Optimizing Fairness in Machine Learning: A Hyperparameter Tuning Approach

Mohammed Abdul Nadeem
abdulnadeemms@gmail.com

***Abstract*—Machine learning models are increasingly utilized in critical areas such as finance, hiring, and criminal justice, yet they often inherit or amplify societal biases, leading to unfair outcomes. Addressing algorithmic fairness is no longer optional but essential for building trustworthy systems. This paper proposes a novel framework that integrates fairness as a primary optimization goal during the hyperparameter tuning phase of model development. Using the FLASH algorithm, a fast sequential model-based optimization technique, we demonstrate that it is possible to simultaneously optimize for predictive accuracy and fairness metrics. Our experiments across multiple real-world datasets reveal that incorporating fairness constraints during model optimization significantly reduces bias without substantially compromising performance. Furthermore, the proposed approach outperforms several established bias mitigation techniques. These findings highlight the critical role of software engineers in embedding fairness into the machine learning lifecycle and present a practical, scalable path toward more equitable AI systems.**

***Index Terms*—Algorithmic bias, fairness, optimization**

I. Introduction

The deployment of machine learning (ML) systems in critical domains such as finance, healthcare, hiring, and criminal justice has surged in recent years. These models often surpass human capabilities in processing large datasets and identifying complex patterns. Nonetheless, their growing influence has revealed an unsettling trend: the reproduction and amplification of societal biases. Instances such as sentiment analysis tools misclassifying statements tied to particular religious or sexual identity groups, or facial recognition technologies exhibiting disproportionately high error rates for darker-skinned women compared to lighter-skinned men, highlight the pressing need for fairness in ML applications. Additionally, automated hiring platforms have been observed to favor male candidates, reflecting underlying gender biases and reinforcing systemic inequalities.

The presence of bias within machine learning models extends beyond ethical considerations; it can lead to outcomes that are legally questionable and socially damaging. Discriminatory algorithmic decisions risk entrenching existing disparities, eroding trust in automated systems, and undermining the societal benefits these technologies promise. While statistical discrimination — the reliance on group characteristics in predictive modeling — is an inherent issue, unchecked biases can exacerbate harmful stereotypes, leading to unjust and inequitable results.

Addressing these concerns has established fairness as a central theme in machine learning research and practice. There is a growing consensus that fairness must be treated as an essential quality attribute, woven throughout the software development lifecycle. Software engineers, traditionally responsible for system reliability, security, and performance, must now also prioritize ethical outcomes by ensuring that their models do not propagate or intensify social inequities. Achieving fairness requires attention to data selection, algorithm design, and evaluation metrics.

In this work, we introduce a method that elevates fairness to a primary objective during the hyperparameter tuning phase of machine learning model development. Hyperparameter optimization, a standard step to improve model performance, is repurposed to balance accuracy and fairness simultaneously. Leveraging the FLASH algorithm — a fast and scalable sequential model-based optimization technique — this study demonstrates that machine learning models can be refined to achieve equitable outcomes without sacrificing predictive power. We validate our approach through experiments on real-world datasets and show that fairness-aware tuning outperforms traditional bias mitigation strategies.

The remainder of the paper is organized as follows: Section II reviews existing research on fairness in machine learning; Section III details our proposed methodology; Section IV presents the experimental design and evaluation; Section V discusses the findings; and Section VII concludes with implications and directions for future research.

II. Background and Motivation

As machine learning systems increasingly influence decisions in sensitive areas such as finance, healthcare, hiring, and criminal justice, concerns regarding algorithmic fairness have gained significant attention. While improving predictive performance has been a primary focus for many machine learning applications, ensuring fairness is equally critical to prevent perpetuating or amplifying societal biases [1], [2]. Software engineers have traditionally been tasked with ensuring the correctness, reliability, security, and efficiency of software systems. With the growing integration of machine learning components, ethical considerations, including fairness, have become a vital aspect of software quality [3], [4]. It is increasingly recognized that fairness should not be treated merely as a post-processing concern but must be integrated into the software engineering life cycle, from design to deployment [5].

Several researchers in the software engineering (SE) community have emphasized the need for ethics-aware software development. Brun and Meliou [3] propose that fairness should be treated as a first-class quality attribute, similar to performance or security. Similarly, Aydemir and Dalpiaz [4] argue for a structured roadmap to incorporate ethical considerations into standard software engineering practices. Furthermore, recent efforts have produced testing frameworks that automatically detect biases in machine learning models. For instance, Galhotra et al. introduced Fairness Testing to identify discrimination in decision-making software [6], while Udeshi et al. developed Directed Fairness Testing to systematically uncover fairness violations [7].

979-8-3315-2037-3/25/$31.00 ©2025 IEEE

However, detecting bias is only the first step. Addressing and mitigating bias requires active intervention during the model-building process. Some works, such as FairTest [8], offer platforms to discover unwarranted associations in data-driven applications, indicating that fairness concerns must be systematically integrated into data analysis and modeling workflows. Software engineers are uniquely positioned to bridge the gap between technical development and ethical standards. By leveraging software engineering principles, including hyperparameter optimization, testing, and system design, engineers can play a pivotal role in ensuring that machine learning systems are both performant and fair. Recognizing fairness as a core software requirement, rather than an optional add-on, is essential for building trustworthy and socially responsible technology.

III. Methodology

A. Hyperparameter Optimization Techniques

Hyperparameter optimization is a crucial step in building effective machine learning models. It involves finding the best combination of hyperparameters that leads to optimal model performance. Several techniques have been developed to automate this process:

1) Grid Search: Grid search is one of the simplest and most widely used hyperparameter tuning methods. It exhaustively searches through a manually specified subset of the hyperparameter space. All possible combinations of parameters are evaluated, and the best-performing configuration is selected. While comprehensive, grid search can become computationally expensive, especially when dealing with high-dimensional spaces, a phenomenon known as the "curse of dimensionality."

2) Random Search: Random search addresses the inefficiency of grid search by selecting random combinations of hyperparameters to evaluate. Although it does not guarantee exhaustive coverage, it often finds better models in less time compared to grid search, particularly when only a few hyperparameters significantly impact performance. Random search is particularly useful when the hyperparameter space is large and sparsely populated with good solutions.

3) Bayesian Optimization: Bayesian optimization offers a more intelligent search strategy by building a probabilistic model (surrogate model) of the objective function and using it to select the most promising hyperparameters to evaluate. It balances exploration (searching uncertain regions) and exploitation (refining around known good regions) to find optimal settings efficiently. Bayesian optimization has been shown to outperform random and grid search, especially for expensive-to-evaluate models.

4) Sequential Model-Based Optimization (SMBO): Sequential Model-Based Optimization (SMBO) formalizes Bayesian optimization into a systematic framework. SMBO iteratively fits a surrogate model to the observed data and uses an acquisition function to select the next set of hyperparameters to evaluate. By sequentially updating the model with new observations, SMBO progressively focuses the search on regions of the hyperparameter space that are likely to yield better performance. SMBO methods are particularly suitable for expensive optimization tasks where each model evaluation is costly.

B. Proposed Approach: FLASH for Fairness Optimization

1) Introduction to FLASH: FLASH (Fast Sequential Model-Based Optimization) is an advanced SMBO technique designed for multi-objective optimization problems. Unlike traditional optimization methods that typically focus on a single performance metric, FLASH models each objective separately using decision trees (specifically CART models) and employs a Maximum Mean acquisition function to guide the search towards promising regions. By modeling objectives independently, FLASH achieves rapid convergence and scalability even for complex optimization landscapes.

2) Why FLASH Fits Fairness Optimization: Fairness optimization in machine learning requires balancing multiple conflicting objectives — improving predictive performance while minimizing bias. Traditional hyperparameter tuning methods often optimize for accuracy alone, neglecting fairness metrics such as Equal Opportunity Difference (EOD) and Average Odds Difference (AOD). FLASH's ability to handle multiple objectives simultaneously makes it particularly well-suited for fairness-aware model development. By treating fairness metrics and predictive performance as parallel goals, FLASH enables the construction of models that are both accurate and equitable. Its efficiency further ensures that the optimization process remains practical, even when fairness constraints add complexity to the search space. **Comparison with Other Techniques:** While FLASH offers multi-objective optimization capabilities, future work should include empirical comparisons with other techniques like Bayesian optimization and random search to quantify its relative advantage in fairness-aware contexts.

IV. Experimental Setup

A. Datasets Description

The evaluation of the proposed method is conducted using three widely studied datasets in fairness research. These datasets have been extensively utilized to investigate fairness in machine learning, particularly with regard to classification tasks. The datasets selected for this study represent different domains: demographic classification, criminal recidivism prediction, and credit risk assessment.

- **Adult Census Income Dataset:** This dataset is derived from the 1994 U.S. Census data and aims to predict whether an individual's income exceeds $50,000 per year. The dataset contains demographic attributes, such as age, education, marital status, and occupation. For fairness evaluation, the protected attributes considered are *sex* (male vs. female) and *race* (white vs. non-white). In this classification task, predicting a higher income ($50K or more) is considered the favorable outcome.
- **COMPAS Recidivism Dataset:** This dataset is used for predicting recidivism risk, specifically the likelihood that an individual will reoffend after being released from prison. The dataset includes criminal history, jail, and demographic information such as age, sex, race, and prior offenses. The protected attributes for fairness evaluation are *sex* (female vs. male) and *race* (Caucasian vs. non-Caucasian). In this study, predicting non-recidivism (i.e., the individual does not reoffend) is treated as the favorable outcome.

- **German Credit Dataset:** This dataset is employed to predict whether an individual is a good or bad credit risk based on various attributes such as age, sex, credit history, and loan purpose. The protected attributes are *sex* (male vs. female) and *age* (old vs. young). In this context, a good credit rating (indicating lower risk) is considered the favorable outcome.

For consistency across datasets, each dataset is randomly split into training (70%), validation (15%), and testing (15%) subsets. Standard preprocessing steps, including normalization of continuous variables and encoding of categorical variables, are applied where necessary to prepare the data for model training and evaluation.

B. Baseline Models

To establish a baseline for model performance and fairness comparison, two standard classification algorithms are utilized. These models represent both simple and more complex methods for binary classification tasks.

- **Logistic Regression (LR):** A widely used linear model that has been a staple in fairness studies due to its simplicity and interpretability. Logistic regression is particularly useful in understanding the relationship between predictors and the outcome, making it an ideal candidate for the initial baseline evaluation.
- **Classification and Regression Trees (CART):** A non-linear model that uses decision trees to capture interactions between features. CART is a more flexible model compared to logistic regression, capable of modeling complex relationships in the data that may not be apparent in a linear model.

Both models are trained with their default hyperparameters to provide an initial performance assessment. Following this, hyperparameter optimization is applied to improve fairness and model performance, adjusting key parameters such as learning rate, regularization strength, and tree depth, depending on the model type.

C. Evaluation Metrics

To assess both the predictive performance and fairness of the models, a set of evaluation metrics is used. These metrics provide insight into how well the models are performing in terms of accuracy, fairness, and their ability to maintain equitable treatment across different groups.

- **Recall:** This metric measures the proportion of actual positive instances correctly identified by the model. It is particularly important when the goal is to minimize false negatives, ensuring that the model correctly identifies positive cases.
- **False Alarm Rate:** Also known as the false positive rate, this metric measures the proportion of negative instances incorrectly classified as positive. A lower false alarm rate indicates that the model is not over-predicting positive outcomes, which is crucial for avoiding unnecessary interventions in practical applications.
- **Average Odds Difference (AOD):** This fairness metric calculates the average of the difference in false positive rates and true positive rates between the privileged and unprivileged groups. AOD provides an indication of whether the model is treating the two groups fairly or if disparities exist in prediction outcomes.
- **Equal Opportunity Difference (EOD):** This metric specifically evaluates the fairness of the model in terms of true positive rates between the privileged and unprivileged groups. The goal is to ensure that both groups are equally likely to be correctly classified as positive outcomes.

For fairness metrics such as AOD and EOD, the goal is to minimize these values. A value of zero for either metric indicates perfect group fairness, where both privileged and unprivileged groups experience equal treatment in terms of classification outcomes. **Statistical Testing:** In future studies, statistical significance testing (e.g., paired t-tests, bootstrap confidence intervals) will be incorporated to assess the robustness of fairness improvements and rule out randomness in observed trends.

D. Baseline Bias Mitigation Methods for Comparison

To benchmark the proposed approach, several existing bias mitigation techniques are employed. These techniques are well-established in fairness research and serve as effective methods for reducing bias in machine learning models. The comparison is conducted to determine whether the proposed method offers improvements over these commonly used approaches.

- **Reweighing:** This pre-processing method adjusts the weights of training instances to promote fairness by assigning higher weights to instances from underrepresented or disadvantaged groups.
- **Optimized Pre-processing:** A technique that learns probabilistic transformations of features and labels to reduce bias before the model is trained. This method focuses on altering the input data to achieve fairer outcomes without modifying the underlying model architecture.
- **Adversarial Debiasing:** An in-processing approach that introduces an adversary during model training. The adversary tries to predict the protected attributes (e.g., sex or race), and the main model is penalized for allowing the adversary to succeed, promoting fairness by reducing reliance on sensitive features.
- **Reject Option Classification:** A post-processing method that modifies the output of the model near the decision boundary to improve fairness. This method is particularly useful when it is necessary to adjust predictions without altering the underlying model structure.

Each of these methods is applied using publicly available implementations, adhering to established best practices in fairness research. The effectiveness of these methods is compared to the proposed approach to understand their relative performance in terms of both predictive accuracy and fairness.

V. Results and Discussion

A. RQ1: Impact of Fairness Optimization on Model Accuracy

The first research question examines the impact of fairness optimization on model accuracy, specifically whether optimizing for fairness leads to a significant reduction in model performance. To answer this, recall and false alarm

H

TABLE I
TIME COST OF HYPERPARAMETER OPTIMIZATION

Dataset	Default Model Time (seconds)	Optimized Model Time (seconds)
Adult	0.56	16.33
Compas	0.15	4.34
German Credit	0.11	3.55

TABLE II
IMPACT OF FAIRNESS OPTIMIZATION ON RECALL AND FALSE ALARM RATES

Dataset	Protected Attribute	Recall (Before)	Recall (After)	False Alarm Change
Adult	Sex	0.38	0.43	+0.04
Adult	Race	0.38	0.39	+0.00
Compas	Sex	0.52	0.53	+0.03
Compas	Race	0.52	0.57	+0.11
German Credit	Sex	0.04	0.06	+0.00
German Credit	Age	0.04	0.10	+0.03

rates are compared before and after applying fairness-oriented hyperparameter optimization.

The results indicate that fairness optimization does not significantly degrade model accuracy in most cases. For the Adult and COMPAS datasets, recall increased slightly after fairness optimization, and the false alarm rates showed minimal changes. The German Credit dataset displayed some improvements in recall for specific protected attributes, although changes in false alarm rates were generally negligible. These findings suggest that fairness optimization can be applied without substantial sacrifices in model accuracy, although minor trade-offs are possible.

B. RQ2: Achieving Trade-offs Between Fairness and Performance

The second research question investigates whether it is possible to optimize both fairness and predictive performance simultaneously. A multi-objective optimization approach was employed, optimizing for recall, false alarm, Average Odds Difference (AOD), and Equal Opportunity Difference (EOD). The results before and after multi-objective optimization are summarized in Table III.

TABLE III
MULTI-OBJECTIVE OPTIMIZATION RESULTS (RECALL, FALSE ALARM, AOD, EOD)

Dataset	Metric	Before	After	Improved
Adult	Recall	0.42	0.41	Slight decrease
Adult	AOD	0.31	0.03	Yes
Adult	EOD	0.49	0.05	Yes
Adult	False Alarm	0.08	0.01	Yes
Compas	Recall	0.52	0.52	No change
Compas	AOD	0.22	0.22	No improvement
Compas	EOD	0.27	0.27	No improvement
Compas	False Alarm	0.28	0.31	Slight increase
German Credit	Recall	0.13	0.13	No change
German Credit	AOD	0.16	0.16	No change
German Credit	EOD	0.03	0.03	No change
German Credit	False Alarm	0.05	0.03	Yes

The results indicate that the Adult dataset experienced substantial improvements in fairness metrics, particularly AOD and EOD, with only a slight decrease in recall. For the COMPAS and German Credit datasets, improvements in fairness metrics were less pronounced, with some metrics remaining unchanged. This suggests that while multi-objective optimization can effectively balance fairness and performance, the success of this approach is dataset-dependent, with some datasets exhibiting greater trade-offs between fairness and performance than others. **Dataset-Specific Effects:** The observed disparity in fairness improvement across datasets (e.g., COMPAS showing no AOD/EOD change) may be due to the inherent structure or bias distribution within the dataset. A deeper investigation into data imbalance and feature influence is essential to explain these inconsistencies.

C. RQ3: Time Efficiency of Hyperparameter Optimization

The final research question examines the computational cost associated with fairness-aware hyperparameter optimization. Table I provides a comparison of the time required for training, tuning, and testing models before and after applying optimization.

Although the optimization process introduces additional computational overhead, the total time required for training and tuning remains relatively modest, with all datasets requiring under 20 seconds for the entire procedure. This demonstrates that the fairness optimization process can be efficiently implemented without substantial increases in runtime. For larger datasets, further studies will be needed to evaluate the scalability and time efficiency of the approach. **Connection to Threats to Validity:** While the FLASH algorithm demonstrated efficient optimization under time constraints and showed varied success across datasets, these results must be interpreted considering the identified threats to validity. For instance, the lack of fairness improvements in the COMPAS dataset (Table III) may relate to dataset-specific characteristics and protected attribute distributions, as highlighted in Section VI.A and VI.B.

VI. THREATS TO VALIDITY

Several potential threats to validity may influence the interpretation of the experimental results. These threats stem from limitations in dataset size, generalizability, model choice, and hyperparameter optimization. **Failure Case Insight:** In the case of the COMPAS dataset, the lack of fairness improvement may be due to tight correlations between sensitive features and outcome labels. Future work should explore causal analysis and feature influence mapping to better understand why certain datasets resist fairness optimization.

A. Dataset Size Limitations

The datasets used in this study—Adult, COMPAS, and German Credit—are relatively small, with 32,000, 10,000, and 1,000 instances, respectively. While these datasets are widely used in fairness research, they may not capture the full diversity and complexity of real-world data, particularly underrepresented subgroups within the protected attributes. Therefore, the findings may not generalize to larger, more complex datasets. Real-world data may also introduce noise, missing values, and other challenges that were not addressed in this study, requiring further validation on more comprehensive datasets.

B. Generalizability

The experiments were conducted on three datasets from specific domains—income prediction, criminal recidivism, and creditworthiness. These datasets represent a narrow scope compared to other domains like healthcare, finance, or social media, where fairness considerations may differ. Moreover, the study focused on a limited set of protected attributes, such as sex and race, which may not encompass all sensitive attributes in other settings. The results may not generalize across different domains or fairness challenges, and further research is needed to explore the applicability of the proposed methods in diverse contexts.

C. Choice of Models and Hyperparameters

This study focused on Logistic Regression (LR) and Classification and Regression Trees (CART). While these models are interpretable and widely used in fairness studies, they represent only a small subset of machine learning models. More complex models, such as ensemble methods and deep learning architectures, were not considered and may offer different fairness-performance trade-offs. Additionally, the study explored a limited set of hyperparameters, which may not capture the full range of model configurations. Broader experimentation with other models and expanded hyperparameter spaces may lead to different results.

D. Hyperparameter Optimization and Model Selection Bias

Model selection and hyperparameter tuning may inadvertently introduce bias into fairness outcomes. For example, optimizing for accuracy without considering fairness constraints could favor majority groups. The choice of fairness metrics during optimization may also affect the results. Future research should focus on fairness-aware hyperparameter optimization techniques that explicitly address the trade-offs between performance and fairness.

E. Impact of Unobserved Confounders

Unobserved confounders, such as implicit biases in data collection or societal structures, could affect both protected attributes and outcomes. These factors may not be captured by the available data and could undermine the effectiveness of fairness optimization techniques. Addressing unobserved confounders through causal inference or external domain knowledge could improve fairness optimization in future research.

VII. Conclusion and Future Work

This study investigated the integration of fairness as an optimization goal during hyperparameter tuning of machine learning models. By applying multi-objective optimization techniques, specifically the FLASH algorithm, it was demonstrated that it is possible to mitigate bias while maintaining predictive performance across several benchmark datasets. Experimental results showed that, in many cases, improvements in fairness metrics such as Average Odds Difference (AOD) and Equal Opportunity Difference (EOD) were achieved with minimal degradation in model recall and false alarm rates. Furthermore, the computational overhead introduced by fairness-aware optimization remained within acceptable bounds, highlighting the practical feasibility of the proposed approach. The findings support the view that fairness should be treated as a primary quality attribute within the software engineering life cycle. Hyperparameter optimization, a standard tool in software analytics, can be effectively leveraged to address ethical considerations alongside traditional performance goals. Future work will aim to extend this study by exploring a broader set of machine learning models, including ensemble methods and deep learning architectures. Evaluations on larger and more diverse datasets will be conducted to assess scalability and generalizability. Additionally, collaboration with domain experts will be prioritized to better understand the significance of protected attributes and the contextual nuances of fairness in different application areas. Expanding the set of fairness metrics and investigating domain-specific definitions of fairness also represent important directions for continued research.

References

[1] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, vol. 104, no. 3, pp. 671–732, 2016.

[2] C. O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing Group, 2016.

[3] Y. Brun and A. Meliou, "Software fairness," in *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2018, pp. 754–759.

[4] F. B. Aydemir and F. Dalpiaz, "A roadmap for ethics-aware software engineering," in *Proceedings of the 1st International Workshop on Ethics in Software Engineering Research and Practice*, 2018, pp. 15–21.

[5] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. Rodolfa, and R. Ghani, "Aequitas: A bias and fairness audit toolkit," *arXiv preprint arXiv:1811.05577*, 2018.

[6] S. Galhotra, Y. Brun, and A. Meliou, "Fairness testing: Testing software for discrimination," in *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, 2017, pp. 498–510.

[7] S. Udeshi, P. Arora, and S. Chattopadhyay, "Automated directed fairness testing," in *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*, 2018, pp. 98–108.

[8] F. Tramèr, V. Atlidakis, R. Geambasu, D. Hsu, J.-P. Hubaux, M. Humbert, A. Juels, and H. Lin, "Fairtest: Discovering unwarranted associations in data-driven applications," in *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2017, pp. 401–416.