fairness. We follow the same approach as described in Sect. 3.7 and only add a model to the ensemble if all constraints are satisfied. This allows a fair comparison of CAML and Auto-Sklearn2.
4. *Spearmint* [19] leverages BO for constrained optimization with Gaussian processes. We use the implementation

by Paleyes et al. [44] and search the same ML hyperparameter space as in our static system.

Furthermore, we evaluate our system with and without dynamic AutoML configuration: CAML *Dynamic* and CAML *Static*:

1. *CAML Static*. The static version covers the full ML hyperparameter space that is inspired by Auto-Sklearn1 [18]. It does not leverage meta-learning to optimize the search space. The details of the ML hyperparameter space are described in Sect. 3.4. We use the same ML hyperparameter ranges as Auto-Sklearn1. Further, the static version always leverages hold-out validation with 33% validation data, which is again the default validation strategy by Auto-Sklearn1. Additionally, it always uses model ensembling and incremental training.
2. *CAML Dynamic* implements our proposed approach. It automatically selects a subset of the full ML hyperparameter space and identifies the hold-out validation fraction, whether to use ensembling, incremental training, and validation split reshuffling.

In the following, we focus on a comparison and insights compared to Auto-Sklearn2 since it is the most similar system compared to ours and considered as one of the strongest systems to date.

4.2 Effectiveness on search time constraints

The most important constraint for AutoML limits the search time, which is a mandatory constraint that AutoML systems require because it is not obvious when to terminate an AutoML system. Therefore, it is crucial that our approach works well for this constraint as it also has to be fulfilled in combination with other constraints. We compare our dynam-

ically configured AutoML system CAML *Dynamic* with the same AutoML system with the default AutoML configuration CAML *Static*. Additionally, we compare our approach to state-of-the-art AutoML systems to show the potential of our idea of constraint-driven AutoML. We note that this is the only type of constraint easily applicable to all other AutoML systems considered in this study.

4.2.1 Performance comparison

Table 1 reports the average balanced accuracy for the meta-test datasets over time and across systems. We focus on search times of up to 60min as most state-of-the-art AutoML systems converge in this time period.

First, it is noticeable that CAML with the default AutoML configuration outperforms TPOT [43]. The reason is that CAML leverages incremental training, which is a multi-fidelity strategy. Therefore, it can yield ML pipelines in a short time, even for large datasets. However, CAML with the default AutoML configuration does not outperform Auto-Sklearn2 [17] and AutoGluon [14] for larger search times. It is noteworthy that Auto-Sklearn2 is a carefully optimized version of the Auto-Sklearn system [18] with a smaller hand-designed configuration space with six model classes. We also report the performance of Auto-Sklearn2 using the full ML hyperparameter space like Auto-Sklearn1. This version achieves significantly worse predictive performance, which shows that the right choice of the ML hyperparameter space is crucial.

Our approach CAML (Dynamic) with meta-learned AutoML configuration outperforms all other systems significantly according to the Mann–Whitney U rank test ($\alpha = 0.05$) until 5 min of search time. Note that both the pool of configurations that we choose the configurations from and the meta-model that chooses the configuration were generated with scenarios until 5 min of search time. This finding shows that our objective of dynamically choosing good AutoML configura-

tions was achieved if the scenarios were in the domain of the meta-training.

In fact, CAML *Dynamic* selects on average only 55 out of 302 ML hyperparameters for the search space and a 5-minute time frame and still achieves a higher average balanced accuracy across all experiments. Interestingly the search space only reduces slightly from here on. Having the 10 s constraint, 51 ML hyperparameters are considered on average, which is only four less than 55 for 5 min.

Yet, the space can also differ significantly between 5 and 1 minutes. Figure 6 shows AutoML configurations that were selected for the dataset “Christine” and “Robert” from the OpenML repository. The visualization follows the hierarchical view that we presented in Sect. 3.4 and displays the obtained configuration space for 1min and 5min search time, respectively. Comparing the ML hyperparameter spaces, we see that in this case the ML hyperparameter space for 1min search time is smaller than for 5min search time. This is because a higher time period allows for the optimization of more ML pipeline parameters.

Additionally, for the dataset “Christine”, our system chooses the validation fraction 0.13, ensembling, and incremental training. The small validation fraction reduces the time for evaluation. Ensembling makes the predictions more robust against noise. Incremental training ensures that the system finds a suitable ML pipeline early. In addition to incremental training, our system also chose to optimize the size of training set to further reduce the iteration overhead.

For the dataset “Robert”, our system chooses the validation fraction 0.54, incremental training, and validation split reshuffling. Validation split reshuffling avoids overfitting. Additionally, our system chose to optimize each class weight individually because the dataset has 10 classes.

Table 2 presents an example that shows the AutoML parameters chosen for the dataset “numerai28.6” under different search time constraints: 10s, 1min, and ≥ 5 min. Since our CAML Dynamic was trained on the data until 5 min, it will pick the same search space for search times greater than

Table 1 Search time constraint: Balanced accuracy averaged across 10 repetitions and 39 datasets comparing CAML to state-of-the-art AutoML systems

Strategy	10 s	30 s	1 min	5 min	30min	1 h
CAML						
Static	0.43 \pm 0.02	0.53 \pm 0.02	0.58 \pm 0.01	0.67 \pm 0.01	0.70 \pm 0.01	0.72 \pm 0.01
Dynamic	0.57 \pm 0.01*	0.67 \pm 0.01*	0.70 \pm 0.01*	0.74 \pm 0.00*	0.77 \pm 0.00	0.77 \pm 0.00
Auto-Sklearn2 opt	0.00 \pm 0.00	0.11 \pm 0.02	0.48 \pm 0.02	0.74 \pm 0.02	0.80 \pm 0.00	0.81 \pm 0.00
Auto-Sklearn2 full space	0.00 \pm 0.00	0.06 \pm 0.02	0.14 \pm 0.02	0.70 \pm 0.03	0.80 \pm 0.00*	0.81 \pm 0.00*
TPOT	0.00 \pm 0.00	0.00 \pm 0.00	0.31 \pm 0.03	0.47 \pm 0.04	0.67 \pm 0.02	0.68 \pm 0.01
AutoGluon	0.33 \pm 0.02	0.41 \pm 0.01	0.49 \pm 0.01	0.62 \pm 0.01	0.77 \pm 0.01	0.79 \pm 0.00
Spearmint	0.24 \pm 0.03	0.36 \pm 0.03	0.43 \pm 0.01	0.60 \pm 0.02	0.69 \pm 0.01	0.72 \pm 0.01

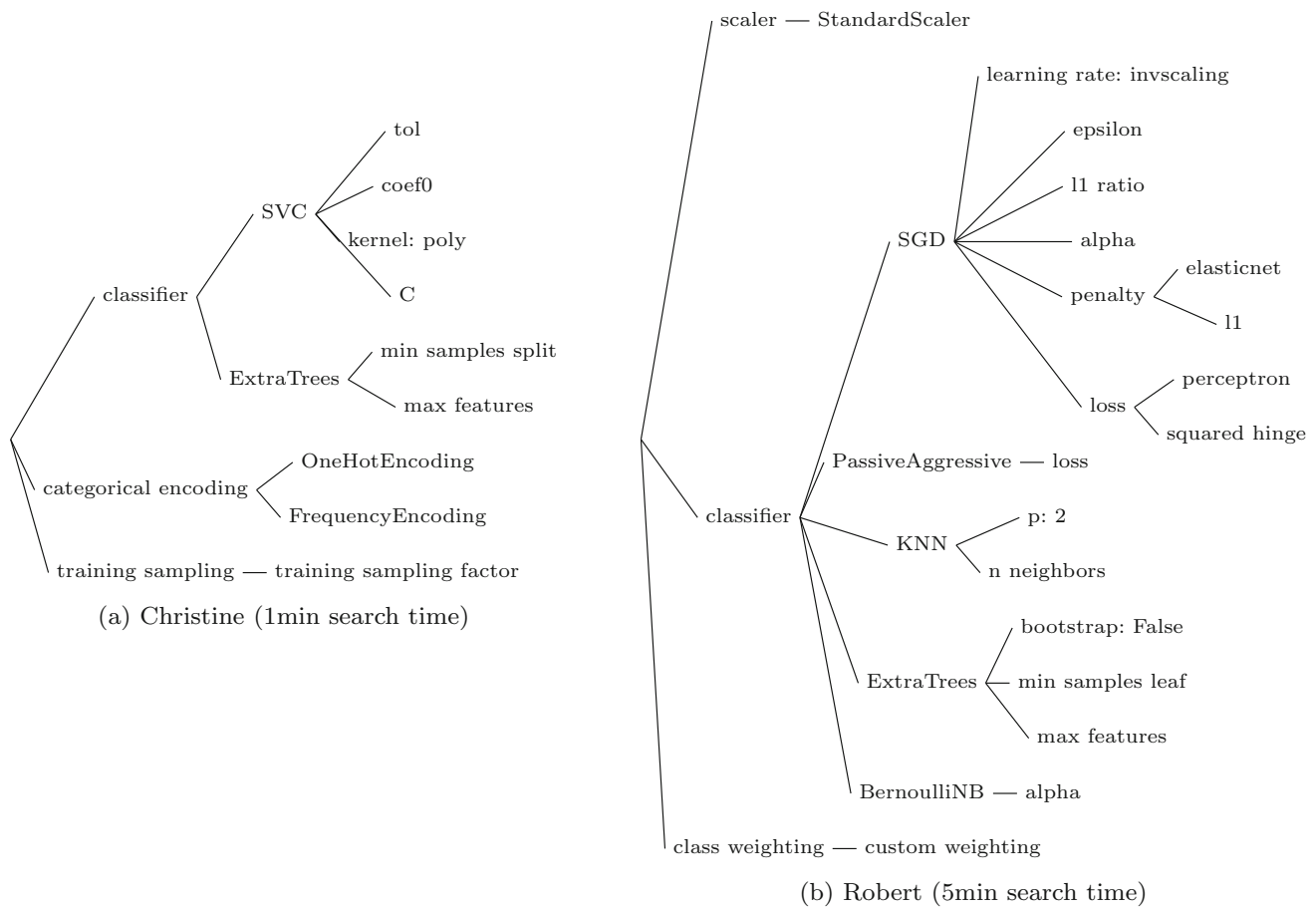


Fig. 6 Examples of ML hyperparameter spaces chosen by CAML Dynamic

5 min, which is why we did not consider higher search time constraints here.

Even for the very short search time of 10s and the rather large dataset with 96,320 instances, the search space still incorporates 9 out of 16 classifiers because of the incremental training, which enables fast skipping of poorly performing ML pipelines. For 1min, our system increases both the search space and the hold-out fraction. With this change, the hold-out validation score evaluation will take more time but will be more accurate. For ≥ 5 min, our system chooses to avoid incremental training. This way, model training will take more time but the models will be trained on more instances and are more likely to achieve a better generalization.

4.2.2 Analyzing the meta-models

To analyze the meta-models, we computed the meta-feature importance based on impurity scores for the trained random forest meta-model. We list the top-15 meta-features in Table 3 for the classification meta-model. The most important meta-features are the constraint thresholds, in particular, for the pipeline size, and inference/training time. These meta-

features are important because different constraints also require different AutoML configurations. This finding supports the aim of this work to consider dynamic AutoML configuration, especially for constrained settings. Another important feature is the hold-out fraction. Especially for large datasets, it is crucial to identify the right sample size to allow the AutoML system to yield any ML pipeline. For instance, for the dataset “KDDCup09 appetency” (50k instances), our method chooses a validation fraction of 7% of the data.

The remaining 8th-15th meta-features all cover dataset-specific meta-features, e.g., about the class distributions and the shape of the data. The meta-features representing the ML hyperparameter search space are less important, e.g., the meta-feature of whether to use a specific categorical encoding is the 37th most important feature.

For the regression meta-model, we list the top-15 meta-features in Table 4. The most important meta-features are similar to the ones for the classification meta-model. However, for the regression meta-model, the meta-feature that describes whether to use a feature *preprocessor* and whether to *incremental training*. Both decisions have a significant

Table 2 AutoML parameters chosen by CAML Dynamic for different search times on the dataset “numera128.6”

10s	1min	≥ 5min
<p>Holdout: 0.45 Ensemble: No Random shuffle: Yes Incremental: Yes</p>	<p>Holdout: 0.61 Ensemble: Yes Random shuffle: No Incremental: Yes</p>	<p>Holdout: 0.46 Ensemble: No Random shuffle: Yes Incremental: No</p>

Table 3 Meta-feature importances of the classification meta-model

Rank	Meta-feature	Importance
1	Pipeline size constraint	0.072
2	Inference time constraint	0.053
3	Training time constraint	0.044
4	Hold-out fraction	0.036
5	Search time constraint	0.023
6	Number of evaluations	0.022
7	Fairness constraint	0.020
8	Hold-out test instances	0.017
9	Evaluation time	0.017
10	instances	0.016
11	ClassProbabilitySTD	0.016
12	DatasetRatio	0.015
13	ClassProbabilityMax	0.015
14	ClassProbabilityMin	0.015
15	ClassEntropy	0.015

Table 4 Meta-feature importances of the regression meta-model

Rank	Meta-feature	Importance
1	Pipeline size constraint	0.072
2	Inference time constraint	0.056
3	Training time constraint	0.052
4	Hold-out fraction	0.043
5	Search time constraint	0.024
6	Preprocessor	0.023
7	Fairness constraint	0.022
8	Number of evaluations	0.022
9	Incremental training	0.021
10	ClassProbabilitySTD	0.020
11	ClassProbabilityMin	0.016
12	ClassEntropy	0.016
13	ClassProbabilityMax	0.015
14	Evaluation time	0.015
15	RatioNominalToNumerical	0.015

impact on how much the given AutoML configuration outperforms the default one.

Table 5 contains statistics about how often our system chooses a specific classifier across the 39 datasets and how

Table 5 Choice of the classifiers: AdaBoost (AdaB.), Bernoulli Naive Bayes (B.NB), Decision Tree (DT), Extra Trees (E.Trees), Gaussian Naive Bayes (G.NB), Histogram-based Gradient Boosting (HGB.), K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Linear Support Vector Classification (LSVC), Multi-

layer Perceptron (MLP), Multinomial Naive Bayes (M.NB), Passive Aggressive (PA), Quadratic Discriminant Analysis (QDA), Random Forest (RF), Stochastic Gradient Descent (SGD), Support Vector Classification (SVC)

Time	AdaB	B.NB	DT	E.Trees	G.NB	HGB	KNN	LDA	LSVC	MLP	M.NB	PA	QDA	RF	SGD	SVC	clf.]
10s	0.69	0.79	0.54	0.90	0.51	0.31	0.62	0.59	0.54	0.49	0.44	0.51	0.67	0.31	0.69	0.74	9.33
30s	0.82	0.77	0.51	0.95	0.67	0.28	0.62	0.69	0.49	0.46	0.54	0.56	0.69	0.28	0.67	0.90	9.90
1 min	0.85	0.74	0.51	0.97	0.62	0.28	0.69	0.72	0.41	0.44	0.56	0.41	0.72	0.28	0.72	0.87	9.79
5 min	0.79	0.82	0.72	0.95	0.59	0.38	0.72	0.54	0.54	0.49	0.54	0.64	0.72	0.36	0.82	0.85	10.46
30 min	0.79	0.82	0.72	0.95	0.59	0.38	0.72	0.54	0.54	0.49	0.54	0.64	0.72	0.36	0.82	0.85	10.46
1 h	0.79	0.82	0.72	0.95	0.59	0.38	0.72	0.54	0.54	0.49	0.54	0.64	0.72	0.36	0.82	0.85	10.46

many classifiers it chooses on average. The first observation is that the meta-model learned that it is beneficial to choose around ten classifiers to achieve high balanced accuracy fast. The Auto-Sklearn2 developers choose only 5 classifiers. However, since our system can decide for every single ML hyperparameter whether to optimize it, the search space stays small in comparison but adjusts itself to the specified dataset. In contrast to building Auto-Sklearn2, this approach is fully automatic and does not require any AutoML systems expertise. Auto-Sklearn2 uses a dynamic chooses the validation strategy. Additionally, its ML hyperparameter space has been manually tuned for accuracy and search time. Thus, users who want to apply Auto-Sklearn2 for a new constrained setting, would need to adjust the ML hyperparameter search space manually again. Further, we see that ExtraTrees are chosen frequently. The reason is that the computation cost is low and the prediction is robust due to ensembling.

For some classifiers, such as MLP and HGB, the frequency stays similar across search time constraints. The reason is twofold: First, using incremental training, we can quickly yield working models for both classifier types that are competitive across search times. Second, CAML identified that these classifier types work well for specific datasets which do not change across constraints. For instance, HGB was chosen for balanced datasets with less than 8 classes and more than 57 numeric features. MLP was chosen for skewed datasets with many categorical features.

Further, for some classifiers, such as LDA and SVC, the frequency increases with increasing search time and then decreases again. For instance, LDA benefits from an increasing number of training instances but is prone to overfitting for unbalanced data if one optimizes it for long enough. The training of SVC is very efficient and therefore, one can train an SVC with many instances in very little time. Therefore, we see a high frequency of 90% for 30s. With increasing search time, other more complex models, such as RF, replace it incrementally.

Table 6 Choice of AutoML parameters

Search time	10s	30s	1 min	5 min
Incremental training	0.97	0.97	0.90	0.82
Ensemble	0.57	0.55	0.56	0.62
Class augmentation	0.37	0.24	0.21	0.10
Validation split reshuffling	0.17	0.29	0.26	0.26

Finally, the frequency across models stays the same because both the pool of configurations that we sample from was generated with a maximum search time of 5mins and training data of the model that chooses the configuration has the same limit.

To understand the interaction among the other AutoML parameters, we report in Table 6 the fraction of datasets that a certain AutoML parameter was applied. First, we see that the choice for incremental training in most cases only decreases slightly with increasing search time. Incremental training ensures that we find ML pipelines independent of the dataset size. Model ensembling is also used frequently because it ensures robustness. Class augmentation is not applied often because most datasets are already balanced. Additionally, its use decreases with increasing search time. The reason might be that with a long enough search time, we can find a suitable model that internally addresses the class imbalance. Finally, validation split reshuffling is increasingly used with increasing search time. Greater search times lead to a higher number of iterations that in turn raise the risk of overfitting and reshuffling can help to reduce this risk. To the best of our knowledge, none of the state-of-the-art systems leverage this reshuffling strategy. Our results show that it is promising and justifies further research.

The choice of feature preprocessors is on par with the Auto-Sklearn2 implementation. Auto-Sklearn2 does not perform any feature preprocessing, and our dynamic approach follows the same strategy. The reason is that feature preprocessing might add more overhead than benefit for the predictive performance.

Table 7 We report the balanced accuracy for 5 min search time averaged across 10 repetitions and test datasets for four constraints

Percentile	2%	4%	8%	16%	32%
Pipeline size	4026B	6651B	8359B	16797B	32266B
Auto-Sklearn2	0.01 \pm 0.00	0.01 \pm 0.00	0.01 \pm 0.00	0.01 \pm 0.00	0.01 \pm 0.00
Spearmint	0.04 \pm 0.02	0.08 \pm 0.02	0.09 \pm 0.03	0.19 \pm 0.03	0.22 \pm 0.03
CAML					
Dynamic	0.25 \pm 0.01	0.39 \pm 0.01*	0.43 \pm 0.00*	0.54 \pm 0.01*	0.63 \pm 0.01*
Static	0.25 \pm 0.01*	0.39 \pm 0.01	0.42 \pm 0.01	0.49 \pm 0.01	0.59 \pm 0.01
Training time	0.009s	0.010s	0.012s	0.019s	0.078s
Auto-Sklearn2	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.01 \pm 0.01
Spearmint	0.00 \pm 0.01	0.01 \pm 0.01	0.00 \pm 0.01	0.00 \pm 0.01	0.05 \pm 0.02
CAML					
Dynamic	0.61 \pm 0.01*	0.62 \pm 0.01*	0.63 \pm 0.01*	0.68 \pm 0.01*	0.71 \pm 0.00*
Static	0.46 \pm 0.02	0.46 \pm 0.02	0.50 \pm 0.02	0.57 \pm 0.02	0.65 \pm 0.01
Inference time	0.00079s	0.00082s	0.00102s	0.00146s	0.00302s
Auto-Sklearn2	0.29 \pm 0.02	0.27 \pm 0.02	0.27 \pm 0.03	0.40 \pm 0.02	0.42 \pm 0.02
Spearmint	0.02 \pm 0.01	0.02 \pm 0.02	0.02 \pm 0.01	0.02 \pm 0.01	0.06 \pm 0.01
CAML					
Dynamic	0.42 \pm 0.02*	0.45 \pm 0.02*	0.57 \pm 0.02*	0.66 \pm 0.01*	0.74 \pm 0.00*
Static	0.25 \pm 0.03	0.26 \pm 0.03	0.38 \pm 0.03	0.52 \pm 0.02	0.64 \pm 0.02
Equal Opportunity	1.000	0.999	0.994	0.981	0.949
Auto-Sklearn2	0.50 \pm 0.00*	0.56 \pm 0.00	0.59 \pm 0.01	0.63 \pm 0.00	0.67 \pm 0.02
Spearmint	0.17 \pm 0.10	0.19 \pm 0.12	0.35 \pm 0.11	0.57 \pm 0.07	0.58 \pm 0.07
CAML					
Dynamic	0.10 \pm 0.00	0.61 \pm 0.05*	0.64 \pm 0.01*	0.67 \pm 0.01*	0.70 \pm 0.00*
Static	0.10 \pm 0.00	0.46 \pm 0.09	0.62 \pm 0.05	0.66 \pm 0.01	0.68 \pm 0.01

4.2.3 Conclusion

Our simple AutoML system with the default configuration is already competitive to state-of-the-art systems, such as AutoGluon and TPOT. This might be due to the fact that our approach leverages incremental training, therefore, can handle large datasets. Second, our dynamic AutoML configuration approach outperforms the same system with the default configuration for all search time constraints. Third, our dynamic approach is head-to-head with the hand-tuned Auto-Sklearn2 system, which was tuned by (Auto)ML system experts.

4.3 Effectiveness on diverse constraint types

To evaluate that our approach also achieves high balanced accuracy for constraint types other than search time, we provide experiments with constraints on ML pipeline size, the training time, the inference time, and equal opportunity (fairness metric). For the experimental setting, we obtain

constraint thresholds for the different constraint types in the following way. We ran random ML tasks for one day and obtained the distributions across all evaluated ML pipelines for all constraint types. On these, we can compute different percentiles to simulate different tight constraints. For training the meta-model, CAML *Dynamic* could freely choose any of the percentiles (with a maximum search time of 5 min). To compare CAML to the baselines, we used fairly tight constraints, i.e., the 2nd, 4th, 8th, 16th, and 32nd percentile of each distribution. We also evaluated the 1st percentile but the results are similar to the 2nd percentile and due to space limitations, we omit the corresponding results.

Other state-of-the-art systems, such as AutoGluon, TPOT, and Auto-Sklearn2, do not natively support these ML application constraints and are not easily extensible because their API only provides access to the ML pipeline predictions. However, we extended the best-performing AutoML system Auto-Sklearn2 to support constraints to allow a comparison to our system. Furthermore, constrained BO systems, such as Spearmint [19], GPflowOpt [28], ADMMBO [2], and Ax [15], support arbitrary constraints for BO. As a representa-

tive of this class of systems, we benchmark Spearmlint with the ML hyperparameter space of our static system.

4.3.1 Performance comparison

Table 7 provides the results of these constraint thresholds. Across constraint types, CAML outperforms both baselines significantly. Only for equal opportunity, Auto-Sklearn2 achieves the best accuracy for very restrictive fairness constraints. The reason is that Auto-Sklearn2 uses Dummy classifiers if it does not find any other model. Dummy classifiers predict only one class. This way it is likely that both the majority and the minority group have very similar true positive rates and therefore very high equal opportunity. However, we decided against including dummy classifiers because users expect an AutoML system to fit actual ML models.

For the constraints inference and training time, our dynamic approach always outperforms our static approach. For pipeline size constraints, the static approach is better for restrictive thresholds. The reason is that pipeline size is more bound to the size of training set size and our default approach always uses incremental training. That means that it starts with a very small training dataset. So, if the pipeline size is not satisfied for such a small set, it will go to the next ML hyperparameter configuration immediately. Our meta-model might be too optimistic and try to avoid incremental training if possible because it has a higher chance of higher accuracies but might miss satisfying the constraints.

For fairness constraints, the dynamic and static approach perform similarly. The reason is that fairness is highly data dependent. Without explicit information about the sensitive attributes, it is harder for the meta-model to decide on the AutoML system configuration. Furthermore, the meta-training for fairness had access to much fewer datasets compared to the other constraints. Additional datasets might help the meta-model to generalize better. However, in case of missing values and fairness constraints, CAML independently learned to choose only median value imputation, which supports the finding by Schelter et al. [49] that mean-value imputation negatively affects fairness.

4.3.2 Analysis

To better understand how our system adapts the ML hyperparameter search space depending on the ML application constraints, we average the chosen classifiers for each ML application constraint and compare it to case using with no ML application constraint in Table 8.

For the pipeline size constraint, CAML avoids models that require more memory, such as extra trees (E.Trees), multi-layer perception (MLP), or the KNN classifier, which needs to store all training instances for inference. For the training time constraint, CAML shifts more to linear models, such as Linear Discriminant Analysis (LDA), Linear Support Vector Classification (LSVC), or Passive Aggressive (PA), because they can be trained faster. For the inference time constraint, CAML chooses significantly more often random forest to be part of the search space because its inference complexity is only $\mathcal{O}(t \log n)$ where t is the number of trees and n is the number of instances. For equal opportunity, CAML avoids models that amplify the bias in the data. For instance, KNN might amplify bias because it always decides based on the majority of the nearest neighbors. These insights confirm that we cannot optimize an AutoML system for one constraint and expect that the same optimization will also benefit other constraints.

4.4 Effectiveness on multiple constraint types

In practice, ML applications can be constrained in multiple dimensions simultaneously. To evaluate our system for multiple constraints simultaneously, we choose two constraint combinations training time/equal opportunity and inference time/pipeline size. For both constraint combinations, we apply all combinations of thresholds that were evaluated in Sect. 4.3. We report the difference in the average balanced accuracy that CAML *Dynamic* outperforms the static variant in Figs. 7 and 8.

In nearly all experiments, CAML *Dynamic* outperforms the static variant or achieves similar predictive performance. Only for very restrictive constraints, such as 100% equal opportunity or 4026B pipeline size, its performance was

Table 8 We report the average percentage difference in choice of ML classifiers depending on the ML application constraint

ML Application Constraint	AdaB.	B.NB	DT	E.Trees	G.NB	HGB.	KNN	LDA	LSVC	MLP	M.NB	PA	QDA	RF	SGD	SVC
None	0.79	0.82	0.72	0.95	0.59	0.38	0.72	0.54	0.54	0.49	0.54	0.64	0.72	0.36	0.82	0.85
Pipeline Size	-0.19	-0.43	-0.15	-0.38	-0.07	+0.06	-0.19	+0.03	-0.01	-0.20	-0.25	-0.25	-0.14	-0.03	-0.24	-0.30
Training Time	+0.07	-0.04	+0.03	+0.00	-0.17	+0.00	+0.03	+0.26	+0.21	-0.09	+0.05	+0.12	+0.03	+0.02	+0.08	+0.09
Inference Time	-0.12	+0.03	-0.01	-0.09	+0.14	+0.12	-0.13	+0.16	+0.14	+0.20	+0.16	+0.06	+0.00	+0.40	-0.19	-0.01
Equal Opp.	-0.23	-0.02	+0.12	-0.11	-0.23	-0.22	-0.52	+0.06	+0.14	+0.19	-0.26	-0.24	-0.12	-0.16	-0.10	-0.01

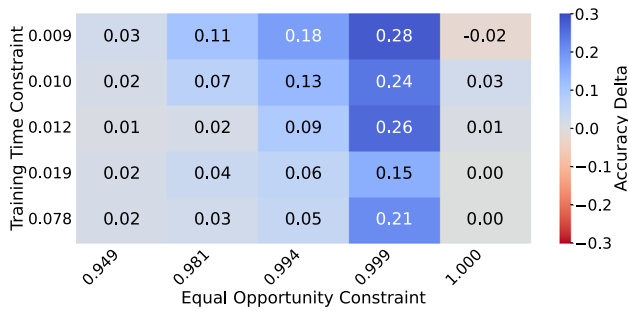


Fig. 7 We apply the constraints training time and fairness simultaneously and report the absolute distance to the average performance between the static and dynamic CAML. Higher numbers are better

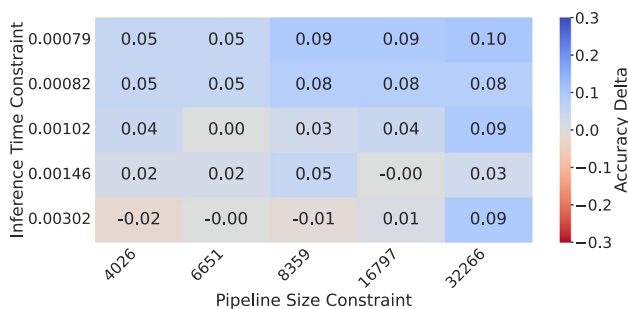


Fig. 8 We apply the constraints inference time and pipeline size simultaneously and report the absolute distance to the average performance between the static and dynamic CAML. Higher numbers are better

slightly lower because these constraints were for some of the test datasets not satisfiable. Overall, the experiments show that CAML *Dynamic* even works for multiple constraints. This finding shows that our meta-learning approach learns how these different constraints interact with each other. In 65% of the cases, CAML chooses AutoML configurations that consider two constraints simultaneously and were never chosen for the cases where we enforced only one of the constraints.

4.5 Alternating vs random sampling

One major design decision of our system is to leverage active learning in addition to random sampling to explore the huge space of AutoML parameters and constraints more efficiently. Table 9 provides the balanced accuracy for meta-models for both sampling approaches across two weeks of training data generation.

The alternating sampling approach outperforms the random sampling *significantly*. The reason is that active learning ensures that we sample along the decision boundary while random sampling ensures the diversity in the training data. Following a purely random sampling strategy results in lower final prediction performance and less consistent gains. For

Table 9 Predictive performance over meta-training time. We report the average balanced accuracy over training time averaged across 10 repetitions and 39 datasets comparing our system with random and alternating sampling

Days	Alternating	Random
2	0.70 ± 0.01*	0.67 ± 0.01
4	0.72 ± 0.01*	0.65 ± 0.02
6	0.70 ± 0.01	0.71 ± 0.01*
8	0.72 ± 0.01*	0.71 ± 0.02
10	0.73 ± 0.01*	0.72 ± 0.01
12	0.73 ± 0.01*	0.69 ± 0.01
14	0.74 ± 0.00*	0.72 ± 0.01

instance, after 12 days of random sampling, we achieve a worse predictive performance than for six days of sampling.

We can leverage Table 9 also to understand the impact of the training time. The numbers show that more training time benefits the meta-model, and even on the 14th day, we gain 1% more in average balanced accuracy. To conclude, alternating sampling outperforms random sampling significantly, and the longer we train, the better the dynamic AutoML configuration works.

4.6 AutoML configuration mining

Another important question for our approach is how many promising AutoML configurations we need to mine to achieve high predictive performance. Therefore, we experiment, for the search time constraint of 5min, with various fractions of the AutoML configurations that we mined within two weeks. We report the results in Table 10. With an increasing number of mined AutoML configurations, the predictive performance increases as well. The accuracy gain in percent might seem small but it is significant according to the Mann–Whitney U rank test. Further, the more constraints we add, the more diverse the pool of mined AutoML configurations needs to be to achieve high predictive performance across all constraints.

4.7 Adjusting to different hardware

To evaluate the described calibration approach in Sect. 3.1.4 that allows us to apply CAML on any machine, we conduct additional experiments on a powerful computer with Intel(R) Core(TM) i7-8565U CPU @ 1.80 GHz and 38 GB RAM. To benchmark both machines, we run CAML Static for the dataset “riccardo” for 10min for 10 times and measure the average validation balanced accuracy across the search time as reported in Fig. 3. We choose this dataset because it has 20k instances and 4k features, and it takes more time to con-

Table 10 Predictive performance over different numbers of mined AutoML configurations for search time of 5min. We report the average balanced accuracy over training time averaged across 10 repetitions and 39 datasets

Fraction	# Configurations	Accuracy
0.0002	3	0.729 \pm 0.00
0.0005	6	0.734 \pm 0.00
0.0010	12	0.736 \pm 0.00
0.0020	23	0.732 \pm 0.00
0.0039	47	0.736 \pm 0.00
0.0078	93	0.734 \pm 0.01
0.0156	186	0.739 \pm 0.01
0.0313	372	0.739 \pm 0.01
0.0625	744	0.739 \pm 0.01
0.1250	1489	0.735 \pm 0.01
0.2500	2978	0.744 \pm 0.00
0.5000	5956	0.743 \pm 0.00
1.0000	11911	0.747 \pm 0.00

Table 11 Search time constraint: Balanced accuracy averaged over 10 repetitions and 39 datasets comparing CAML to hardware adjusted CAML

CAML Strategy	10 s	30 s	1 min	5 min
Static	0.59 \pm 0.01	0.66 \pm 0.01	0.68 \pm 0.01	0.71 \pm 0.01
Dynamic	0.63 \pm 0.01	0.70 \pm 0.01	0.72 \pm 0.01	0.76 \pm 0.00
Dynamic (adjusted)	0.67 \pm 0.01	0.71 \pm 0.01	0.72 \pm 0.01	0.76 \pm 0.00

verge to find a well-performing model. If the data is too simple, we could not compare the convergence across time well. Then, we conduct the experiments for search times from 10 s to 5 min on the new machine with and without hardware adjustment as reported in Table 11. First, we see that even without hardware adjustment the CAML Dynamic still outperforms the static one. With hardware adjustment, for 10 s and 30 s, the average balanced accuracy improves by 4% and 1% accordingly. So, we conclude for small search budgets, hardware adjustment does improve our system. This finding also reduces the cost of the offline benchmark because CAML can run the benchmark for at most 10 min and turn off mapping for larger search times.

4.8 Impact of the number of constraints on meta-training

To analyze the impact of the number of constraints on the meta-learning performance, we run meta-training with up to 5 constraints, 4 ML application constraints and the mandatory search time constraint.

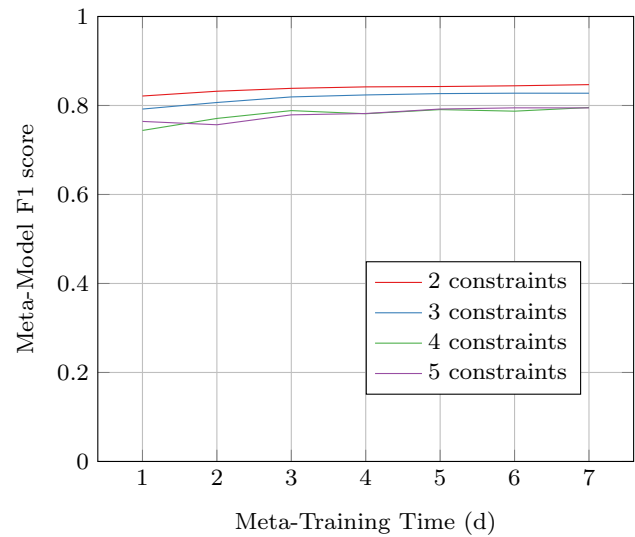


Fig. 9 Meta-Training Model F1 score under a varying number of constraints and different durations of meta-training

To compare the performance across constraint combinations, for each combination, we randomly pick random configurations, constraint thresholds, and datasets and evaluate whether the corresponding configuration performs better than the default configuration. We gather this test set for 1 week for each constraint combination. Then, for each number of constraints, we apply our alternating meta-training for 1 week. Finally, we report the F1 score of the meta-model on each test set reporting whether the corresponding configuration outperformed the default configuration at least once. We report the corresponding averaged F1 scores across 5 repetitions for each day of the week of meta-training in Fig. 9.

With increasing meta-training duration, the F1 score of the meta-model increases for all considered numbers of constraints. After one week, all constraint combinations reach an F1 score higher than 79%. As expected, by adding additional constraints, the F1 score slightly decreases by up to 3%, which is rather marginal considering the multiplied complexity of the resulting search space. This analysis can help to estimate how much time it may take to generate sufficient training data for larger search spaces.

5 Related work

Our work on constraint-driven AutoML combines research from various areas of optimization, AutoML, and meta-learning.

Constrained Optimization. One direction of work addresses constrained optimization by learning a surrogate model that estimates whether sampled configurations violate the corresponding constraints [2, 19, 28, 35, 46]. However, this

approach has two downsides. First, it requires the surrogate models to learn the constraints each time from scratch. Second, it cannot adjust the parameters of the AutoML systems, such as the validation approach or the search strategy, to the corresponding ML task.

Meta-Learning. A more effective approach is to learn upfront whether a given ML pipeline satisfies a well-known constraint, such as training time [37]. This approach does not require learning the constraint each time from scratch. Still, it does not adjust the AutoML parameters. Another direction is to meta-optimize the AutoML parameters. For instance, Lindauer et al. [34] optimize the parameters of hyperparameter optimization. However, they do not consider constraints. Further, Auto-Sklearn 2 [17] only supports predicting discrete strategy decisions using pairwise modeling. Therefore, their approach does not support continuous AutoML hyperparameters and does not scale to hundreds of settings. This scalability issue also hinders joint strategy prediction as the combinatorial space is too huge. Van Rijn et al. leverage meta-learning to identify the most important hyperparameter for various ML models individually [48]. However, they do not consider constraints. Alpine Meadow [54] uses the history of the quality and cost of all so far run pipelines to warm-start search, but can also not handle constraints.

Accelerating AutoML. Further, there is a large effort in the data management community to speed up AutoML systems. For instance, Li et al. propose to leverage search space decomposition [33]. Yakovlev et al. propose to leverage proxy models, iteration-free optimization, and adaptive data reduction to accelerate hyperparameter optimization [60]. Another well-known approach to speed up hyperparameter optimization is to leverage successive halving [16, 31]. It starts by evaluating many configurations on a small budget and incrementally chooses the best half of the configurations to evaluate them on a bigger budget. Xin et al. leverage caching to accelerate hyperparameter optimization [59]. However, their strategies cannot be applied in case of validation split reshuffling. Nakandala et al. propose a new parallel SGD execution strategy to speed up hyperparameter optimization for SGD-based models [38]. Hilprecht et al. [25] propose to make ML pipelines end-to-end differentiable to avoid costly Bayesian optimization. SystemDS [4, 5] allows users to specify ML programs in a declarative R-like language and compiles it to highly efficient hardware-specific code that can be distributed. Shah et al. [53] extensively benchmark feature-type detection that is important because the downstream AutoML system is dependent on the right feature-type classification. The aforementioned systems and algorithms are orthogonal to our contribution as they do not consider the search space of AutoML but optimize the computation for training and parameter tuning.

6 Conclusion

We proposed integrating constraints as a first-class citizen into AutoML—a paradigm that we call constraint-driven AutoML. As the constraints set limitations on the hyperparameter search, we proposed an approach to dynamically change the AutoML search space for the constraints at hand. To achieve this goal, we leverage active meta-learning. To explore the huge space of datasets, AutoML configurations, and constraints, we sample those combinations that benefit the meta-model. To show the full benefit of this approach, we develop a simple adjustable AutoML system, CAML, that exposes its whole ML hyperparameter space as binary AutoML parameters to have a task-specific search space. This way, CAML *Dynamic* can decide for every single ML hyperparameter whether it should be optimized or not. It automatically chooses an ML hyperparameter space for search time constraints that is similar to the space covered by the hand-tuned Auto-Sklearn2 [17] system. Overall, our new approach allows for configurable generic AutoML systems that dynamically adjust to the task and constraints at hand, and thus further increase the applicability of AutoML systems in practical application.

Acknowledgements This work was funded by the German Ministry for Education and Research as BIFOLD - Berlin Institute for the Foundations of Learning and Data (ref. 01IS18025A and ref. 01IS18037A), by the Federal Ministry of Education and Research (BMBF), Germany under the project AI service center KISSKI (grant no. 01IS22093C) and by the Federal Ministry of the Environment, Nature Conservation, Nuclear Safety and Consumer Protection, Germany under the project GreenAutoML4FAS (grant no. 67KI32007A).

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: a next-generation hyperparameter optimization framework. In: SIGKDD (2019)
2. Ariaifar, S., Coll-Font, J., Brooks, D.H., Dy, J.G.: ADMMBO: bayesian optimization with unknown constraints using ADMM. *J. Mach. Learn. Res.* **20**, 123:1–123:26 (2019)

3. Bergstra, J.S., Bardenet, R., Bengio, Y., Kégl, B.: Algorithms for hyper-parameter optimization. In: *NeurIPS*, pp. 2546–2554 (2011)
4. Boehm, M., Antonov, I., Baunsgaard, S., Dokter, M., Ginhör, R., Innerebner, K., Klezin, F., Lindstaedt, S.N., Phani, A., Rath, B., Reinwald, B., Siddiqui, S., Wrede, S.B.: SystemDS: a declarative machine learning system for the end-to-end data science lifecycle. In: *CIDR* (2020)
5. Boehm, M., Dusenberry, M., Eriksson, D., Evfimievski, A.V., Manshadi, F.M., Pansare, N., Reinwald, B., Reiss, F., Sen, P., Surve, A., Tatikonda, S.: Systemml: declarative machine learning on spark. *Proc. VLDB Endow.* **9**(13), 1425–1436 (2016)
6. Caruana, R., Niculescu-Mizil, A., Crew, G., Ksikes, A.: Ensemble selection from libraries of models. In: *ICML*, vol. 69 (2004)
7. Castiello, C., Castellano, G., Fanelli, A.M.: Meta-data: Characterization of input features for meta-learning. In: *MDAI*, vol. 3558, pp. 457–468 (2005)
8. Chaudhuri, K., Monteleoni, C., Sarwate, A.D.: Differentially private empirical risk minimization. *JMLR* **12**, 1069–1109 (2011)
9. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
10. Delangue, C., et al.: Hugging face (2023). <https://huggingface.co>
11. Derakhshan, B., Mahdiraji, A.R., Rabl, T., Markl, V.: Continuous deployment of machine learning pipelines. In: *Advances in Database Technology - 22nd International Conference on Extending Database Technology, EDBT 2019, Lisbon, Portugal, March 26–29, 2019*, pp. 397–408 (2019)
12. Ding, F., Hardt, M., Miller, J., Schmidt, L.: Retiring adult: new datasets for fair machine learning. *Adv. Neural Inf. Process. Syst.* **34**, 6478–90 (2021)
13. Elluswamy, A.: Occupancy networks. <https://www.youtube.com/watch?v=jPCV4GKX9Dw> (2022)
14. Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., Smola, A.J.: Autogluon-tabular: robust and accurate automl for structured data. *CoRR abs/2003.06505* (2020)
15. Facebook: Adaptive experimentation platform (2021). <https://ax.dev/>
16. Falkner, S., Klein, A., Hutter, F.: BOHB: robust and efficient hyperparameter optimization at scale. In: *ICML*, vol. 80, pp. 1436–1445 (2018)
17. Feurer, M., Eggenberger, K., Falkner, S., Lindauer, M., Hutter, F.: Auto-sklearn 2.0: Hands-free automl via meta-learning. *JMLR* **23**(261), 1–61 (2022)
18. Feurer, M., Klein, A., Eggenberger, K., Springenberg, J.T., Blum, M., Hutter, F.: Efficient and robust automated machine learning. In: *NeurIPS*, pp. 2962–2970 (2015)
19. Gelbart, M.A., Snoek, J., Adams, R.P.: Bayesian optimization with unknown constraints. In: *UAI*, pp. 250–259 (2014)
20. Ghodsnia, P., Bowman, I.T., Nica, A.: Parallel I/O aware query optimization. In: *SIGMOD*, pp. 349–360 (2014)
21. Ghosh, D., Gupta, P., Mehrotra, S., Yus, R., Altowim, Y.: JENNER: just-in-time enrichment in query processing. *Proc. VLDB Endow.* **15**(11), 2666–2678 (2022)
22. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: *NeurIPS*, pp. 3315–3323 (2016)
23. He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: *IJCNN*, pp. 1322–1328 (2008)
24. Hilprecht, B., Hammacher, C., Reis, E., Abdelaal, M., Binnig, C.: Diffml: End-to-end differentiable ML pipelines. *CoRR abs/2207.01269* (2022)
25. Hilprecht, B., Hammacher, C., Reis, E., Abdelaal, M., Binnig, C.: Diffml: End-to-end differentiable ML pipelines. In: *DEEM/SIGMOD*, pp. 7:1–7:7 (2023)
26. Kaoudi, Z., Quiané-Ruiz, J.A., Thirumuruganathan, S., Chawla, S., Agrawal, D.: A cost-based optimizer for gradient descent optimization. In: *SIGMOD*, pp. 977–992 (2017)
27. Kelly, M., Longjohn, R., Nottingham, K.: UCI ml repository (2023). <https://archive.ics.uci.edu>
28. Knudde, N., van der Herten, J., Dhaene, T., Couckuyt, I.: Gpflowopt: A bayesian optimization library using tensorflow. *arXiv preprint arXiv:1711.03845* (2017)
29. Kumar, A., Boehm, M., Yang, J.: Data management in machine learning: Challenges, techniques, and systems. In: *SIGMOD*, pp. 1717–1722 (2017). <https://doi.org/10.1145/3035918.3054775>
30. Lévesque, J.C.: Bayesian hyperparameter optimization: overfitting, ensembles and conditional spaces (2018)
31. Li, L., Jamieson, K.G., DeSalvo, G., Rostamizadeh, A., Talwalkar, A.: Hyperband: a novel bandit-based approach to hyperparameter optimization. *J. Mach. Learn. Res.* **18**, 185:1–185:52 (2017)
32. Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: challenges, methods, and future directions. *IEEE Signal Process. Mag.* **37**(3), 50–60 (2020)
33. Li, Y., Shen, Y., Zhang, W., Jiang, J., Li, Y., Ding, B., Zhou, J., Yang, Z., Wu, W., Zhang, C., Cui, B.: Volcanoml: speeding up end-to-end automl via scalable search space decomposition. *Proc. VLDB Endow.* **14**(11), 2167–2176 (2021)
34. Lindauer, M., Feurer, M., Eggenberger, K., Biedenkapp, A., Hutter, F.: Towards assessing the impact of bayesian optimization's own hyperparameters. In: *IJCAI 2019 DSO Workshop* (2019). [arXiv:1908.06674](https://arxiv.org/abs/1908.06674)
35. Liu, S., Ram, P., Vijaykeerthy, D., Bouneffouf, D., Bramble, G., Samulowitz, H., Wang, D., Conn, A., Gray, A.G.: An ADMM based framework for automl pipeline configuration. In: *AAAI*, pp. 4892–4899 (2020)
36. Mehra, A., Mandal, M., Narang, P., Chamola, V.: Reviewnet: a fast and resource optimized network for enabling safe autonomous driving in hazy weather conditions. *IEEE Trans. Intell. Transp. Syst.* **22**(7), 4256–4266 (2021)
37. Mohr, F., Wever, M., Tornede, A., Hullermeier, E.: Predicting machine learning pipeline runtimes in the context of automated machine learning. *PAMI* (2021)
38. Nakandala, S., Zhang, Y., Kumar, A.: Cerebro: a data system for optimized deep learning model selection. *Proc. VLDB Endow.* **13**(11), 2159–2173 (2020)
39. Neutatz, F.: Constraint-Driven AutoML. <https://github.com/BigDaMa/DeclarativeAutoML> (2022)
40. Neutatz, F.: Search space (2023). <https://github.com/BigDaMa/DeclarativeAutoML/blob/main/images/treespace.pdf>
41. Neutatz, F., Biessmann, F., Abedjan, Z.: Enforcing constraints for machine learning systems via declarative feature selection: an experimental study. In: *SIGMOD*, pp. 1345–1358 (2021)
42. Nishihara, R., Moritz, P., Wang, S., Tumanov, A., Paul, W., Schleier-Smith, J., Liaw, R., Niknami, M., Jordan, M.I., Stoica, I.: Real-time machine learning: the missing pieces. In: *HotOS*, pp. 106–110 (2017)
43. Olson, R.S., Moore, J.H.: TPOT: A tree-based pipeline optimization tool for automating machine learning. In: *Automated Machine Learning-Methods, Systems, Challenges, The Springer Series on Challenges in Machine Learning*, pp. 151–160 (2019)
44. Paleyes, A., Pullin, M., Mahsereci, M., McCollum, C., Lawrence, N., González, J.: Emulation of physical processes with emukit. In: *NeurIPS* (2019)
45. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in Python. *JMLR* **12**, 2825–2830 (2011)
46. Perrone, V., Donini, M., Kenthapadi, K., Archambeau, C.: Fair bayesian optimization. *arXiv preprint arXiv:2006.05109* (2020)

47. Ré, C.: Overton: a data system for monitoring and improving machine-learned products. In: CIDR (2020)
48. van Rijn, J.N., Hutter, F.: Hyperparameter importance across datasets. In: KDD, pp. 2367–2376 (2018)
49. Schelter, S., He, Y., Khilnani, J., Stoyanovich, J.: FairPrep: promoting data to a first-class citizen in studies on fairness-enhancing interventions. In: EDBT, pp. 395–398 (2020)
50. Sculley, D., al.: Kaggle (2023). <https://www.kaggle.com>
51. Settles, B.: Active learning literature survey (2009)
52. Shafique, M., Theodorides, T., Reddy, V.J., Murmann, B.: Tinyml: current progress, research challenges, and future roadmap. In: DAC, pp. 1303–1306 (2021)
53. Shah, V., Lacanlale, J., Kumar, P., Yang, K., Kumar, A.: Towards benchmarking feature type inference for automl platforms. In: SIGMOD, pp. 1584–1596 (2021)
54. Shang, Z., Zraggen, E., Buratti, B., Kossmann, F., Eichmann, P., Chung, Y., Binnig, C., Upfal, E., Kraska, T.: Democratizing data science through interactive curation of ml pipelines. In: SIGMOD, pp. 1171–1188 (2019)
55. Sparks, E.R., Venkataraman, S., Kaftan, T., Franklin, M.J., Recht, B.: Keystoneml: optimizing pipelines for large-scale advanced analytics. In: ICDE, pp. 535–546 (2017)
56. Thornton, C., Hutter, F., Hoos, H.H., Leyton-Brown, K.: Autoweka: combined selection and hyperparameter optimization of classification algorithms. In: KDD, pp. 847–855 (2013)
57. Vanschoren, J.: Meta-learning. In: Automated Machine Learning- Methods, Systems, Challenges, The Springer Series on Challenges in Machine Learning, pp. 35–61 (2019)
58. Vanschoren, J., van Rijn, J.N., Bischl, B., Torgo, L.: OpenML: networked science in machine learning. SIGKDD Explor. **15**(2), 49–60 (2013)
59. Xin, D., Macke, S., Ma, L., Liu, J., Song, S., Parameswaran, A.: Helix: holistic optimization for accelerating iterative machine learning. PVLDB **12**(4), 446–460 (2018)
60. Yakovlev, A., Moghadam, H.F., Moharrer, A., Cai, J., Chavoshi, N., Varadarajan, V., Agrawal, S.R., Idicula, S., Karnagel, T., Jinturkar, S., et al.: Oracle automl: a fast and predictive automl pipeline. PVLDB **13**(12), 3166–3180 (2020)
61. Yang, J., He, Y., Chaudhuri, S.: Auto-pipeline: synthesize data pipelines by-target using reinforcement learning and search. Proc. VLDB Endow. **14**(11), 2563–2575 (2021)
62. Yu, Y., Qian, H., Hu, Y.: Derivative-free optimization via classification. In: AAAI, pp. 2286–2292 (2016)
63. Zhang, J.M., Harman, M., Ma, L., Liu, Y.: Machine learning testing: survey, landscapes and horizons. IEEE Trans. Softw. Eng. (2020)
64. Zhang, S., Yang, F., Zhou, D., Zeng, X.: An efficient asynchronous batch bayesian optimization approach for analog circuit synthesis. In: DAC, pp. 1–6 (2020)

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