Fix Fairness, Don’t Ruin Accuracy: Performance Aware Fairness Repair using AutoML

Giang Nguyen
gnguyen@iastate.edu
Dept. of Computer Science
Iowa State University
Ames, Iowa, USA

Sumon Biswas
sumonb@cs.cmu.edu
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA

Hridesh Rajan
hridesh@iastate.edu
Dept. of Computer Science
Iowa State University
Ames, Iowa, USA

ABSTRACT

Machine learning (ML) is increasingly being used in critical decision-making software, but incidents have raised questions about the fairness of ML predictions. To address this issue, new tools and methods are needed to mitigate bias in ML-based software. Previous studies have proposed bias mitigation algorithms that only work in specific situations and often result in a loss of accuracy. Our proposed solution is a novel approach that utilizes automated machine learning (AutoML) techniques to mitigate bias. Our approach includes two key innovations: a novel optimization function and a fairness-aware search space. By improving the default optimization function of AutoML and incorporating fairness objectives, we are able to mitigate bias with little to no loss of accuracy. Additionally, we propose a fairness-aware search space pruning method for AutoML to reduce computational cost and repair time. Our approach, built on the state-of-the-art Auto-Sklearn tool, is designed to reduce bias in real-world scenarios. In order to demonstrate the effectiveness of our approach, we evaluated our approach on four fairness problems and 16 different ML models, and our results show a significant improvement over the baseline and existing bias mitigation techniques. Our approach, Fair-AutoML, successfully repaired 60 out of 64 buggy cases, while existing bias mitigation techniques only repaired up to 44 out of 64 cases.

CCS CONCEPTS
• Software and its engineering → Search-based software engineering; • Computing methodologies → Machine learning.

KEYWORDS
Software fairness, bias mitigation, fairness-accuracy trade-off, machine learning software, automated machine learning

1 INTRODUCTION

Recent advancements in machine learning have led to remarkable success in solving complex decision-making problems such as job recommendations, hiring employees, social services, and education [2, 10, 11, 21, 22, 24, 38, 39, 51, 53, 55, 56, 63]. However, ML software can exhibit discrimination due to unfairness bugs in the models [3, 6]. These bugs can result in skewed decisions towards certain groups of people based on protected attributes such as race, age, or sex [30, 31].

To address this issue, the software engineering (SE) community has invested in developing testing and verification strategies to detect unfairness in software systems [1, 8, 30, 31, 62]. Additionally, the machine learning literature contains a wealth of research on defining different fairness criteria for ML models and mitigating bias [12, 20, 26, 37, 48, 50, 64, 65]. Various bias mitigation methods have been proposed to build fairer models. Some approaches mitigate data bias by adapting the training data [15, 16, 52]; some modify ML models during the training process to mitigate bias [19, 32, 41, 58, 60], and others aim to increase fairness by changing the outcome of predictions [1, 62, 66].

Despite these efforts, current bias mitigation techniques often come at the cost of decreased accuracy [6, 42]. Their effectiveness varies based on datasets, fairness metrics, or the choice of protected attributes [18, 25, 26, 37]. Hort et al. proposed Fairea [42], a novel approach to evaluate the effectiveness of bias mitigation techniques, which found that nearly half of the evaluated cases received poor effectiveness. Moreover, evaluations by Chen et al. also showed that in 25% of cases, bias mitigation methods reduced both ML performance and fairness [18].

Recent works [34, 42, 60] have shown that parameter tuning can successfully fix fairness bugs without sacrificing accuracy. By finding the best set of parameters, parameter tuning can minimize the error between the predicted values and the true values to reduce bias. This helps to ensure that the model is not overly simplified or too complex, which can lead to underfitting (high bias) or overfitting (low accuracy), respectively. By tuning the parameters, we can find the right balance between bias and accuracy, which leads to a model that generalizes well to different data or fairness metric. However, it is challenging to identify which parameter setting achieves the best fairness-accuracy trade-off [34].

Recent advancements in AutoML technology [28, 29, 43] have made it possible for both experts and non-experts to harness the power of machine learning. AutoML proves to be an effective option for discovering optimal parameter settings; however, currently there is a lack of focus on reducing bias within the AutoML techniques. Thus, we pose the following research questions: Is it possible...
to utilize AutoML for the purpose of reducing bias? Is AutoML effective in mitigating bias? Does AutoML outperform existing bias reduction methods? Is AutoML more adaptable than existing bias mitigation techniques?

We introduce Fair-AutoML, a novel technique that utilizes AutoML to fix fairness bugs in machine learning models. Unlike existing bias mitigation techniques, Fair-AutoML addresses their limitations by enabling efficient and fairness-aware Bayesian search to repair unfair models, making it effective for a wide range of datasets, models, and fairness metrics. The key idea behind Fair-AutoML is to use AutoML to explore as many configurations as possible in order to find the optimal fix for a buggy model. Particularly, Fair-AutoML enhances the potential of AutoML for fixing fairness bugs in two novel techniques: by generating a new optimization function that guides AutoML to fix fairness bugs without sacrificing accuracy, and by defining a new search space based on the specific input to accelerate the bug-fixing process. Together, these contributions enable Fair-AutoML to effectively fix fairness bugs across various datasets and fairness metrics. We have implemented Fair-AutoML on top of Auto-Sklearn [29], the state-of-the-art AutoML framework.

Fair-AutoML aims to effectively address the limitations of existing bias mitigation techniques by utilizing AutoML to efficiently repair unfair models across various datasets, models, and fairness metrics. We conduct an extensive evaluation of Fair-AutoML using 4 widely used datasets in the fairness literature [1, 31, 62] and 16 buggy models collected from a recent study [6]. The results demonstrate the effectiveness of our approach, as Fair-AutoML successfully repairs 60 out of 64 buggy cases, surpassing the performance of existing bias mitigation techniques which were only able to fix up to 44 out of 64 bugs in the same settings and training time.

Our main contributions are the following:

- We have proposed a novel approach to fix unfairness bugs and retain accuracy at the same time.
- We have proposed methods to generate the optimization function automatically based on an input to make AutoML fixing fairness bugs more efficiently.
- We have pruned the search space automatically based on an input to fix fairness bugs faster using AutoML.
- We have implemented our approach in a SOTA AutoML, Auto-Sklearn [29]. The artifact is available here [33].

The paper is organized as follows: §2 describes the background, §3 presents a motivation, §4 indicates the problem definition, §5 shows the Fair-AutoML approaches, §6 presents the our evaluation, §7 discusses the limitations and future directions of Fair-AutoML, §8 discusses the threats to validity of Fair-AutoML, §9 concludes, and §10 describes the artifact.

2 BACKGROUND

We begin by providing an overview of the background and related research in the field of software fairness.

2.1 Preliminaries

2.1.1 ML Software. Given an input dataset $D$ split into a training dataset $D_{\text{train}}$ and a validation dataset $D_{\text{val}}$, a ML software system can be abstractly viewing as mapping problem $M_{\text{train}}: x \rightarrow y$ from inputs $x$ to outputs $y$ by learning from $D_{\text{train}}$. ML developers aims to search for a hyperparameter configuration $\lambda^*$ and complementary components $c^*$ for model $M$ to obtain optimal fairness-accuracy on $D_{\text{val}}$. The complementary components can be ML algorithms combined with a classifier i.e., pre-processing algorithms.

2.1.2 AutoML. Given the search spaces $\Lambda$ and $C$ for hyperparameters and complementary components, AutoML aims to find $\lambda^*$ and $c^*$ to obtain the lowest value of the cost function (Equation 1):

$$M = \arg \min_{\lambda \in \Lambda, c \in C} \text{Cost}(M_{\lambda, c}, D_{\text{val}})$$ (1)

$$\lambda^*, c^* = \arg \min_{\lambda, c} \text{Loss}(M_{\lambda, c}, D_{\text{train}})$$ (2)

2.1.3 Measures. We consider a problem, where each individual in the population has a true label in $y = \{0, 1\}$. We assume a protected attribute $z = \{0, 1\}$, such as race, sex, age, where one label is privileged (denoted 0) and the other is unprivileged (denoted 1). The predictions are $\hat{y} \in \{0, 1\}$ that need to be not only accurate with respect to $y$ but also fair with respect to the protected attribute $z$.

Accuracy Measure. Accuracy is given by the ratio of the number of correct predictions by the total number of predictions.

Accuracy = (# True positive + # True negative) / # Total

Fairness Measure. We use four ways to define group fairness metrics, which are widely used in fairness literature [4, 5, 30]:

- The Disparate Impact (DI) is the proportion of the privileged group with the favorable label divided by the proportion of the unprivileged group with the favorable label [26, 64].

$$DI = \frac{Pr[\hat{y} = 1 | z = 1]}{Pr[\hat{y} = 1 | z = 0]}$$

The Statistical Parity Difference (SPD) quantifies the disparity between the favorable label’s probability for the privileged group and the favorable label’s probability for the privileged group [12].

$$SPD = Pr[\hat{y} = 1 | z = 0] - Pr[\hat{y} = 1 | z = 1]$$

The Equal Opportunity Difference (EOD) measures the disparity between the true-positive rate of the unprivileged group and the privileged group.

$$TPR_u = Pr[\hat{y} = 1 | y = 1, z = 0]; TPR_p = Pr[\hat{y} = 1 | y = 1, z = 1]$$

$$EOD = TPR_u - TPR_p$$

The Average Absolute Odds Difference (AOD) is the mean of the difference of true-positive rate and false-positive rate among the unprivileged group and privileged group [37].

$$FPR_u = Pr[\hat{y} = 1 | y = 0, z = 0]; FPR_p = Pr[\hat{y} = 1 | y = 0, z = 1]$$

$$AOD = \frac{1}{2} \times |FPR_u - FPR_p| + |TPR_u - TPR_p|$$

To use all the metrics in the same setting, DI has been plotted in the absolute value of the log scale, and SPD, EOD, AOD have been plotted in absolute value [16, 42]. Thus, the bias score of a model is measured from 0, with lower scores indicating more fairness.

2.2 Related Work

2.2.1 Bias mitigation. SE and ML researchers has developed various bias mitigation methods to increase fairness in ML software divided into three categories [30, 40]:

- **Pre-processing** approaches reduce bias by pre-processing the training data. For instance, *Fair-SMOTE* [15] addresses data bias by removing biased labels and balancing the distribution of positive
and negative examples for each sensitive attribute. **Reweighting** [48] decreases bias by assigning different weights to different groups based on the degree of favoritism of a group. **Disparate Impact Remover** [26] is a pre-processing bias mitigation technique that aims to reduce bias by editing feature values.

**In-processing** approaches reduce bias by modifying ML models during the training process. For example, **Parfait-MI** [60] present a search-based solution to balance fairness and accuracy by tuning hyperparameter to approximate the twined Pareto curves. **MAAT** [19] is an ensemble approach aimed at improving the fairness-performance trade-off in ML software. Instead of combining models with the same learning objectives as traditional ensemble methods, MAAT merges models that are optimized for different goals.

**Post-processing** approaches change the outcome of prediction to reduce bias. This technique unfavors privileged groups’ instances and favors those of unprivileged groups lying around the decision boundary. For example, **Equalized Odds** [37] reduces the value of EOD by modifying the output labels. **Fax-AI** [35] eliminates direct discrimination in machine learning models by limiting the use of certain features, thereby preventing them from serving as surrogates for protected attributes. **Reject Option Classification** [49] prioritizes instances from the privileged group over those from the unprivileged group that are situated on the decision boundary with high uncertainty.

Previous efforts have made significant progress in reducing bias; however, they come at the cost of decreased accuracy and their results can vary depending on the datasets and fairness metrics. Our proposal, **Fair-AutoML**, aims to strike a balance between accuracy and bias reduction and demonstrate generalizability across various datasets and metrics.

2.2.2 **Search space pruning.** Search space pruning involves reducing the size or complexity of the search space in optimization or machine learning tasks. Pruning techniques are employed to accelerate the optimization process of AutoML by eliminating unpromising or redundant options, thus focusing computational resources on more promising areas of the search space. For example, Feurer et al. [29] introduce Auto-Sklearn 2.0, a novel approach aimed at enhancing the performance of Auto-Sklearn. This advancement involves constraining the search space to exclusively comprise iterative algorithms, while eliminating feature preprocessing. This strategic adjustment streamlines the implementation of successive halving, as it reduces the complexity to a single fidelity type: the number of iterations. Otherwise, the incorporation of dataset subsets as an alternative fidelity would require additional consideration. Another innovative contribution comes from Cambronero et al., who introduces AMS [19]. This method capitalizes on the wealth of source code repositories to streamline the search space for AutoML. Notably, AMS harnesses the power of unspecified complementary and functionally related API components. By leveraging these components, the search space for AutoML is pruned effectively. Diverging from prior research efforts, **Fair-AutoML** distinguishes itself by leveraging data characteristics to effectively trim down the search space. Notably, existing techniques in search space pruning primarily target accuracy enhancement within AutoML. In contrast, our innovative pruning methodology within **Fair-AutoML** is uniquely directed towards repairing unfair models.

![Figure 1: Baseline fairness-accuracy trade-off [42]](image)

2.2.3 **AutoML extension.** AutoML aims to automate the process of building a high-performing ML model, but it has limitations. It can be costly, time-consuming to train, and produces complex models that are difficult to understand. To address these limitations, software engineering researchers have developed methods to enhance AutoML performance, such as AMS [13] and Manas [54]. AMS utilizes source code repositories to create a new search space for AutoML, while Manas mines hand-developed models to find a better starting point for AutoML. The goal of these methods is to improve AutoML to maximize the accuracy. Different from these methods, **Fair-AutoML**, built on top of **Auto-Sklearn** [29], is the first to focus on repairing unfair models.

3 **MOTIVATION.**

The widespread use of machine learning in software development has brought attention to the issue of fairness in ML models. Although various bias mitigation techniques have been developed to address this issue, they have limitations. These techniques suffer from a poor balance between fairness and accuracy [42], and are not applicable to a wide range of datasets, metrics, and models [25, 26, 37]. To gain a deeper understanding of these limitations, we evaluate six different bias mitigation techniques using four fairness metrics, four datasets, and six model types. The evaluation criteria are borrowed from FairA [42] and are presented in Table 1.

**FairA** is designed to assess the trade-off between fairness and accuracy of bias mitigation techniques. The methodology of **FairA** is demonstrated in Figure 1, where the fairness and accuracy of a bias mitigation technique on a dataset are displayed in a two-dimensional coordinate system. The baseline is established by connecting the fairness-accuracy points of the original model and the mitigation models on the dataset. **FairA** evaluates the performance of the mitigation technique by altering the original model predictions and replacing a random subset of the predictions with other labels. The mutation degree ranges from 10% to 100% with a step-size of 10%. The baseline classifies the fairness-accuracy trade-off of a bias mitigation technique into five regions: lose-lose trade-off (**lose**), bad trade-off (**bad**), inverted trade-off (**inv**), good trade-off (**good**), and win-win trade-off (**win**). A technique reducing both accuracy and fairness would fall into the lose-lose trade-off region. If the trade-off is worse than the baseline, it would fall into the bad trade-off region. If the trade-off is better than the baseline, it would fall into the
The rest of this work describes our approach, Fair-AutoML, that addresses the limitations of both existing bias mitigation methods and AutoML. As a demonstration, Fair-AutoML achieved good performance in 100% of the 16 buggy cases in the Adult dataset, while 75% of the mitigation cases showed a good fairness-accuracy trade-off, and the remaining 25% exhibited an improvement in accuracy without sacrificing bias reduction.

### 4 PROBLEM DEFINITION

This work aims to utilize AutoML to address issues of unfairness in ML software by finding a new set of configurations for the model that achieves optimal fairness-accuracy trade-off. Because fairness is an additional consideration beyond accuracy, the problem becomes a multi-objective optimization problem, requiring a new cost function that can optimize both fairness and accuracy simultaneously. To achieve this, we use a technique called weighted-sum scalarization (Equation 3) [23], which allows us to weigh the importance of different objectives and create a single scalar cost function:

$$ A = \sum_{i=1}^{n} c_i \cdot \beta_i $$

where, $\beta_i$ denotes the relative weight of importance of $c_i$:

$$ \sum_{i=1}^{n} \beta_i = 1 $$

In this work, we use a cost function (or objective function) that is a weighted-sum scalarization of two decision criteria: bias and accuracy. This cost function, as shown in Equation 5, assigns weights to bias and accuracy in the cost function allowing us to adjust the trade-off between the two criteria according to the specific problems:

$$ \text{Cost}(M_{\lambda, C}, D(z)) = \beta \cdot f + (1 - \beta) \cdot (1 - a) $$

We analyze the output of the buggy ML software (including bias and accuracy) to create a suitable cost function for each input. By analyzing the output, we are able to automatically estimate the weights of the cost function in order to balance fairness and accuracy for a specific problem. To the best of our knowledge, this is the first work that applies output analysis of the software to AutoML to repair unfair ML models.

However, using AutoML can be costly and time-consuming. To address this issue, we propose a novel method that automatically create new search spaces $A*$ and $C*$ based on different inputs to accelerate the bug-fixing process of AutoML. These new search spaces are smaller in size compared to the original ones, $|A| < |A|$ and $|C| < |C|$. Particularly, as shown in Equation 6, Fair-AutoML takes as input a ML model and a dataset with a protected attribute $z$, and aims to find $\lambda*$ and $c*$ in the smaller search space, in order to minimize the cost value:

$$ M = \arg\min_{\lambda \in \Lambda_{A*}, c \in C*} \text{Cost}(M_{\lambda, C}, D_{bad}(z)) $$

The technique of search space pruning in Fair-AutoML utilizes data characteristics to enhance bug-fixing efficiency. By shrinking the search spaces based on input analysis, Fair-AutoML can find better solutions more quickly. A set of predefined modifications to the ML model are pre-built and used as a new search space for new input datasets, reducing the time needed to fix buggy models. Our approach is based on previous works in AutoML [29], but updated and modified to tackle bias issues. To the best of our knowledge,
we are the first to propose a search space pruning technique for fairness-aware AutoML.

5 FAIR-AUTOML

This section describes a detailed description of key components of Fair-AutoML (Figure 2): the dynamic optimization function (steps 1-3) and the search space pruning (steps 4-13).

5.1 Dynamic Optimization for Bias Elimination

We strive to eliminate bias in unfair models by utilizing Equation 5 as the objective function and determining the optimal value of $\beta$ to minimize the cost function. In this section, we propose an approach to automatically estimate the optimal value of $\beta$ for a specific dataset and a targeted model. This method ensures efficient correction of fairness issues while maintaining high predictive accuracy.

5.1.1 Upper bound of the cost function. To estimate the optimal value of $\beta$, the first step is to determine the upper bound of the cost function. This can be done by using a "pseudo-model", which is the 100% mutation degree model [42], as shown in the Figure 1. In other words, the pseudo-model always achieves the accuracy on any binary classification problem as follows:

$$a_0 = \max(P(Y = 1), P(Y = 0))$$  \hspace{1cm} (7)

Given an input, the pseudo-model achieves an accuracy of $a_0$ and a bias value of $f_0$ on that input. We define the cost function, Cost, of the buggy ML model with accuracy $a$ and bias value $f$ on the input. As AutoML tries different hyperparameter configurations to fix the model, the values of $a$ and $f$ may change over time. The upper bound of the cost function is defined as Equations 8 and 9:

$$\text{Cost}(M_{L_c}, D(z)) < \beta * f_0 + (1 - \beta) * (1 - a_0)$$  \hspace{1cm} (8)

$$\Rightarrow \beta * f + (1 - \beta) * (1 - a) < \beta * f_0 + (1 - \beta) * (1 - a_0)$$  \hspace{1cm} (9)

The upper bound of the cost function is defined with the goal of repairing a buggy model so that its performance falls within a good/win-win trade-off region of fairness and accuracy. In other words, the accuracy of the repaired model must be better than the pseudo-model in terms of the cost function’s value. Since the pseudo-model has zero bias ($f_0 = 0$), the upper bound of the cost function is defined as follows (Equation 10):

$$\beta * f + (1 - \beta) * (1 - a) < (1 - \beta) * (1 - a_0)$$  \hspace{1cm} (10)

5.1.2 Lower bound of $\beta$. In this work, we desire to optimize the value of $\beta$ in order to minimize bias as much as possible. The cost function used by Fair-AutoML is designed to balance accuracy and fairness, and increasing $\beta$ will place more emphasis on reducing bias. However, simply setting $\beta$ to its highest possible value is not a viable option, as it may lead to low predictive accuracy and overfitting. We cannot accept models with poor predictive accuracy regardless of their low bias [36, 57]. To overcome this challenge, we aim to find the lower bound of $\beta$, which can be done based on the upper bound of the cost function. From Equation 10, we get:

$$\beta < \frac{a - a_0}{a - a_0 + f}$$  \hspace{1cm} (11)

However, if the value of $\beta$ is smaller than $\frac{a - a_0}{a - a_0 + f}$, the optimization function Cost will always meet its upper bound condition. If the value of $\beta$ always satisfies the upper bound condition of the cost function regardless of accuracy and fairness, we can obtain a better optimization function by either increasing accuracy or decreasing bias. In this case, we cannot guide AutoML to produce a lower bias. Therefore, to guide AutoML produces an output with improved fairness, we set a lower bound for $\beta$ as Equation 12:

$$\beta \geq \frac{a - a_0}{a - a_0 + f}$$  \hspace{1cm} (12)

The intuition being that our method aims to increase the chance for AutoML to achieve better fairness. However, by setting $\beta < \frac{a - a_0}{a - a_0 + f}$ and $a > a_0$ (we aim to find a model which has better accuracy than the pseudo-model), any value of bias ($f$) can satisfy upper bound condition of the cost function, which lower chance to obtain fairer models of AutoML. To increase this chance, we set $\beta \geq \frac{a - a_0}{a - a_0 + f}$ and $a > a_0$. In this case, AutoML need to find better models that has lower bias to satisfy Equation 10. In other words, this lower bound condition indirectly forces bayesian optimization to search for lower bias models.

5.1.3 $\beta$ estimation. The final step is estimating the value of $\beta$ based on its lower bound condition. Suppose that the buggy model achieves an accuracy of $a_1$ and a bias value of $f_1$ on that input. From the beginning, we have: $a = a_1$ and $f = f_1$. In that time, the lower bound of $\beta$ is $L = \frac{a_1 - a_0}{a_1 - a_0 + f_1}$, so we have:

$$\beta = L + k, k \in [0, 1 - L]$$  \hspace{1cm} (13)

We present a greedy algorithm for estimating the value of $\beta$, which is detailed in Algorithm 1. Given a dataset $D$ with a protected attribute $z$ and a buggy model $M$ (Line 1), we start by measuring the lower bound of $\beta$. Next, we run Fair-AutoML on the input under time constraint $t$ with a value of $\beta$ set to $\frac{a_1 - a_0}{a_1 - a_0 + f_1}$ (Line 2-8). As the algorithm searches, whenever Fair-AutoML finds a candidate model that meets the condition Cost $< Cost_0$ (Lines 10-12), the value of $\beta$ is slightly increased by $a$ (Line 10-12). If after N tries, Fair-AutoML cannot find a model that satisfies the condition, the final value of $\beta$ is set to $\beta = \beta - a$ for the remaining search time to prevent overfitting from an excessively high value of $\beta$ (Lines 13-15). The algorithm returns the best model found (Line 16).
We propose a solution to speed up the Bayesian optimization process when datasets that contain multiple protected attributes, such as age, sex, and race [17]. In total, we collected 231 fairness datasets. Thirdly, for each mined dataset, we use all available protected attributes. For example, when dealing with datasets that contain multiple protected attributes, such as age, sex, and race [17]. In total, we collected 231 fairness datasets. Thirdly, for each mined dataset, we use all available protected attributes. For example, when dealing with datasets that contain multiple protected attributes, such as age, sex, and race [17]. In total, we collected 231 fairness datasets. Thirdly, for each mined dataset, we use all available protected attributes.

Algorithm 1 Greedy Weight Identifier

1. Input: a dataset $D$ with protected attribute $z$, buggy model $M$ hyperparameterized by $\lambda$, the increment value $\alpha$, the searching time $t$ and the threshold $N$
2. $\beta = \frac{a_{i} - a_{0}}{a_{i} - a_{0} + f_{i}}$
3. $\text{Cost}(M_{j_{x_{c}}, D}(z)) = \beta \cdot f + (1 - \beta) \cdot (1 - a)$
4. $\text{Cost}_{0}(M_{j_{x_{c}}, D}(z)) = (1 - \beta) \cdot (1 - a_{0})$
5. count = 0
6. checker = False
7. while $t$ do
8. $\text{for} \ i \in \Lambda$
9. count = count + 1
10. if $\text{Cost}(M_{j_{x_{c}}, D}(z)) < \text{Cost}_{0}(M_{j_{x_{c}}, D}(z))$ then
11. if checker = False then
12. $\beta = \beta + \alpha$
13. count = 0
14. if count $\geq N$ and checker = False then
15. $\beta = \beta - \alpha$
16. checker = True
17. return $M_{j_{x}}$

5.2 Search Space Pruning for Efficient Bias Elimination

We propose a solution to speed up the Bayesian optimization process in Fair-AutoML by implementing search space pruning. This technique takes advantage of data characteristics to automatically reduce the size of the search space in AutoML, thus improving its efficiency. Our approach includes two phases: the offline phase and the online phase. The offline phase trains a set of inputs multiple times to gather a collection of hyperparameters and complementary components for each input, forming a pre-built search space. In the online phase, when a new input is encountered, it is matched against the inputs stored in our database to find a matching pre-built search space, which is then utilized to repair the buggy model. This approach effectively replaces the original search space of Fair-AutoML, making the Bayesian optimization process much faster. Search space pruning has already been successfully applied before [13, 28]; however, this is the first application of data characteristics to prune the search space for fairness-aware AutoML.

5.2.1 Offline Phase. This phase constructs a set of search spaces for Fair-AutoML based on different inputs. It is important to note that the input format in the offline phase must match that of the online phase, which includes a dataset with a protected attribute and a ML model. This ensures that the pre-built search spaces created in the offline phase can be effectively utilized in the online phase.

Input. In the offline phase, we collect a set of inputs to build search spaces for Fair-AutoML. The inputs are obtained as follows. Firstly, we mine machine learning datasets from OpenML, considering only the 3425 active datasets that have been verified to work properly. Secondly, to ensure that the mined datasets are relevant to the fairness problem, we only collect datasets that contain at least one of the following attributes: age, sex, race [17]. In total, we collected 231 fairness datasets. Thirdly, for each mined dataset, we use all available protected attributes. For example, when dealing with datasets that contain multiple protected attributes, such as age, sex, and race [17]. In total, we collected 231 fairness datasets. Thirdly, for each mined dataset, we use all available protected attributes.

Algorithm 2 Database Building

1. Input: a dataset $D$ with protected attribute $z$, a model $M$ with default hyperparameters $\lambda$. Running time $t$.
2. $d = \emptyset$
3. dev = 1
4. database = {}
5. $\text{space} = \emptyset$
6. count = 0
7. while count $\leq n$ do
8. count = count + 1
9. while $t$ do
10. $M_{j*} = \text{arg minCost}(M_{j}, D(z))$
11. $d = d \cup M_{j*}$
12. $\text{kBestPipelines} = \text{top_k}(d)$
13. $\text{mBestComponents} = \text{top_m}(\text{kBestPipelines})$
14. $\text{for} \ model \in \text{kBestPipelines} \ do$
15. $\text{for} \ para \in \text{model} \ do$
16. $\text{no_outliers} = \emptyset$
17. for $i \in \text{space}[\text{para}]$ do
18. if $\text{para}$ is numerical then
19. $\text{no_outliers} = \emptyset$
20. for $i \in \text{space}[\text{para}]$ do
21. if $|i - \text{space}[\text{para}]| < \text{dev} \cdot \text{sigma}(\text{space}[\text{para}])$ then
22. $\text{no_outliers} = \text{no_outliers} \cup \text{space}[\text{para}][i]$
23. $\text{space}[\text{para}] = [\text{min}(	ext{no_outliers}), \text{max}(	ext{no_outliers})]$
24. $\text{database}[\text{input}] = (\text{space}, \text{mBestComponents})$
25. return database

Database building. To build a pre-defined search space database, we use the algorithm outlined in Algorithm 2 to obtain a pre-built search space for each collected input in order to fix the buggy model. This process involves training a fairness dataset with a specific protected attribute and ML model multiple times using Fair-AutoML, collecting the top $k$ best pipelines found, and extracting parameters from these pipelines. In particular, we use Fair-AutoML to train the fairness dataset with a specific protected attribute and a ML model for $n$ iterations (Line 7-11). We then gather the top $k$ best pipelines, including a classifier and complementary components, found by Fair-AutoML according to the optimization function’s value (Line 12). This results in $k \times n$ total pipelines. From these pipelines, we extract and store the $m$ most frequently used complementary components in the database (Line 13). For each classifier parameter, we also store its value (Lines 14-16). This results in $k \times n$ values being stored for each hyperparameter. If a hyperparameter is categorical and its values are sampled from a set of different values, we store all its unique values in the database. If a hyperparameter is numerical and its values are sampled from a uniform distribution, we remove any outliers and store the range of values from the minimum to the maximum in the database (Lines 17-23). After this process, we have collected the pre-built search space for the input (Lines 24-25). We believe that two similar inputs may have similar buggy models and
We consider that the most similar dataset to the new dataset is

Fair-AutoML

search space pruning effective in fixing fairness bugs? To answer this question, we assess the performance of Auto-Sklearn, both with and without the dynamic optimization function and search space pruning, to demonstrate the impact of each proposed approach.

6.1 Experiment

6.1.1 Benchmarks. We evaluated our method using real-world fairness bugs sourced from a recent empirical study [6], with our benchmark consisting of 16 models collected from Kaggle covering five distinct types: XGBoost (XGB), Random Forest (RF), Logistic Regression (LRG), Gradient Boosting (GBC), Support Vector Machine (SVC). We use four popular datasets for our evaluation [10, 61, 62]:

The Adult Census (race) [44] comprised of 32,561 observations and 12 features that capture the financial information of individuals from the 1994 U.S. census. The objective is to predict whether an individual earns an annual income greater than 50K. The Bank Marketing (age) [45] has 41,188 data points with 20 features including information on direct marketing campaigns of a Portuguese banking institution. The classification task aims to identify whether the client will subscribe to a term deposit. The German Credit (sex) [46] has 1000 observations with 21 features containing credit information to predict good or bad credit. The Titanic (sex) [47] has 891 data points with 10 features containing individual information of Titanic passengers. The dataset is used to predict who survived the Titanic shipwreck.

6.1.2 Evaluated Learning Techniques. We examined the performance of Auto-AutoML and other supervised learning methods addressing discrimination in binary classification including all three types of bias mitigation techniques and Auto-ML techniques.

Bias mitigation methods. We investigate all three types of bias mitigation methods: pre-processing, in-processing, post-processing. We select widely-studied bias mitigation methods for each category:

- The pre-processing includes Reweighing (R) [48], Disparate Impact Remover (DIR) [26].
- The in-processing includes Parfait-ML (PML) [60].
- The post-processing includes Equalized Odds (EO) [37], FaXAI (FAX) [35], Reject Option Classification (ROC) [49].

Auto-Sklearn. We explore the efficiency of Auto-Sklearn (AS) [29] on mitigating bias in unfair model. Although, Auto-Sklearn does not seek to decrease bias, we compare its performance with Fair-AutoML to demonstrate the efficiency of our techniques in guiding Auto-ML to repair fairness bugs.

Fair-AutoML. We create 4 versions of Fair-AutoML in this evaluation representing for Fair-AutoML with different cost functions:

- T1 uses $\beta \times DI + (1 - \beta) \times (1 - accuracy)$ as a cost function.
- T2 uses $\beta \times SPD + (1 - \beta) \times (1 - accuracy)$ as a cost function.
- T3 uses $\beta \times EOD + (1 - \beta) \times (1 - accuracy)$ as a cost function.
- T4 uses $\beta \times AOD + (1 - \beta) \times (1 - accuracy)$ as a cost function.

6.1.3 Experimental Configuration. Experiments were conducted using Python 3.6 on Intel Skylake 6140 processors. Fair-AutoML leverages the capabilities of Auto-Sklearn [29], taking advantage of

Algorithm 3 Input Matching

<table>
<thead>
<tr>
<th>Line</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Input: a input dataset $D$ with the protected attribute $z$, the number of data points $p$, the number of features $f$, lower bound $L$, a buggy model $M$, and a database.</td>
</tr>
<tr>
<td>2</td>
<td>dist = {}</td>
</tr>
<tr>
<td>3</td>
<td>for $d_i$ in database do</td>
</tr>
<tr>
<td>4</td>
<td>dist[$d_i$] = $</td>
</tr>
<tr>
<td>5</td>
<td>similarDataset = min(dist, key=dist.get)</td>
</tr>
<tr>
<td>6</td>
<td>dist = {}</td>
</tr>
<tr>
<td>7</td>
<td>for $z_i$ in similarDataset do</td>
</tr>
<tr>
<td>8</td>
<td>dist[$d_i$] = $</td>
</tr>
<tr>
<td>9</td>
<td>similarAttribute = min(dist, key=dist.get)</td>
</tr>
<tr>
<td>10</td>
<td>similarModel = $M$ with default parameter</td>
</tr>
<tr>
<td>11</td>
<td>return similarDataset, similarAttribute, similarModel</td>
</tr>
</tbody>
</table>

This search space pruning is built based on the best models found by Fair-AutoML from similar inputs, making it a reliable solution for fixing buggy models.

5.1 Online Phase. In this phase, we describe the design of the experiments to evaluate the efficiency of Fair-AutoML. We pose several research questions to evaluate the performance of Fair-AutoML and existing bias mitigation techniques on different dataset/metrics to assess the adaptability of an AutoML system in fixing fairness issues.

RQ1: Is Fair-AutoML effective in fixing fairness bugs? To answer this question, we quantify the number of fairness bugs that Fair-AutoML is able to repair compared to existing methods, allowing us to assess the capability of an AutoML system in fixing fairness issues.

RQ2: Is Fair-AutoML more adaptable than existing bias mitigation techniques? The adaptability of a bias mitigation technique indicates its performance across a diverse range of datasets/metrics. So, we analyze the effectiveness of Fair-AutoML and
### Table 2: Trade-off assessment results of Fair-AutoML, Auto-Sklearn, and mitigation techniques

<table>
<thead>
<tr>
<th>Model</th>
<th>Metric</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>R</th>
<th>DR</th>
<th>PM</th>
<th>EOD</th>
<th>FAX</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult Census</td>
<td></td>
<td>DIF Di</td>
<td>0.096</td>
<td>0.111</td>
<td>0.166</td>
<td>0.095</td>
<td>0.058</td>
<td>0.337</td>
<td>inv</td>
<td>inv</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SPD Di</td>
<td>0.023</td>
<td>0.024</td>
<td>0.038</td>
<td>0.048</td>
<td>0.016</td>
<td>0.055</td>
<td>inv</td>
<td>inv</td>
<td>0.047</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EOD Di</td>
<td>0.019</td>
<td>0.014</td>
<td>0.020</td>
<td>0.008</td>
<td>0.030</td>
<td>0.041</td>
<td>inv</td>
<td>inv</td>
<td>0.041</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AOD Di</td>
<td>0.028</td>
<td>0.033</td>
<td>0.042</td>
<td>0.032</td>
<td>0.016</td>
<td>0.060</td>
<td>0.002</td>
<td>0.004</td>
<td>0.047</td>
</tr>
<tr>
<td>XGB</td>
<td></td>
<td>DIF Di</td>
<td>0.912</td>
<td>0.023</td>
<td>0.001</td>
<td>0.007</td>
<td>0.008</td>
<td>0.014</td>
<td>0.035</td>
<td>inv</td>
<td>inv</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SPD Di</td>
<td>0.035</td>
<td>0.003</td>
<td>0.019</td>
<td>0.009</td>
<td>0.019</td>
<td>0.008</td>
<td>inv</td>
<td>inv</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EOD Di</td>
<td>0.036</td>
<td>0.041</td>
<td>0.030</td>
<td>0.003</td>
<td>0.038</td>
<td>0.062</td>
<td>0.003</td>
<td>0.044</td>
<td>inv</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AOD Di</td>
<td>0.055</td>
<td>0.044</td>
<td>0.037</td>
<td>0.003</td>
<td>0.005</td>
<td>0.066</td>
<td>0.007</td>
<td>0.020</td>
<td>bad</td>
</tr>
<tr>
<td>Bank Marketing</td>
<td></td>
<td>DIF Di</td>
<td>0.010</td>
<td>0.004</td>
<td>0.001</td>
<td>0.000</td>
<td>0.018</td>
<td>0.004</td>
<td>inv</td>
<td>inv</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SPD Di</td>
<td>0.033</td>
<td>0.001</td>
<td>0.007</td>
<td>0.002</td>
<td>0.021</td>
<td>0.009</td>
<td>inv</td>
<td>inv</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EOD Di</td>
<td>0.003</td>
<td>0.039</td>
<td>0.007</td>
<td>0.022</td>
<td>0.008</td>
<td>0.006</td>
<td>0.001</td>
<td>0.004</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AOD Di</td>
<td>0.012</td>
<td>0.002</td>
<td>0.001</td>
<td>0.000</td>
<td>0.008</td>
<td>0.016</td>
<td>0.008</td>
<td>0.029</td>
<td>inv</td>
</tr>
<tr>
<td>German Credit</td>
<td></td>
<td>DIF Di</td>
<td>0.023</td>
<td>0.022</td>
<td>0.012</td>
<td>0.011</td>
<td>0.016</td>
<td>0.006</td>
<td>0.034</td>
<td>0.049</td>
<td>0.046</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SPD Di</td>
<td>0.098</td>
<td>0.348</td>
<td>0.140</td>
<td>0.092</td>
<td>0.017</td>
<td>0.005</td>
<td>0.003</td>
<td>0.009</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EOD Di</td>
<td>0.021</td>
<td>0.007</td>
<td>0.007</td>
<td>0.008</td>
<td>0.018</td>
<td>0.006</td>
<td>0.003</td>
<td>0.007</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AOD Di</td>
<td>0.043</td>
<td>0.040</td>
<td>0.046</td>
<td>0.038</td>
<td>0.054</td>
<td>0.020</td>
<td>0.041</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Titanic</td>
<td></td>
<td>DIF Di</td>
<td>0.098</td>
<td>0.130</td>
<td>0.128</td>
<td>0.014</td>
<td>0.166</td>
<td>0.021</td>
<td>0.010</td>
<td>0.179</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SPD Di</td>
<td>0.395</td>
<td>0.571</td>
<td>0.545</td>
<td>0.545</td>
<td>0.545</td>
<td>0.545</td>
<td>0.015</td>
<td>0.651</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EOD Di</td>
<td>0.274</td>
<td>0.440</td>
<td>0.445</td>
<td>0.446</td>
<td>0.446</td>
<td>0.446</td>
<td>0.446</td>
<td>0.648</td>
<td>0.038</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AOD Di</td>
<td>0.477</td>
<td>0.556</td>
<td>0.601</td>
<td>0.602</td>
<td>0.602</td>
<td>0.602</td>
<td>0.602</td>
<td>0.602</td>
<td>0.602</td>
</tr>
<tr>
<td>LRG</td>
<td></td>
<td>DIF Di</td>
<td>0.743</td>
<td>1.673</td>
<td>1.497</td>
<td>0.882</td>
<td>0.597</td>
<td>0.214</td>
<td>0.643</td>
<td>bad</td>
<td>2.557</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SPD Di</td>
<td>0.101</td>
<td>0.552</td>
<td>0.597</td>
<td>0.113</td>
<td>0.063</td>
<td>0.017</td>
<td>0.015</td>
<td>0.015</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EOD Di</td>
<td>0.467</td>
<td>0.057</td>
<td>0.021</td>
<td>0.009</td>
<td>0.171</td>
<td>0.275</td>
<td>0.623</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AOD Di</td>
<td>0.140</td>
<td>0.562</td>
<td>0.179</td>
<td>0.101</td>
<td>0.067</td>
<td>0.214</td>
<td>0.420</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Each cell shows the accuracy/bias difference between the original and repaired models. For accuracy, accuracy difference = new accuracy - old accuracy. For bias (DI, SPD, EOD, AOD), bias difference = old bias - new bias. Thus, a positive value indicates an improvement in bias/accuracy in the repaired model compared to the original and vice versa. For bias, if a method falls into either the good region (regular numbers) or the win-win region (solid numbers), the bias difference value will be provided. If it falls into any other region, the region type will be indicated. The values highlighted in blue denote the most effective bug fixing method. The data from this table is divided and analyzed in depth in Tables 3, 4, 6.

### Table 3: Proportion of Fair-AutoML, Auto-Sklearn, and mitigation techniques that fall into each mitigation region

<table>
<thead>
<tr>
<th>Method</th>
<th>Fairness Metric</th>
<th>FAX</th>
<th>AOD</th>
<th>T4</th>
<th>R</th>
<th>DR</th>
<th>PM</th>
<th>EOD</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fairness Metric</td>
<td>Dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adult Census</td>
<td>Bank Marketing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>German Credit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Titanic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 4: Proportion of Fairness metric = # buggy cases of a metric fall into a region / # buggy cases of that dataset.

The proportions in this table are determined based on the data presented in Table 2: proportion for fairness metric = # buggy cases of a metric fall into a region / # buggy cases of that dataset.

### Table 5: Proportion of Fairness metric = # buggy cases of a metric fall into a region / # buggy cases of that dataset.

The proportions in this table are determined based on the data presented in Table 2: proportion for fairness metric = # buggy cases of a metric fall into a region / # buggy cases of that dataset.

its automatic optimization of the best ML model for a given dataset. We tailored Auto-Sklearn to better fit our method in two ways: (1) its search space was restricted to the type of the faulty classifier - for example, if the faulty classifier is Random Forest, Auto-Sklearn will only optimize the hyperparameters and identify complementary components for that specific classifier. (2) The faulty model was set as the default model for Auto-Sklearn. These modifications are features of Auto-Sklearn that we utilized.
6.2 Effectiveness (RQ1)

We evaluate the effectiveness of Fair-AutoML by comparing it with Auto-Sklearn and existing bias mitigation techniques based on Fairea baseline. The comparisons are based on the following rules:

- **Rule 1**: A model is considered successfully repaired when its post-mitigation mean accuracy and fairness falls into win-win/good trade-off regions.
- **Rule 2**: A model that falls in the win-win region is always better than one falling into any other region.
- **Rule 3**: If two models are in the same trade-off region, the one with lower bias is preferred.

Our comparison rules for bug-fixing performance were established based on Fairea and our evaluations. Firstly, we define a successful bug fix as a fixed model that falls within the win-win or good trade-off regions, as these regions demonstrate improved fairness-accuracy trade-offs compared to the baseline in Fairea. Secondly, when comparing successfully fixed models in different trade-off regions (win-win versus good), we consider the win-win models to be superior as they offer improved fairness and accuracy. Lastly, for models that fall within the same trade-off region, the one with lower bias is deemed to be better, as our goal is to fix unfair models. Our evaluations then consider two aspects of the bug-fixing performance: the number of successful bug fixes and the number of times a bias mitigation method outperforms others.

### 6.2.1 Is Fair-AutoML effective in fixing fairness bugs?

The results presented in Table 4 show that Fair-AutoML was effective in resolving 60 out of 64 (94%) fairness bugs, while Auto-Sklearn only fixed 28 out of 64 (44%) and bias mitigation techniques resolved up to 44 out of 64 (69%). This indicates that Auto-Sklearn alone was not effective in reducing bias, however, our methods were successful in enhancing AutoML to repair fairness bugs. Moreover, Fair-AutoML was able to repair more cases than other bias mitigation techniques, which often resulted in lower accuracy for lower bias. This highlights the effectiveness of our approaches in guiding AutoML towards repairing models for better trade-off between fairness and accuracy compared to the Fairea baseline.

### 6.2.2 Does Fair-AutoML outperform bias reduction techniques?

Fair-AutoML demonstrated superior performance in fixing fairness bugs compared to other bias mitigation techniques. The results presented in Table 4 indicate that 63 out of 64 buggy cases were fixed by Fair-AutoML, Auto-Sklearn, or bias mitigation techniques. Among the repaired buggy cases, Fair-AutoML outperformed other techniques 19 times (30%). On the other hand, Auto-Sklearn outperformed Fair-AutoML and bias mitigation techniques only 4 times (6%), and bias mitigation techniques outperformed other techniques 10 times at most (16%). This highlights that Fair-AutoML is often more effective in improving fairness and accuracy simultaneously or reducing more bias than other bias mitigation techniques.

### 6.3 Adaptability (RQ2)

To assess the adaptability of Fair-AutoML, we measure the proportions of each evaluated tool that fall into each fairness-accuracy trade-off region in different categories: fairness metric and dataset (Table 3). To further evaluate the adaptability of Fair-AutoML, instead of using our prepared models and datasets, we used the benchmark [59] of Parfait-ML to evaluate Fair-AutoML. Particularly, we evaluate Fair-AutoML and Parfait-ML on three different ML models (Decision Tree, Logistic Regression, Random Forest) on two datasets (Adult Census and COMPAS) (Table 3 and Figure 3).

#### 6.3.1 Is Fair-AutoML more adaptable than existing bias mitigation techniques and Auto-Sklearn?

Table 3 shows Fair-AutoML demonstrates exceptional repair capabilities across various datasets and fairness metrics, with a high rate of success in fixing buggy models. For example, in the Adult Census, Bank Marketing, German Credit, and Titanic datasets, Fair-AutoML (T4) repaired 100%, 82%, 94%, and 94% of the models, respectively. Similarly, in the DI, SPD, EOD, and AOD fairness metrics, Fair-AutoML (T4) achieved repair rates of 100%, 94%, 82%, and 94%. On the other hand, bias mitigation methods often show inconsistent results. For instance, Equalized Odds repaired all buggy cases in Adult Census but none in Bank Marketing.

In fact, our methods effectively guide AutoML in hyperparameter tuning to reduce bias, leading to superior repair performance across different datasets and metrics.

#### 6.3.2 Is Fair-AutoML effective in fixing fairness bugs on other bias mitigation methods benchmark?

Based on evaluation of Parfait-ML [60], we only use accuracy and EOD as evaluation metrics for this evaluation. To make a fair comparison with Parfait-ML, we utilize the version of Fair-AutoML that incorporates EOD and accuracy as its cost function (T3). The results are displayed in Table 4.

### Table 4: Fair-AutoML (FA) vs bias mitigation methods in fixing fairness bugs

<table>
<thead>
<tr>
<th># bugs fixed</th>
<th>FA</th>
<th>AS</th>
<th>R</th>
<th>DLR</th>
<th>PML</th>
<th>EOD</th>
<th>FAX</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td># best models</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>7</td>
<td>3</td>
</tr>
</tbody>
</table>

The results in this table are derived from the data presented in Table 2. The row # bugs fixed indicates the number of cases where the technique falls into either the win-win or good trade-off region. The row # best models represents the number of instances where a bias mitigation technique outperforms all other methods.
Table 5: Accuracy and fairness achieved by Fair-AutoML and Parfait-ML on Parfait-ML’s benchmark

<table>
<thead>
<tr>
<th>Data</th>
<th>Decision Tree</th>
<th>Logistic Regression</th>
<th>Random Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T3</td>
<td>PML</td>
<td>T3</td>
</tr>
<tr>
<td>Adult</td>
<td>0.16</td>
<td>0.02</td>
<td>0.10</td>
</tr>
<tr>
<td>Compas</td>
<td>0.16</td>
<td>0.02</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Figure 3: Accuracy and fairness achieved by Fair-AutoML (green circle) and Parfait-ML (orange circle) with Decision Tree (left) and logistic regression (right) on Adult dataset (Parfait-ML’s benchmark). The blue line shows the Fair-Auto baseline and red lines define the trade-off regions.

In this work, we bring particular attention to the fairness-accuracy tradeoff while mitigating bias in ML models. Many works in the area only optimize fairness metrics by sacrificing accuracy, and do not consider the tradeoff rigorously. However, as shown by recent work [42], trivial mutation methods can also achieve fairness if accuracy is compromised in different magnitudes. Therefore, a rigorous evaluation method is necessary to demonstrate that the tradeoff is beneficial. Another limitation of existing tools is not generalizing over different ML classifiers (e.g., LRG, GBC, RF, XGB), multiple fairness metrics, and dataset characteristics. To that end, we leveraged the recent progress of AutoML in the context and achieved better tradeoff than SOTA methods. We believe that our approach

Table 6: Trade-off assessment results of Auto-Sklearn, FAv1, and FAv2

<table>
<thead>
<tr>
<th>Metric</th>
<th>Model</th>
<th>FAv1</th>
<th>FAv2</th>
<th>Model</th>
<th>FAv1</th>
<th>FAv2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loss</td>
<td>RF</td>
<td>0.058</td>
<td>0.100</td>
<td>0.009</td>
<td>0.100</td>
<td>0.009</td>
</tr>
<tr>
<td>Accuracy</td>
<td>XGB</td>
<td>0.017</td>
<td>0.017</td>
<td>0.004</td>
<td>0.017</td>
<td>0.004</td>
</tr>
<tr>
<td>EOD</td>
<td>RF</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Bias</td>
<td>RF</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>F1</td>
<td>RF</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>FAv1</td>
<td>RF</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>FAv2</td>
<td>RF</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Loss</td>
<td>XGB</td>
<td>0.017</td>
<td>0.017</td>
<td>0.004</td>
<td>0.017</td>
<td>0.004</td>
</tr>
<tr>
<td>Accuracy</td>
<td>XGB</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>EOD</td>
<td>XGB</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Bias</td>
<td>XGB</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>F1</td>
<td>XGB</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>FAv1</td>
<td>XGB</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>FAv2</td>
<td>XGB</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

The data in Table 6 is created in the same ways as Table 2. For each method in the good trade-off region and win-win region (bold number), a trade-off measurement value is given, for other regions the region type is displayed. The values in blue, orange, and black indicate the top 1, top 2, top 3 bug fixing tools, respectively. The tools successfully repair the bug. The same rules described in RQ1 is applied in this evaluation.

6.4.1 Are dynamic optimization function and search space pruning effective in fixing fairness bugs? From Table 6, our results show that the dynamic optimization function approach in Fair-AutoML helps fix buggy models more efficiently. Comparing the performance in fixing fairness bugs, FAv1 outperforms Auto-Sklearn 39 times, while Auto-Sklearn outperforms FAv2 only 7 times. The search space pruning approach in Fair-AutoML also contributes to more efficient bug fixing, as FAv2 outperforms both FAv1 and Auto-Sklearn 46 and 55 times respectively, while FAv1 and Auto-Sklearn only outperform FAv2 14 and 4 times respectively.

7 DISCUSSION

In this work, we bring particular attention to the fairness-accuracy tradeoff while mitigating bias in ML models. Many works in the area only optimize fairness metrics by sacrificing accuracy, and do not consider the tradeoff rigorously. However, as shown by recent work [42], trivial mutation methods can also achieve fairness if accuracy is compromised in different magnitudes. Therefore, a rigorous evaluation method is necessary to demonstrate that the tradeoff is beneficial. Another limitation of existing tools is not generalizing over different ML classifiers (e.g., LRG, GBC, RF, XGB), multiple fairness metrics, and dataset characteristics. To that end, we leveraged the recent progress of AutoML in the context and achieved better tradeoff than SOTA methods. We believe that our approach
is versatile and can be applied to various ML problems. Particularly, the dynamic optimization function approach remains versatile across various datasets and models. Furthermore, the search space pruning approach is refined through pre-constructed database and a matching mechanism, that capitalizes on diverse datasets stored in repositories such as OpenML or Kaggle.

We implemented Fair-AutoML on top of Auto-Sklearn to ensure its wide applicability on ML algorithms. State-of-the-art bias mitigation techniques also primarily use classic ML algorithms [6, 7, 15, 16, 19, 60] that are supported by Auto-Sklearn. These models are more suitable than the DL models since the fairness critical tasks in prior works commonly use tabular datasets. Should one desire to explore alternative model types not directly supported by Auto-Sklearn, they can adopt the general ML model adoption of Auto-Sklearn [27].

Our approach also outlines several opportunities towards leveraging AutoML and search-based software engineering to ensure fairness in new ML models that are becoming available. First, the greedy weight identifier algorithm’s performance might suffer for complex models due to computational costs (Algorithm 1). Second, search space pruning quantitatively estimates the similarity of datasets based on data characteristics. Thus, if we do not have a dataset similar enough to the input dataset, AutoML may not perform well. To address this, we plan to regularly update our database with new datasets. Lastly, constructing suitable search spaces, particularly for resource-intensive methods like deep learning, could entail significant computational expenses. Further works are needed to maximize the versatility and effectiveness of our approach over novel fairness-critical tasks. One key direction is to combine Fair-AutoML with other bias mitigation techniques, such as integrating Fair-AutoML’s model with pre-processing bias mitigation methods to enhance overall pipeline fairness. Additionally, integrating Fair-AutoML with ensemble learning could improve both performance and fairness by capturing a broader range of biases and patterns. These directions could significantly amplify the impact of this work, making Fair-AutoML a potent tool for promoting fairness and equity in machine learning across various domains.

8 THREATS TO VALIDITY

Construct Validity: The choice of evaluation metrics and existing mitigation techniques may pose a threat to our results. We mitigate this threat by employing a diverse range of metrics and mitigation methods. First, we have used accuracy and four most recent and widely-used fairness metrics to evaluate Fair-AutoML and the state-of-the-art. These metrics have been commonly applied in the software engineering community [15, 16, 19, 60]. Second, we demonstrate the superiority of Fair-AutoML over state-of-the-art methods in different categories: pre-processing, in-processing, and post-processing, which are most advanced techniques from the SE and ML communities. For evaluating fairness and applying these mitigation algorithms except Parfait-ML [60], we have used Alif360 toolkit. For evaluating Parfait-ML, we have used its original implementation. We create a baseline using the original Fairea implementation, enabling us to conduct a comprehensive comparison between our approach and existed mitigation methods. In the future, we intend to explore supplementary performance metrics and extend our analysis to incorporate additional mitigation techniques for a more comprehensive evaluation.

External Validity: To ensure an equitable comparison with cutting-edge bias mitigation techniques, we leverage a diverse array of real-world models, datasets, and evaluation scenarios. Particularly, we utilize a practical benchmark comprising 16 real-world models thoughtfully curated by prior research [6]. Then, these meticulously chosen models undergo evaluation using four extensively studied datasets in the fairness literature [10, 61, 62]. We conducted experiments under identical setups and subsequently validated our findings [6]. In addition to assessing Fair-AutoML against alternative methods within our established settings and benchmarks, we subject Fair-AutoML to evaluation using the Parfait-ML [60] benchmark, a leading-edge bias mitigation framework.

9 CONCLUSION

We present Fair-AutoML, an innovative system that enhances existing AutoML frameworks to resolve fairness bugs. The core concept of Fair-AutoML is to optimize the hyperparameters of faulty models to resolve fairness issues. This system offers two novel technical contributions: a dynamic optimization function and a search space pruning approach. The dynamic optimization function dynamically generates an optimization function based on the input, enabling AutoML to simultaneously optimize both fairness and accuracy. The search space pruning approach reduces the size of the search space based on the input, resulting in faster and more efficient bug repair. Our experiments show that Fair-AutoML outperforms Auto-Sklearn and conventional bias mitigation techniques, with a higher rate of bug repair and a better fairness-accuracy trade-off. In the future, we plan to expand the capabilities of Fair-AutoML to include deep learning problems, beyond the scope of the current study.

10 DATA AVAILABILITY

To increase transparency and encourage reproducibility, we have made our artifact publicly available. All the source code and evaluation data with detailed descriptions can be found here [33].

ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation under Grant CCF-15-18897, CNS-15-13263, CNS-21-20448, CCF-19-34884, and CCF-22-23812. Additionally, our sincere gratitude extends to the reviewers for their invaluable insights, which significantly contributed to enhancing the quality of the paper.
Fix Fairness, Don’t Ruin Accuracy: Performance Aware Fairness Repair using AutoML


