Why Do Machine Learning Practitioners Still Use Manual Tuning? A Qualitative Study

Niklas Hasebrook[1], **Felix Morsbach**[1], **Niclas Kannengießer**[1,2], **Jörg Franke**[3], **Frank Hutter**[3], **and Ali Sunyaev**[1,2]

[1]Institute of Applied Informatics and Formal Description Methods, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]KASTEL Security Labs, Karlsruhe Institute of Technology, Karlsruhe, Germany
[3]Machine Learning Lab, University of Freiburg, Freiburg, Germany

ABSTRACT

Current advanced hyperparameter optimization (HPO) methods, such as Bayesian optimization, have high sampling efficiency and facilitate replicability. Nonetheless, machine learning (ML) practitioners (e.g., engineers, scientists) mostly apply less advanced HPO methods, which can increase resource consumption during HPO or lead to underoptimized ML models. Therefore, we suspect that practitioners choose their HPO method to achieve different goals, such as decrease practitioner effort and target audience compliance. To develop HPO methods that align with such goals, the reasons why practitioners decide for specific HPO methods must be unveiled and thoroughly understood. Because qualitative research is most suitable to uncover such reasons and find potential explanations for them, we conducted semi-structured interviews to explain why practitioners choose different HPO methods. The interviews revealed six principal practitioner goals (e.g., increasing model comprehension), and eleven key factors that impact decisions for HPO methods (e.g., available computing resources). We deepen the understanding about why practitioners decide for different HPO methods and outline recommendations for improvements of HPO methods by aligning them with practitioner goals.

Keywords: Machine Learning, Hyperparameter Optimization, Automated Machine Learning

1 INTRODUCTION

The performance of machine learning (ML) models strongly depends on appropriate hyperparameter values [1, 2, 3, 4, 5, 6, 7]. Finding appropriate hyperparameter values is, however, challenging for ML practitioners (e.g., scientists or engineers from academia or industry), for example, because of large search spaces and dependencies between hyperparameters values. Such challenges require highly experienced ML practitioners to carry out extensive searches for appropriate hyperparameter values.

There are many methods for hyperparameter optimization (HPO), such as grid search, random search, and Bayesian optimization. While all of these can find equivalent hyperparameter values for ML models, they satisfy different characteristics, such as replicability and sample efficiency, to different degrees. Replicability refers to the deterministic processing of inputs by an HPO method that produces identical outputs and is especially paramount for ML research to guarantee comparability of ML models and methods [8, 9]. For example, grid search can be easily replicated if the search space and sampling rate are given. In contrast, manual tuning is hard to replicate because decisions for manual hyperparameter tuning are often influenced by unconscious factors that can be hard to explicate, e.g., personal experiences [10] and tacit knowledge [11, 12] as well as intuition [13]. Sampling efficiency refers to the required number of trials for finding appropriate hyperparameter values for an ML model. Extant research shows that the sampling efficiency of Bayesian optimization is superior to random search [14] and grid search [15] and random search has better sampling efficiency than grid search [16]. Despite the superiority of specific HPO methods over others, ML practitioners tend to apply HPO methods with lower sampling efficiency or manual tuning [17], where sampling efficiency can hardly be evaluated.

The dominant use of HPO methods that are not replicable or have low sampling efficiency indicates that practitioners may aim to achieve goals beyond improving ML model performance and let them perceive different HPO methods as more or less beneficial to achieve these goals. However, often these goals remain unclear, which hinders the emergence of best practices that guide practitioners in their selection for HPO methods.

Existing research on HPO in ML has taken a foremost technical stance by presenting HPO methods and related improvements [e.g., 18, 19, 20]. Typical works focus on the performance of ML models (e.g., in terms of accuracy, model size, running time) and mostly show superiority of HPO methods over others in a quantitative way. Thereby, technical contributions to HPO in ML support the understanding about the functioning of individual HPO methods and implications on their performance. Due to the performance-centric research on HPO, goals of practitioners beyond performance optimization, as well as decision factors that cause practitioners to choose individual HPO methods, remain unclear. The insufficient understanding about practitioners' reasons for using specific HPO methods hinders the consideration of practitioner goals in the development of HPO methods and respective software libraries. To support the reasonable selection of HPO methods and to aid the practitioner-centric advancement of HPO methods, we answer the following research question: *Why do ML practitioners choose HPO methods with different characteristics?*

To answer this question, we carried out semi-structured interviews with ten ML experts following methodological recommendations from prior research [21, 22, 23]. Two scientists independently analyzed the transcripts of the interviews using thematic analysis [24, 25] to unveil goals targeted by practitioners in HPO, HPO methods applied to reach these goals, and decision factors that cause practitioners to perceive these methods as suitable to reach their goals. Our contributions are three-fold:

- We identified six principal goals sought by practitioners in HPO for ML.
- We deepen the understanding about why practitioners use different HPO methods to reach their goals and how eleven key decision factors can influence practitioners in their decisions for HPO methods.
- We support research on more practitioner-centric HPO methods by collating perceived benefits and drawbacks of HPO methods and presenting recommendations for the advancement of HPO methods.

2 METHODS

Qualitative research is suitable to make observations and generate potential explanations for these observations by explaining their occurrence [26]. Hence, we chose an explorative, qualitative research approach and conducted semi-structured expert interviews to identify principal goals of practitioners in applying HPO and understand why practitioners decide for specific HPO methods to achieve these goals. We carried out the interviews under consideration of methodological recommendations from prior research [21, 22, 23], for example, to be open-minded and not to bias interviewees in their responses.

We individually emailed 80 potential interviewees, who we considered to be ML experts with sufficient experience. We carefully selected these interviewees among contacts from ongoing research projects, authors of scientific studies for which an ML model has been implemented, and company contacts. Among the 80 contacted ML experts, ten agreed to be interviewed. These ML experts came from six organizations, situated in five sectors, and had an average work experience in ML of about five years. For more detailed information about the interviewees, please see Appendix A.

In preparation for the interviews, we created an interview guide, including a motivation for our study, an introduction to HPO, and questions to be answered during the interview. We sent this guide to each interviewee prior to their interview and also asked the interviewees to choose an ML project they were familiar with. The selection of familiar ML projects was important to us in order to contextualize the interviewees' statements and, in case of ambiguity, to understand their actual meaning.

The interviews took between 22 and 61 minutes, with an average time of 35 minutes. In the interviews, we asked the ML experts how they selected hyperparameters for optimization and how they tuned these hyperparameters. Furthermore, we asked about decision factors that impacted their selections of HPO methods. We recorded each interview and took notes on statements that seemed particularly important to our study. We carefully transcribed each interview by hand to ensure high quality of the transcripts. Each transcript was prove-read by another co-author, who again listened to the respective recording of the interview and also considered notes taken during the interview. The final transcripts comprised a total count of 40,509 words.

Two co-authors independently analyzed the transcripts using thematic analysis [25, 24] to identify HPO methods used by the interviewees, their goals targeted through HPO, and decision factors that form a basis for decisions for specific HPO methods. Thematic analysis comprises six phases: (1) familiarize yourself with the data, (2) generate initial codes, (3) search for themes, (4) review themes, (5) define and name themes, and (6) produce the report. After familiarizing ourselves with our transcripts and notes (1), we coded the transcripts (2) to identify HPO methods applied by practitioners, to extract practitioners goals targeted in HPO (e.g., ensure comparability, increase model performance), and decision factors that impacted the interviewees' decisions for HPO methods. We incorporated decision factors to better understand contextual factors that influenced practitioner decisions for HPO methods in the sense of Gigerenzer and Brighton [27].

Two authors independently read the transcripts, copied quotes relevant for our study in individual rows of an Excel sheet, and labeled the quote with a name, a so-called code, that expresses a potentially relevant HPO method, goal, or decision factor. After the first coding iteration, we had collated 241 preliminary codes related to HPO methods, goals, and decision factors. We discussed the preliminary codes in a group of three scientists to clarify the intended meaning of each code. Based on the agreed understanding about the codes, we first harmonized the codes so that no different codes had the same semantics and, second, formulated a detailed description of each code. Subsequently, we checked the preliminary codes and descriptions for mutual exclusiveness, merged codes with overlapping meaning but different labels, and adjusted corresponding descriptions. For example, we merged the decision factors *knowledge about Bayesian optimization* and *knowledge about grid search* into the decision factor *HPO method comprehension*. After these refinements, our set of codes included six goals (e.g., decrease practitioner effort), four HPO methods (e.g., manual tuning), and eleven decision factors (e.g., lab routines) that are grouped into three higher level themes (e.g., social environment).

During the coding of HPO methods, we identified

Bayesian optimization, grid search, and random search as HPO methods. In addition, we identified actions applied by practitioners when tuning hyperparameters without formalized approach, e.g., *choose a learning rate on a logarithmic scale between 0.001 and 0.1*, *select Adam optimizer*, and *use symmetric encoder decoder architecture*. We recognized that HPO methods incorporating such actions could not be fully explicated by the interviewees to replicate their methods. The interviewees explained the difficulties in explicating these manual HPO methods by various situational factors and the importance of intuition in decision-making. The interviewees felt that these factors and intuition are hard to explicate retroactively, which is coherent with research in psychology [13, 10]. Nevertheless, to present our results about HPO methods that to a large extent rely on intuition, experience, and unconscious situational cues, we refer to the class of not (yet) explicated HPO methods as *manual tuning*. In summary, we grouped 152 codes into 40 preliminary themes (e.g., Bayesian optimization, satisfy requirements, cost of objective function) that are associated with HPO methods, goals, and decision factors.

Next, we searched for preliminary themes (3) into which we grouped the identified codes. To this end, we iteratively formulated themes based on common characteristics of codes searched and differences between them. If a code did not suite an existing theme, we created a new theme. For example, we assigned the decision factor *available computational resources* to the theme *technical environment*, while we created a new theme *own knowledge* for the decision factor *HPO method familiarity*.

For the review of the developed themes (4), we discussed the preliminary themes. We identified minor inconsistencies in the descriptions of the preliminary themes and refined the descriptions of the preliminary themes to offer a set of distinct groups associated with HPO methods, goals, and decision factors. Subsequently, we developed an intuitive name for each theme and a precise description (5). Finally, we collated the final set of 13 themes into three categories: HPO methods (four themes), principal goals (six themes), and decision factors (three themes) in this work (6).

3 RESULTS

The interviewees applied four HPO methods (i.e., manual tuning, grid search, random search, and Bayesian optimization; see Appendix B) to achieve six goals under consideration of eleven decision factors. In the following, we first briefly explain the HPO methods used. Second, we explain the identified decision factors that influence practitioners in their choice of HPO methods. Third, considering the decision factors, we describe the goals (see Table 1) sought by the practitioners, describe which methods have been used to reach these goals (see Table 2), and explain why the respondents decided for which HPO methods.

3.1 Used HPO Methods

Manual Tuning Manual tuning refers to a set of HPO methods, where a practitioner decides hyperparameter configurations based on personal knowledge (i.e., explicit and implicit), and external influences (e.g., results from literature). The dependence of manual HPO methods on individual practitioner experience and even unconscious rationals for decisions make such methods very individual to practitioners, make the explication of applied methods difficult, and, thus, decreases replicability of such methods [8]. Usually, only intermediate data (e.g., used hyperparameter values) can be used to replicate manual tuning, while reasons for the selection of these values remain unclear. Because formalization of HPO methods within manual tuning is difficult, the number of HPO methods applied by practitioners is unknown. In addition, the difficulty in explicating HPO methods used in manual tuning makes it difficult to evaluate their individual sampling efficiency. Therefore, we are not aware of any hard evidence for manual tuning to outperform even advanced methods, such as Bayesian optimization. However, many published accounts show that advanced HPO methods can outperform methods in manual tuning in certain use cases [28, 1, 4, 5, 6].

Grid Search Grid search refers to the process of evaluating the Cartesian product of a finite set of values for each hyperparameter to find appropriate values for an ML model. Every possible combination of hyperparameter values included in the defined subset of the search space is evaluated [7, 29]. To use grid search, practitioners need to decide which hyperparameters to include in the search space, as well as their respective value ranges. In contrast to manual tuning, grid search allows to replicate an experiment, because a deterministic procedure selects hyperparameter configurations to be evaluated in HPO. For replication, the originally applied search space and sampling rate must be known. The sampling efficiency of grid search tends to be lower compared to random search and Bayesian optimization [15, 30, 14].

Random Search Random search refers to the process of sampling random hyperparameter configurations from a defined search space until a specified budget for the search is exhausted [16, 7]. In preparation to use random search, practitioners define the search space for the HPO by selecting hyperparameters to be tuned and corresponding value ranges. Random search can be replicated if the used search space, the randomness generator, and the corresponding seed are known. Random search has been shown to reach better sampling efficiency in high-dimensional search spaces than grid search if some hyperparameters have a larger influence on the performance of the ML model than others [16].

Bayesian Optimization Bayesian optimization refers to the process of using a sequential approach based on a surrogate model to find appropriate hyperparameter values for an ML model in a defined search space [e.g., 31, 32, 7, 33] . The surrogate model predicts the performance of different

Table 1. Principal goals for hyperparameter optimization

Code	Description
Decrease Practitioner Effort	The state in which a practitioner applies an HPO approach for training an ML model that requires less resources compared to other HPO approaches (e.g., time for learning a new HPO method or implementing corresponding software tools).
Decrease Necessary Computations	The state where an ML model is trained with an HPO method that requires less computational resources than other methods but still is sufficiently useful for a given purposes.
Increase Model Comprehension	The state where a practitioner is able to predict changes in an ML model's behavior caused by altering hyperparameter values based on an understanding about the inner workings of the ML model.
Increase Model Performance	The state where a refined version of an ML model outperforms its original model in terms of a specified metric.
Satisfy Requirements	The state where the development and training of an ML model fulfills social and technical constraints imposed by stakeholders and the environment.
Target Audience Compliance	The state where the applied HPO and the resulting ML model fulfills the expectations of addressees.

hyperparameter configurations based on evaluated samples. To select a hyperparameter configuration to be evaluated in a subsequent iteration within HPO, an acquisition function uses these performance predictions to rank hyperparameter configurations according to their expected utility. To use Bayesian optimization, practitioners need to select hyperparameters to be tuned and their respective value ranges. Additionally, practitioners need to decide for a surrogate model, such as a Gaussian process, and an acquisition function, such as the expected improvement. HPO based on Bayesian optimization can be replicated if the search space, the acquisition function, and the surrogate model, including its hyperpriors, are known. Several studies have shown that Bayesian optimization can achieve higher sampling efficiency than grid search and random search [e.g., 15, 34, 14].

3.2 Decision Factors

Own Knowledge Practitioner decisions for HPO methods depend on their *own knowledge*, which refers to internal knowledge of a practitioner about HPO or ML models that guides the practitioner in HPO. We identified three decision factors related to own knowledge: *personal experiences*, *model comprehension*, and *HPO method comprehension*.

Personal experiences refers to the available internal knowledge that has been generated by past activities (e.g., personal best practices for solving a specific type of problem). Practitioners tend to use HPO methods with which they made positive experiences.

Model comprehension refers to the ability to predict changes in an ML model's behavior caused by altering hyperparameter values based on an understanding about the inner workings of the ML model. The perceived level of model comprehension plays an important role; practitioners that perceive their level of model comprehension as high, stated to have chosen manual tuning. They claim that based on their model comprehension, they are able to find appropriate sets of hyperparameters without the need for an extensive HPO. The interviewees perceived current HPO methods as not taking advantage of known effects of hyperparameters:

> ”*Relationships between hyperparameters are often deducible, but optimizers [here: HPO method libraries] usually do not support functionalities for this.*” (Scientist #40)

Practitioners, who perceive their own level of model comprehension as low, tend to use random search or Bayesian optimization, because they do not perceive their own knowledge as sufficient to outperform these with manual tuning.

HPO method comprehension refers to the degree to which practitioners understand HPO methods. Practitioners tend to neglect HPO methods they do not sufficiently understand. For example, two interviewees stated to have disregarded Bayesian optimization because they feel to have not sufficiently understood its inner workings. In addition, another interviewee perceived random search as uncontrolled, which caused them to decide against it. Grid search is, however, perceived as very simple, easy to understand and implement, which is the reason two interviewees gave for using it.

Social Environment The choice for an HPO method is also influenced by the social environment of ML practitioners, including five decision factors: *acceptance of proficient methods*, *lab routines*, *literature*, *shared opinions*, and *tension for resources*.

Acceptance of advanced methods refers to the extent to

Table 2. Overview of HPO methods and their use to reach practitioner goals

Goals	Manual tuning	Grid search	Random search	Bayesian optimization
Decrease Practitioner Effort	x	x		
Decrease Necessary Computations	x			
Increase Model Comprehension	x			
Increase Model Performance	x	x	x	x
Satisfy Requirements	x			
Target Audience Compliance	x	x		

which advanced HPO methods, such as Bayesian optimization, are valued by a target group. A low acceptance of advanced HPO methods in a community targeted by a practitioner can make them choose manual tuning or avoid extensive HPO entirely.

> ”*I believe it [here: HPO] is just not as valued. I believe that if you say I spent two weeks doing HPO, you will get looked at a bit strange.*” (Scientist #90)

Lab routines refer to the degree to which members of a lab always perform HPO in a similar manner because of manifested habits. The interviewees explained to have chosen HPO methods that are considered as commonly used in their labs or by their peers. In various laboratories and communities, different HPO methods are applied so frequently that their use becomes habitual. For example, manual tuning was commonly used in one research group, while Bayesian optimization was considered the primarily applied method in another one. The interviewees associated with those communities applied the respectively manifested HPO methods. This indicates that the immediate social environment has a noticeable influence on practitioner's HPO method choices.

Literature refers to the knowledge acquired on the basis of published text documents (e.g., articles, blog entries, papers). Practitioners are guided in their choice of HPO methods by recommendations from or what is considered state of the art in the literature that pertains to their ML model. All practitioners that primarily based their decisions on literature, chose Bayesian optimization, because the literature attests Bayesian optimization a high sampling efficiency [e.g. 14].

Shared opinions refers to the knowledge acquired on the basis of advice by peers (e.g., colleagues). For example, a PhD student stated in the interview that they had used Bayesian optimization because a peer had told them it was superior to random search.

Tension for shared resources refers to the degree to which limited compute resources cause conflicts between practitioners. The availability of only shared resources can cause tensions among colleagues, for example, when practitioners need to compete for compute resources to perform HPO. Such tensions caused one scientist in academia to choose manual tuning in order to avoid arguing with colleagues over computing resources.

Technical Environment Decision factors associated with the technical environment refer to technical boundaries, such as limited computational resources, that guide a practitioner in selecting a HPO method. The interviewees stated three decision factors associated with the technical environment: *available compute resources*, *cost of the objective function*, and *parallization possibilities*.

Available compute resources refer to the amount of compute resources available for HPO. Practitioners choose manual tuning when faced with limited available compute resources. They perceive that in combination with a high level of model comprehension, they can outperform other HPO methods.

Practitioners choose HPO methods depending on the *cost of the objective function* they seek to optimize (i.e., the actual training of a neural network). The cost of the objective function refers to the amount of compute resources required to evaluate a single point within the hyperparameter space. Similar to limited compute resources, the interviewees chose manual tuning when faced with expensive objective functions. If the interviewees perceive their level of model comprehension as high, they perceive manual tuning as more efficient in such situations.

Parallization possibilities for HPO methods refer to the degree to which multiple independent ML models can be simultaneously evaluated. Limited parallization possibilities can, for example, be caused by software licence limitations. Two interviewees chose Bayesian optimization if parallelization of HPO was not possible. Moreover,a practitioner stated that they opted to choose Bayesian optimization if their objective function is expensive and parallization of HPO is not possible.

3.3 Goals and How Practitioners Reached Them

Decrease Practitioner Effort Practitioner effort is decreased when a practitioner applies an HPO method that comes with a smaller overhead, e.g., in terms of time for learning a new HPO method or integrating HPO methods into workflows. To decrease practitioner effort, the interviewees applied grid search and manual tuning in their reference projects. One interviewee perceived grid search to be faster to implement and easier to use compared to Bayesian optimization, because using Bayesian optimization would have required the interviewee to learn an HPO method they were not experienced with, which aligns with findings from prior research [35]. In particular, practitioners stated to have applied manual tuning on their local machines to avoid efforts related to the integration of libraries for more advanced HPO methods into cluster infrastructures.

> ”*HPO is time-consuming sometimes, because it requires some extra lines of code to wrap all your models with this HPO method, and then set up the scripts to run them on the cluster.*” (Scientist #83)

Decrease Necessary Computations In HPO, extensive searches for hyperparameter values in large search spaces can require a vast amount of compute resources. A decrease in necessary computation is achieved when an HPO method is applied that requires less compute resources than other methods but is still sufficiently useful.

If the compute resources are too limited, the exploration of a large search space is not possible. To nevertheless perform HPO, practitioners need to decrease the number of necessary computations. To decrease the number of necessary function evaluations, three scientists stated to have used manual tuning. They perceived that with a high level of model comprehension, manual tuning is superior to Bayesian optimization or random search when available compute resources are limited.

Practitioners also rely on their model comprehension for defining the search space to decrease the number of necessary computation. For example, when compute resources are limited, they decide for, or against, the inclusion of hyperparameters into the search space based on the perceived importance of the hyperparameter. Additionally, the decision on which hyperparameters to include in the search space can be supported by literature. If the literature recommended suitable default values for hyperparameters, these hyperparameters are often assigned to those default values and kept constant, while others are tuned in HPO.

Increase Model Comprehension Increase model comprehension refers to reaching the state where a practitioner is able to predict changes in an ML model's behavior caused by altering hyperparameter values based on an understanding about the inner workings of the model. To increase their model comprehension, the interviewees reported to have applied manual tuning. Their decision for manual tuning was impacted by their perceived low comprehension about their ML model. The interviewees claimed that manual tuning facilitates the improvement of their understanding about hyperparameter influences on their ML models, because they could formulate hypothesis about hyperparameter influences and evaluate them immediately. Thereby, practitioners are able to improve their model comprehension iteratively by tuning hyperparameter values, observe influences of these values on their ML models, and test their hypotheses.

Increase Model Performance Performance of an ML model is increased when a refined version of the model outperforms its original model in terms of a specified metric. To improve the model performance, practitioners choose different HPO methods.

The interviewees chose manual tuning or grid search to find good hyperparameters, for example when prototyping a novel model.

Low model comprehension makes it difficult for practitioners to predict challenges they will encounter in HPO, especially in a prototyping setting. To better react to occurrences of unforeseen challenges, practitioners choose manual tuning. For example, manual tuning can facilitate spotting and correcting mistakes when errors occur during the development of a novel model type because feedback loops are faster compared to those of advanced HPO methods:

> ”*Because we altered the standard architecture as a whole, we were not really sure what problems we will face. So that was one of the reasons to stick with manual tuning.*” (Scientist #75)

The interviewees reported to have chosen random search and Bayesian optimization to finalize or maximize the performance of ML models.

> ”*If the only concern is to find the best model possible and no one asks how I got there and I do not have a lot of time, I probably would use random search.*” (Scientist #87)

Satisfy Requirements The goal to satisfy requirements refers to the state where the development and training of an ML model fulfills social and technical constraints imposed by stakeholders (e.g., business clients, ethics commissions) and the environment (e.g., compute resources). The interviewees described that their decisions for HPO methods were influenced by the goal to fulfill such requirements. For example, one interviewee reported to prefer manual tuning to meet hard-to-formalize requirements, such as a smooth behaviour of the model output. The interviewee felt that it was easier to react to criticism of stakeholders with manual tuning. In this sense, manual tuning appears to allow for a higher degree of agility compared to other HPO methods.

Target Audience Compliance In ML research, different communities differently value various aspects of the research project. The interviewees explained to have decided for HPO methods to comply with expectations of their target audiences regarding applied HPO methods and the resulting ML model. For example, an academic stated that they perceived the use of advanced HPO methods and extensive HPO as not being valued by their community. According to the interviewee, their community encourages the use of pre-trained ML models in combination with manual fine-tuning to avoid extensive HPO. Although the interviewee perceived Bayesian optimization as more suitable, they felt discouraged by the attitude of their community and applied manual tuning instead.

In a similar vein, two academics perceived Bayesian optimization as uncommon in their research communities and felt the need to explain Bayesian optimization in scientific works on their ML model. However, explaining Bayesian optimization would have resulted in exceeding the page limit for their paper. Therefore, the interviewees explained to have decided against Bayesian optimization but used grid search to comply with their target audience. They felt that a detailed explanation of the grid search was not necessary because it was sufficiently widespread. This left more space for the actual content in their paper.

4 DISCUSSION

Our findings show that practitioner decisions for different HPO methods pertain to six primary goals, while these decisions can vary due to influences of eleven identified decision factors. While most goals are only addressed by particular HPO methods, *increase model performance* is the only goal that can be reached by all identified HPO methods. Across all the statements of the interviewees, we observed consensus in two perceptions. First, practitioners perceive manual tuning as particularly beneficial to simply get an ML model to fit training data (e.g., in the development of prototypical ML models). Second, HPO methods with high sampling efficiency (i.e., random search or Bayesian optimization) are perceived suitable to maximize the generalization performance of ML models. These observations indicate that practitioners first tend to increase their ML model comprehension. Building on that comprehension, practitioners subsequently define search spaces for HPO. To finalize ML models in their performance, practitioners appear to acknowledge that Bayesian optimization can more reliably optimize hyperparameter values in large search spaces than themselves, which confirms recent studies [e.g., 4, 5, 6].

Manual tuning is used by ML practitioners to address all identified goals (see Table 2). This broad applicability of manual tuning for reaching manifold goals may be a justification for the dominant use of manual tuning [17, 35]. According to the interviewees' perceptions, for example, the goal *increase model comprehension* could only be achieved by manual tuning. Practitioners preferred manual tuning over advanced HPO methods because advanced HPO methods hardly support practitioners in improving their model comprehension.

Based on our findings, we derived two possible improvements for advanced HPO methods:

Generation of hyperparameter influence reports In order to aid model comprehension, HPO libraries should generate reports about the importance of individual hyperparameters after, or even better, during the HPO run. For the generation of such reports, many methods are already available, such as functional ANOVA [36], ablation [37], parallel coordinates plots [20], local parameter importance [38], and partial dependence plots [39]. Including such reports into HPO libraries can facilitate leveraging the benefits of advanced HPO methods (e.g., high sampling efficiency), while still helping practitioners increase model comprehension.

Utilization of human model comprehension To increase efficiency of HPO methods, HPO methods like Bayesian optimization should allow the incorporation of comprehension of practitioners about ML models, whose hyperparameter values are to be optimized. Practitioners should be enabled to input their knowledge about behaviors of ML models into HPO libraries prior to HPO on a case-by-case basis. For example, practitioners could specify their perceived hyperparameter importance or influences between hyperparameters. Furthermore, practitioner knowledge could be directly incorporated into the search strategy of advanced HPO methods. Promising work in this direction includes various methods for integrating prior knowledge into Bayesian optimization. This can be achieved by directly specifying priors about the location of the optimum [40, 41, 42, 43], or structural priors, e.g., in the form of log-transformations of hyperparameters [44], monotonicity constraints [45], or warping of hyperparameters [46].

5 CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

Conclusions We reported on a qualitative study on the reasons why practitioners prefer different HPO methods. With this study, we make three key contributions. First, we deepened the understanding about what goals practitioners aim to achieve by HPO (e.g., ensure comparability, increase model comprehension; see Table 1). Second, we describe how practitioners decide for HPO methods depending on their targeted goals and relevant decision factors. These findings can guide research on the reasonable integration of human decision-making into AutoML tools in order to increase their spectrum of functions and efficiency. Third, by collating practitioners' reasons for using HPO methods, we support research on AutoML in developing more practitioner-centric software applications for HPO. This can increase the adoption of sophisticated HPO methods and improve the quality of ML research (e.g., by improving the replicability of experiments).

Limitations The results presented in this study can be generalized at a limited scale. We applied semi-structured interviews as a qualitative and explorative research approach. Such interviews mainly rely on the interviewees' experiences, knowledge, perceptions, and capabilities to verbalize responses to our questions. In this sense, our results may be biased by the ML experts interviewed in this study despite our efforts to reduce such biases (e.g., by not asking leading questions and asking the interviewees for clarifications of potentially misinterpreted statements). We aimed to reduce biases in the analysis of the interviews by having two scientists independently code the transcripts of the interviews and, then, discuss their codes to agree on a shared understanding. However, despite these efforts we cannot guarantee to have fully prevented our results from being biased. Moreover, our results may not be comprehensive due to the small set of ten ML experts; e.g., additional HPO methods are not considered. To increase the comprehensiveness of the findings presented in this work, additional interviews or focus group workshops should be conducted.

Future Work We deem further investigations of human decision-making in ML as a promising direction for future research. The interviewed practitioners reported actions they applied in HPO, which are agnostic to the choice of HPO method, such as choosing a set of hyperparameters to tune and defining corresponding search ranges. The interviewed practitioners applied very alike actions for choosing HPO methods, hyperparameters, and hyperparameter values with similar reasoning and stated to have achieved their goals by these applied actions. Therefore, we assume the presence of still unclear best practices for actions applied during HPO. Since the interviewees mostly stated that they did not consciously compare different HPO methods but achieved sufficient outcomes, we deem the identification of heuristics applied in their decision-making [27, 10] in HPO as of great potential to advance AutoML research. By identifying such heuristics, a better understanding about how practitioners choose HPO methods can be reached to advance AutoML software applications by automated selection of best suitable HPO methods for ML models with individual characteristics, uses, and other contextual factors (e.g., available computing power). In future research, we will build on the findings presented in this work and seek to identify such human heuristics, implement them in algorithms for AutoML, and evaluate these algorithms in comparison to the performance of human decision-making. We also derived two very promising paths for future work that would strongly increase the benefit of current HPO tools to practitioners: (1) to increase model comprehension, automatically generate hyperparameter influence reports, and (2) to exploit human model comprehension, allow the integration of human knowledge into HPO tools.

REFERENCES

[1] Gábor Melis, Chris Dyer, and Phil Blunsom. “On the State of the Art of Evaluation in Neural Language Models”. In: *International Conference on Learning Representations*. 2018.

[2] Mario Lucic et al. “Are GANs Created Equal? A Large-Scale Study”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. 2018.

[3] Peter Henderson et al. “Deep Reinforcement Learning That Matters”. In: AAAI’18/IAAI’18/EAAI’18. New Orleans, Louisiana, USA, 2018. ISBN: 978-1-57735-800-8.

[4] Yutian Chen et al. “Bayesian optimization in alphago”. In: *arXiv preprint arXiv:1812.06855* (2018).

[5] Baohe Zhang et al. “On the Importance of Hyperparameter Optimization for Model-based Reinforcement Learning”. In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. Ed. by Arindam Banerjee and Kenji Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, 13–15 Apr 2021, pp. 4015–4023.

[6] Arlind Kadra et al. “Well-tuned Simple Nets Excel on Tabular Datasets”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al. 2021.

[7] Matthias Feurer and Frank Hutter. “Hyperparameter Optimization”. In: *Automated Machine Learning: Methods, Systems, Challenges*. Ed. by Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. Cham, 2019, pp. 3–33. ISBN: 978-3-030-05318-5. DOI: 10.1007/978-3-030-05318-5_1.

[8] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. “A metric learning reality check”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 681–699.

[9] Katharina Eggensperger et al. “HPOBench: A Collection of Reproducible Multi-Fidelity Benchmark Problems for HPO”. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 2021.

[10] Amos Tversky and Daniel Kahneman. “Judgment under Uncertainty: Heuristics and Biases”. In: *Science* 185.4157 (Sept. 1974), pp. 1124–1131. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.185.4157.1124.

[11] John Seely Brown and Paul Duguid. “Organizing Knowledge”. In: *California Management Review* 40.3 (Apr. 1998), pp. 90–111. ISSN: 0008-1256, 2162-8564. DOI: 10.2307/41165945.

[12] Michael Polanyi and Amartya Sen. *The tacit dimension*. 2009. ISBN: 978-0-226-67298-4.

[13] Gerald Carl Helmstadter. *Research concepts in human behavior: Education, psychology, sociology.* 1970.

[14] Ryan Turner et al. “Bayesian Optimization Is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020”. In: *arXiv:2104.10201 [cs, stat]* (Aug. 2021). arXiv: 2104.10201 [cs, stat].

[15] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. “Practical bayesian optimization of machine learning algorithms”. In: *Advances in neural information processing systems* 25 (2012).

[16] James Bergstra and Yoshua Bengio. “Random search for hyper-parameter optimization.” In: *Journal of machine learning research* 13.2 (2012).

[17] Xavier Bouthillier and Gaël Varoquaux. *Survey of machine-learning experimental methods at NeurIPS2019 and ICLR2020*. Research Report. Inria Saclay Ile de France, Jan. 2020.

[18] Stefan Falkner, Aaron Klein, and Frank Hutter. “BOHB: Robust and efficient hyperparameter optimization at scale”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 1437–1446.

[19] Lisha Li et al. “Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization”. In: *Journal of Machine Learning Research* 18.1 (Jan. 2017), pp. 6765–6816. ISSN: 1532-4435.

[20] Daniel Golovin et al. “Google Vizier: A Service for Black-Box Optimization”. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’17. Halifax, NS, Canada, 2017, pp. 1487–1495. ISBN: 9781450348874. DOI: 10.1145/3097983.3098043.

[21] Michele J. McIntosh and Janice M. Morse. “Situating and Constructing Diversity in Semi-Structured Interviews”. In: *Global Qualitative Nursing Research* 2 (Nov. 2015), p. 233339361559767. ISSN: 2333-3936, 2333-3936. DOI: 10.1177/2333393615597674.

[22] K. Louise Barriball and Alison While. “Collecting data using a semi-structured interview: a discussion paper”. In: *Journal of Advanced Nursing* 19.2 (Feb. 1994), pp. 328–335. ISSN: 0309-2402, 1365-2648. DOI: 10.1111/j.1365-2648.1994.tb01088.x.

[23] Raymond L. Gorden. *Interviewing: strategy, techniques, and tactics*. Rev. ed. The Dorsey series in sociology. 1975. ISBN: 978-0-256-01511-9.

[24] Virginia Braun and Victoria Clarke. “Using thematic analysis in psychology”. In: *Qualitative Research in Psychology* 3.2 (Jan. 2006), pp. 77–101. DOI: 10.1191/1478088706qp063oa.

[25] Virginia Braun and Victoria Clarke. “Thematic analysis”. In: *APA Handbooks in Psychology®. APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* American Psychological Association, 2012, pp. 57–71. DOI: 10.1037/13620-004.

[26] Anselm Strauss and Juliet Corbin. *Basics of qualitative research techniques*. 1998.

[27] Gerd Gigerenzer and Henry Brighton. “Homo Heuristicus: Why Biased Minds Make Better Inferences”. In: *Topics in Cognitive Science* 1.1 (Jan. 2009), pp. 107–143. ISSN: 17568757, 17568765. DOI: 10.1111/j.1756-8765.2008.01006.x.

[28] Matthias Feurer, Matthias Klein, and Frank Hutter. *Winning the AutoML Challenge with Auto-sklearn*. 2016. URL: https://www.kdnuggets.com/2016/08/winning-aut

[29] Douglas C Montgomery. *Design and analysis of experiments*. 2017.

[30] Katharina Eggensperger et al. “Towards an empirical foundation for assessing bayesian optimization of hyperparameters”. In: *NIPS workshop on Bayesian Optimization in Theory and Practice*. Vol. 10. 3. 2013.

[31] Eric Brochu, Vlad M Cora, and Nando De Freitas. “A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning”. In: *arXiv preprint arXiv:1012.2599* (2010).

[32] Bobak Shahriari et al. “Taking the Human Out of the Loop: A Review of Bayesian Optimization”. In: *Proceedings of the IEEE* 104.1 (2016), pp. 148–175. DOI: 10.1109/JPROC.2015.2494218.

[33] Roman Garnett. *Bayesian Optimization*. in preparation. Cambridge University Press, 2022.

[34] Aaron Klein et al. “Fast bayesian optimization of machine learning hyperparameters on large datasets”. In: *Artificial Intelligence and Statistics*. PMLR. 2017, pp. 528–536.

[35] Koen van der Blom et al. “AutoML Adoption in ML Software”. In: July 2021. URL: https://openreview.net/attachment?id=D5H5Ljwv

[36] Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. “An Efficient Approach for Assessing Hyperparameter Importance”. In: *International Conference on Machine Learning*. Jan. 2014, pp. 754–762.

[37] A. Biedenkapp et al. “Efficient Parameter Importance Analysis via Ablation with Surrogates”. In: *Proceedings of the Thirty-First Conference on Artificial Intelligence*. Ed. by S.Singh and S. Markovitch. 2017, pp. 773–779.

[38] André Biedenkapp et al. “Cave: Configuration assessment, visualization and evaluation”. In: *International Conference on Learning and Intelligent Optimization*. Springer. 2018, pp. 115–130.

[39] Julia Moosbauer et al. “Explaining Hyperparameter Optimization via Partial Dependence Plots”. In: *Proceedings of the international conference on Neural Information Processing Systems (NeurIPS)*. Dec. 2021.

[40] C. Li et al. “Incorporating Expert Prior Knowledge into Experimental Design via Posterior Sampling”. In: *ArXiv* (2020). DOI: https://arxiv.org/abs/2002.11256.

[41] A. Ramachandran et al. “Incorporating expert prior in Bayesian optimisation via space warping”. In: *Knowledge-Based Systems* 195 (2020), p. 105663.

[42] A. Souza et al. “Bayesian Optimization with a Prior for the Optimum”. In: *Machine Learning and Knowledge Discovery in Databases. Research Track ECML PKDD 2021, Bilbao, Spain, September 13-17, 2021, Proceedings, Part III*. Vol. 12977. Lecture Notes in Computer Science. 2021, pp. 265–296.

[43] Anonymous. “πBO: Augmenting Acquisition Functions with User Beliefs for Bayesian Optimization”. In: *Submitted to The Tenth International Conference on Learning Representations*. under review. 2022.

[44] Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. “Sequential model-based optimization for general algorithm configuration”. In: *International conference on learning and intelligent optimization*. Springer. 2011, pp. 507–523.

[45] C. Li et al. “Accelerating Experimental Design by Incorporating Experimenter Hunches”. In: *IEEE International Conference on Data Mining*. 2018, pp. 257–266.

[46] J. Snoek et al. “Input Warping for Bayesian Optimization of Non-Stationary Functions”. In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research 2. 22–24 Jun 2014, pp. 1674–1682.

A OVERVIEW OF INTERVIEWEES

Table 3. Demographic overview of interviewees

Field	Highest Degree of Education	Years of Experience	Focus in AI
Academia	Master (7), Bachelor (1)	< 2 (1), 2-4 (2), 5-7 (4), 8-10 (1)	RL (1), CV (5), NLP (1), Genomics (1)
Engineer	Master (1), Bachelor (1)	2-4 (1), 8-10 (1)	CV (1), NLP (2)

B HYPERPARAMETER OPTIMIZATION METHODS

Table 4. Identified hyperparameter optimization methods

Code	Description
Manual Tuning	Select hyperparameter configuration perceived as productive
Grid Search	From a defined search space, select equally spaced hyperparameter configuration
Random Search	From a defined search space, sample random hyperparameter configuration
Bayesian Optimization	From a defined search space, a surrogate model iteratively predicts productive hyperparameter configuration based on previous observations

C PRINCIPAL DECISION FACTORS

Table 5. Overview of principal decision factors for practitioner decisions for HPO methods

Theme	Description
Own Knowledge	Internal knowledge of a practitioner about HPO or ML Models that guides a practitioner in HPO
Social Environment	Statements and attitudes (e.g., opinions, recommendations) of individuals or social groups (e.g., labs) that guide a practitioner in HPO
Technical Environment	Technical boundaries (e.g., limited computational resources) that guide a practitioner in HPO

Table 6. Overview of principal decision factors for practitioner decisions for HPO methods

Theme	Code	Description
Own Knowledge	HPO Method Comprehension	The self-perceived level of knowledge a practitioner has about the inner-workings of a HPO method
Own Knowledge	Model Comprehension	The self-perceived level of understanding about the inner workings of an ML model with which a practitioner is able to predict changes in the behavior of the model caused by altering hyperparameter values
Own Knowledge	Personal Experience	The available internal knowledge that has been generated by past activities (e.g., personal best practices for solving a specific type of problem)
Social Environment	Acceptance of Advanced Methods	The extent to which advanced HPO methods, such as Bayesian optimization, are valued by a target group
Social Environment	Lab Routines	The degree to which members of a lab always perform HPO in a similar manner because of manifested habits
Social Environment	Literature	The knowledge acquired on the basis of published text documents (e.g., articles, blog entries, papers)
Social Environment	Shared Opinions	The knowledge acquired on the basis of advice by peers (e.g., colleagues)
Social Environment	Tension for Resources	The degree to which limited compute resources cause conflicts between practitioners
Technical Environment	Available Compute Resources	The amount of compute resources available for HPO
Technical Environment	Cost of Objective Function	The amount of compute resources required to evaluate a single point within the hyperparameter space
Technical Environment	Parallization Possibilities	The degree to which multiple independent ML models can be simultaneously evaluated