Contextual Fairness-Aware Practices in ML: A Cost-Effective Empirical Evaluation

Alessandra Parziale*, Gianmario Voria*, Giammaria Giordano*,
Gemma Catolino*, Gregorio Robles†, Fabio Palomba*
*Software Engineering (SeSa) Lab, Department of Computer Science - University of Salerno, Salerno, Italy
†Universidad Rey Juan Carlos, Madrid, Spain

***Abstract*—As machine learning (ML) systems become central to critical decision-making, concerns over fairness and potential biases have increased. To address this, the software engineering (SE) field has introduced bias mitigation techniques aimed at enhancing fairness in ML models at various stages. Additionally, recent research suggests that standard ML engineering practices can also improve fairness; these practices, known as fairness-aware practices, have been cataloged across each stage of the ML development life cycle. However, fairness remains context-dependent, with different domains requiring customized solutions. Furthermore, existing specific bias mitigation methods may sometimes degrade model performance, raising ongoing discussions about the trade-offs involved.**

In this paper, we empirically investigate fairness-aware practices from two perspectives: contextual and cost-effectiveness. The contextual evaluation explores how these practices perform in various application domains, identifying areas where specific fairness adjustments are particularly effective. The cost-effectiveness evaluation considers the trade-off between fairness improvements and potential performance costs. Our findings provide insights into how context influences the effectiveness of fairness-aware practices. This research aims to guide SE practitioners in selecting practices that achieve fairness with minimal performance costs, supporting the development of ethical ML systems.

***Index Terms*—Machine Learning Fairness; Fairness-Aware Practices; Cost-Effectiveness; Empirical Software Engineering.**

I. Introduction

Machine Learning (ML) applications continue to spread in diverse contexts, becoming integral in business operations due to their impact on efficiency, decision-making, and innovation [1]–[3]. However, this trend has raised ethical concerns around *fairness*—the expectation that models make unbiased decisions [4]. Bias in training data can lead models to make unfair decisions, presenting ethical and legal risks [5], [6].

In response to these challenges, the software engineering (SE) research community, particularly within *SE for Artificial Intelligence*, has proposed multiple bias mitigation techniques, i.e., methods designed to reduce or eliminate biases in machine learning models by operating on data or algorithms. These solutions have been categorized as *pre-processing*, *in-processing*, and *post-processing* [7], based on the ML development stage in which they operate. Different research evaluated these solutions empirically [8]–[10], demonstrating their efficacy in mitigating bias. As fairness is a *context-dependent* issue, i.e., different ethical concerns arise in different contexts [11], most of the solutions proposed so far have been evaluated under specific settings, including different application domains.

Nevertheless, fairness-specific adjustments have implications for the *economic sustainability of businesses*. This dimension refers to the financial sustainability of the software over time [12]. In this regard, the application of these bias mitigation solutions may impact operating costs in terms of customer satisfaction, particularly with decreased *performances* of the resulting models [4], [13]. De Martino et al. [12] performed a benchmark study of the implications of applying bias mitigation solutions on other sustainability dimensions, such as the economic one, highlighting that applying these algorithms involves complex trade-offs, particularly between fairness and performance.

Drifting apart from specific bias mitigation methods, recent research in SE highlighted how fairness in ML can be addressed by carefully selecting common engineering practices during the development life cycle [14]. These practices have been defined as *fairness-aware practices*—common practices that have a positive impact on the fairness level of an ML model—and they have been cataloged in the six stages of an engineered ML development life cycle [15]. These start from early stages, with practices like 'Multi-objective Optimization' or 'Data Balancing' in *'Requirements Elicitation'* and *'Data Preparation'*, to practices like 'Model Outcomes Analysis' in the final stage of *'Model Maintenance & Evolution'*.

Furthermore, these practices have been evaluated through a survey with expert ML developers [16], assessing the extent to which these have a positive impact on fairness, how often they are applied, and the perceived effort to be implemented. Results show that the majority of these practices have a positive impact on fairness, according to practitioners. Still, only a few of them are frequently applied despite not requiring high effort to be implemented [16]. Additionally, these practices appear to offer a distinct advantage in managing the fairness-performance trade-off, as they can enhance both fairness and model performance without the potential performance costs sometimes associated with bias mitigation techniques [17].

Stemming from these considerations, this work presents an empirical evaluation of fairness-aware practices with two main focuses. First, we perform a *contextual* evaluation, in which we select multiple application domains to understand if specific areas require the application of specific fairness-aware practices. Second, we introduce a novel evaluation metric to

perform the trade-off analysis targeting the *cost-effectiveness* [18] of these practices. Such an evaluation is inspired by previous work in SE [19] and aims at representing the trade-off between the benefits of implementing a fairness practice in relation to its cost in terms of performance loss.

To conduct our study, we select common datasets presented and used in previous research [8], [11], [12] but spanning over different contexts of application, e.g., Finance with the German credit dataset [20] or Law with the COMPAS dataset [11]. Afterward, we select a set of fairness-aware practices based on insights by practitioners regarding their positive impact on fairness and frequency of application [16]. Finally, we perform extensive experimentation with datasets and practices evaluating fairness and performance metrics to establish the cost of mitigating bias fairness through these practices. Results show that fairness-aware practices in ML models vary in effectiveness and cost-effectiveness across domains and datasets. Domains like *Finance* showed significant fairness improvements, while *Economic* showed none. Practices like *Iterative Imputer*, *Simple Imputer*, and *Oversampling* balanced fairness and performance well, while *Mutation Testing* was less effective. These results highlight the importance of context in selecting and evaluating fairness practices.

Paper Structure. Section II reviews the research literature relevant to our study, highlighting the key distinctions that enable our work to push the state of the art forward. Section III outlines the research questions and describes the method used to address them. In Section IV, we present and analyze the study's findings and discuss the implications for both researchers and practitioners. Section V addresses the primary limitations of the study and the strategies we employed to mitigate them. Lastly, Section VI offers concluding remarks.

II. Background and Related Work

ML fairness, defined as the absence of bias against protected groups in automated decision-making systems [4], has rapidly gained importance in the *Software Engineering* field. This heightened focus is reflected in a diverse and growing body of research that explores fairness from multiple perspectives [6]–[8], [21]. As ethical controversies surrounding ML applications continue to surface [22], [23], they underscore an urgent need to prioritize fairness in ML practices across the field.

Background. Research has proposed various *bias mitigation approaches* throughout the ML pipeline, categorized as pre-, in-, and post-processing techniques. Pre-processing methods address bias in training data, with solutions like Chakraborty et al.'s [24] FAIR-SMOTE, a synthetic data augmentation method that preserves model performance, and reweighting techniques by Kamiran and Calders [25], which adjust instance weights to enhance fairness. In-processing techniques modify learning algorithms to mitigate bias during training; for instance, Zhang et al. [26] utilized an adversarial approach, while Chakraborty et al. [27] balanced fairness and performance via multi-objective optimization. Finally, post-processing methods adjust outputs after training to ensure fairness. Galhotra et al. [28] proposed THEMIS, which identifies bias using input perturbation, and Udeshi et al. [29] introduced AEQUITAS to enhance bias detection efficiency. Black-box and white-box fairness testing approaches, such as those by Aggarwal et al. [30] and Zhang et al. [31], employ adversarial sampling to detect and address biases.

Related Works. Recent research has advanced benchmark studies for ML bias mitigation techniques. Hort et al. [10] introduced FAIREA, a tool to benchmark bias mitigation methods, focusing on five pre- and in-processing algorithms and non-functional requirements. Chen et al. [32] used Fairea in a large-scale study with seven algorithms, finding that mitigation methods can reduce ML accuracy, with effectiveness varying by task, model, protected attribute, and metric set. Zhang and Sun [9] adapted ML fairness methods for multiple protected attributes, assessing six algorithms. Recently, Chen et al. [8] benchmarked fairness improvements for multiple protected attributes across eight techniques, while Hort et al. [7] proposed a new approach to enhance both fairness and accuracy. Finally, De Martino et al. [12] benchmarked bias mitigation algorithms and explored the trade-offs among social sustainability—fairness—, economic sustainability, and environmental sustainability.

Our Contribution.

In this work, we empirically evaluate fairness-aware practices and their cost in performance loss. We advance research by (1) evaluating underexplored fairness-aware practices rather than specific bias mitigation techniques, (2) performing a context-dependent and cost-effective evaluation of these solutions, and (3) providing practitioners with practical recommendations on the specific set of fairness-aware practices to apply in their context.

III. Research Design and Methods

The *goal* of this study is to evaluate the effectiveness of fairness-aware practices in mitigating bias across different contexts, with the *purpose* of assessing their impact and understanding any performance trade-offs. The study considers the *perspective* of both researchers and practitioners. Researchers are interested in the implications of these practices on performance, contributing to the broader discourse on bias mitigation in ML models. Practitioners, meanwhile, seek practical recommendations for embedding fairness-aware practices into their workflows to build fair ML systems.

A. Research Questions

Our empirical study was centered around two main research questions. First, we aimed to quantitatively verify the positive impact of fairness-aware practices on bias mitigation, complementing previous qualitative studies that relied on expert opinions [16]. This builds on prior research assessing specific bias mitigation methods [8], [12], which reported that practitioners viewed these practices as beneficial for enhancing ML model fairness. We sought to expand on these findings by conducting experiments in diverse contexts to understand the

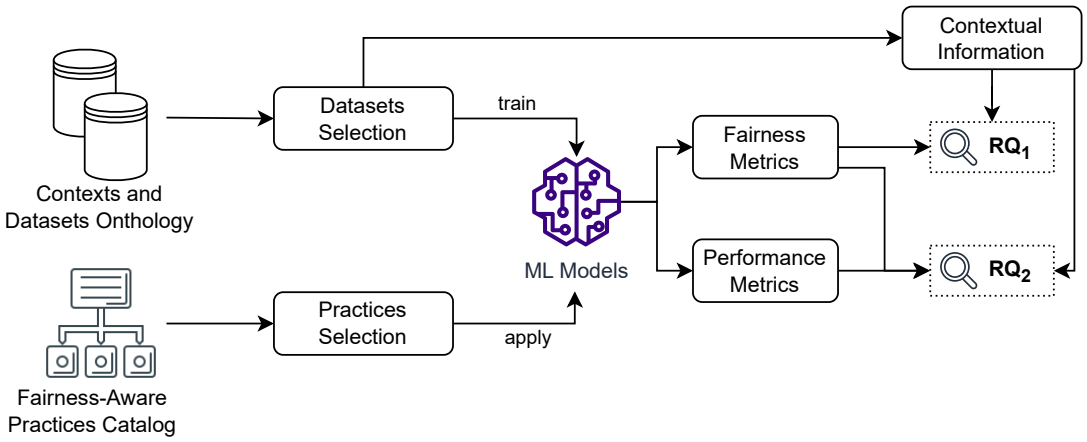

Fig. 1. Overview of the research method proposed for our study.

contextual dependency of fairness practices. This led us to our first research question:

> **RQ$_1$. Contextual Fairness Evaluation.**
>
> *To what extent can fairness-aware practices mitigate bias in different contexts?*

Our second objective was to examine the performance trade-offs associated with fairness-aware practices, as the balance between fairness and performance remains a challenge in fairness research [12]. To this aim, we designed a cost-effective analysis [18], i.e., a method for assessing the cost-effectiveness of an intervention by calculating the ratio of its cost to its effectiveness, considering the benefit as fairness gains and the cost as performance loss. This is crucial for both researchers and practitioners, as it can guide cost-effective recommendations for selecting fairness practices that balance fairness and efficiency in ML systems. This motivated our second research question:

> **RQ$_2$. Cost-Effective Evaluation.**
>
> *What is the cost in terms of performance loss against fairness improvements given by the application of fairness-aware practices?*

By exploring these two critical aspects, our study aims to advance fair ML by evaluating practical, easy-to-implement fairness-aware practices in terms of their bias mitigation efficacy across contexts and their performance cost-effectiveness. Figure 1 illustrates an overview of the research approach used to address these questions, with subsequent sections detailing the study objects and methods. Our reporting followed the guidelines of Wohlin et al. [33] and the *ACM/SIGSOFT Empirical Standards* [34],[1] specifically the *"General Standard"* guidelines given the nature of our study.

[1]Available at: https://github.com/acmsigsoft/EmpiricalStandards.

B. Practices Selection

The fairness-aware practices [14] evaluated in this study were selected based on a recent survey of ML experts [16], which assessed these practices' impact on fairness, frequency of application, and perceived implementation effort. By leveraging these findings, we identified a set of practices that offered an optimal balance of high impact, moderate to frequent application, and manageable implementation effort. These practices were deemed suitable for an in-depth, quantitative evaluation, particularly to understand their efficacy in enhancing fairness across various contexts and sensitive attributes. Below, we detail each practice selected, the reasoning behind its inclusion, and the specific implementation choices.

- Data Balancing [35]: Experts rated data balancing as a technique with a medium-to-high impact on fairness, achievable with relatively low effort [16]. The data balancing techniques selected for implementation include *LabelEncoder*, *Oversampling*, and *Undersampling* due to their popularity [36].
- Parameter Regularization [37], [38]: Despite experts identifying parameter regularization as requiring considerable effort to be implemented, it was also noted for its potentially high impact on fairness [16].
- Data Transformation [39]: This practice, frequently utilized according to expert feedback, demands a medium-to-high level of effort but has shown potential for positive impact in different contexts [16]. Techniques chosen for implementing data transformation include *IterativeImputer*, *SelectBest*, and *SimpleImputer*.
- Metamorphic/Mutation Testing [40]: This practice was selected due to its potential positive impact on fairness combined with low implementation effort [16].

This selection allowed us to explore practices' impact in diverse settings and provide recommendations for practitioners aiming to build fair ML models with minimal performance trade-offs. As for some practices, different solutions have been proposed, such as Oversampling and Undersampling for Data Balancing, and we ended up with the selection of *eight*

different techniques. The code for all the practices is available alongside the experiments in our online appendix [41].

C. Contexts and Datasets Selection

To conduct a comprehensive evaluation of fairness-aware practices across different application domains, we selected contexts from established fairness research, each represented by a widely used fairness-related dataset [11], [42]. These datasets, chosen for their relevance and prevalence in the field, provide specific contextual settings, enabling us to assess the effectiveness of bias mitigation methods within distinct contexts. Below, we detail the datasets selected, the contexts they represent, and the identified sensitive attributes.

- *COMPAS Dataset*: The COMPAS dataset is a risk assessment tool containing data from 2013 and 2014. It is used to estimate the likelihood of recidivism for defendants and is categorized under the *Recidivism Prediction context* [11]. The sensitive attributes are "Sex" and "Race", with the disadvantaged groups identified as "African-American" for "Race" and "Female" for "Sex".
- *German Credit Dataset* [20]: This dataset is used for credit risk assessment, predicting whether a loan applicant is a good or bad credit risk. It falls under the *Finance context* [11]. The sensitive attributes are "GenderStatus" and "Age", with disadvantaged groups identified as "Female-Divorced-Separated-Married" for "GenderStatus" and individuals under the age of 40 for "Age".
- *Adult Dataset* [43]: Created to analyze U.S. population characteristics—such as occupation, education, age, sex, and race—this dataset, derived from census data, aims to predict whether an individual's income exceeds $50,000 per year. It represents the *Economics context* [11]. The sensitive attributes in this dataset are "Race" and "Sex", with the disadvantaged groups being "Black" for "Race" and "Female" for "Sex".
- *Bank Marketing Dataset* [44]: This dataset comprises information on direct marketing campaigns by a Portuguese bank from 2008 to 2013, with the objective of predicting whether a client would subscribe to a bank deposit. It falls within the *Marketing context* [11]. The sensitive attributes are "Marital Status" and "Age", with disadvantaged groups being "Married" for "Marital Status" and individuals under 40 for "Age".
- *Communities and Crime Dataset* [45]: Containing socioeconomic data from 46 U.S. states, this dataset is used to predict the total number of violent crimes (including murder, rape, robbery, and assault). It represents the *Crime context* [11], with the sensitive attribute being "Race" and the disadvantaged group identified as "Black".

Each of these datasets serves as a context driver for our evaluation, enabling us to observe the impact of fairness-aware practices in settings specific to "law" or "economics and business." For instance, assessing practices on a dataset like COMPAS offers insights into how these practices operate within the law context, particularly the one of recidivism prediction. This approach allows us to provide practical insights into the contextual performance and potential trade-offs of fairness-aware methods across distinct application domains.

D. Data Collection and Analysis

After selecting fairness-aware practices and fairness-related datasets, we experimented with their combination to answer our research questions. In this preliminary empirical study, we focused on *classification* tasks, as they were the most applied in fairness research [11]. Nonetheless, it is worth noting that the selected datasets may be exploited for other tasks as well, e.g., *clustering* or *anomaly detection* [11]. Concerning the *classification* task that we implemented in both our **RQs**, we leveraged the work by Fabris et al. [11] from which we selected the contexts, as they also provided a classification of fairness-related datasets alongside the specific tasks and models that have been used to evaluate them. Hence, we finally selected the *Random Forest* model, as it was among the most commonly used for the chosen datasets. All the data and code used to perform the experiments and evaluate the results are available in our online appendix [41].

RQ_1 — Contextual Fairness Evaluation. To address our first research question, we conducted experiments using Random Forest for the *classification* task. We began by establishing a baseline: for each dataset, we trained the model without applying any fairness-aware practices and evaluated its fairness level. To assess model fairness, we used three widely recognized metrics from fairness research [46]: *Average Odds Difference*, which measures the absolute difference between the rates of correct and incorrect classifications across two groups; *False Discovery Rate Difference*, which indicates disparities in false positive rates between distinct groups; and *Disparate Impact*, defined as the ratio of positive outcomes in the protected group to those in the non-protected group. For each dataset, all the metrics were computed on one single protected attribute, selected by considering the ones that were most frequently evaluated in the literature [11]. We selected "Sex" for the COMPAS and Adult datasets, "Age" for the Bank Marketing dataset, "Gender" for the German credit dataset, and "Race" for the Communities and Crime dataset. This process resulted in an initial set of five baseline experiments.

Subsequently, we applied each selected fairness-aware practice individually to each dataset and re-trained the same ML model used for the baseline. For each new model, we recalculated the three fairness metrics, leading to 40 additional experiments. Including the baseline, in our data collection phase, we conducted 45 experiments in total.

To answer $\mathbf{RQ_1}$ and understand if the application of fairness-aware practices may increase the models' fairness level in different contexts, we analyzed such data comparing the baseline's results with the metrics computed after the application of the practices in different contexts. To verify the significance of these results, we finally applied statistical tests. We first assessed the normality of our data to determine the appropriate statistical methods. Using the Shapiro-Wilk test [47] with a significance level of $\alpha = 0.05$, we found that not all the studied datasets conformed to a normal

distribution, which led us to apply non-parametric methods. Hence, we used the Wilcoxon signed-rank test [48] to assess differences between the baselines and experiments involving fairness-aware practices, testing the null hypothesis that no significant differences exist. Given the limited sample sizes, further evaluation was not deemed necessary. The application of the Wilcoxon test allowed us to compute p-values and assess statistical significance directly.

RQ_2 — Cost-Effective Evaluation. To answer our second research question, we followed the same steps as the first one to collect the data. However, in this case, our objective was to evaluate the model's performance. Hence, we collected standard performance metrics of the trained models such as *Precision*, *Recall*, *F1-score*, and *Accuracy* [49]. By collecting these data, we ended up with a dataset of experiments composed of 45 rows: for each of the five datasets, we trained an ML model and collected four performance metrics and three fairness metrics, repeating this experiment nine times—one for the baseline and eight for each fairness-aware practice selected. After computing all these experiments and collecting the metrics, we continued with the data analysis phase.

In the context of **RQ_2**, our objective was to evaluate the performance-fairness trade-off of applying fairness-aware practices. Hence, the data analysis process slightly changed from the first research question. We applied a *cost-effectiveness analysis* [18], a technique used to evaluate the cost-effectiveness ratio of an intervention by dividing its *costs* by its *effectiveness*. In our case, we evaluated each intervention—the application of a fairness-aware practice to an ML model training—based on its *effectiveness* in enhancing fairness relative to its *cost* in terms of model performance loss. This approach enabled us to quantify the trade-offs between improved fairness and reduced predictive accuracy, helping identify the most efficient techniques for maintaining both fairness and performance.

For each experiment with a fairness-aware practice application, we calculated two measures: (1) the *cost*, computed as the difference in performance between the baseline *(B)* model—without intervention—and the models with the fairness-aware practices applied *(I)*, and (2) the *effectiveness* measured by calculating the difference in fairness metrics between the models with the fairness intervention and the baseline models.

With these two measures, we computed a *cost-effectiveness* ratio by dividing the performance cost by the effectiveness in improving fairness as follows:

$$\text{Cost-effectiveness} = \frac{\text{Performance}_B - \text{Performance}_I}{\text{Fairness}_I - \text{Fairness}_B}$$

We used this metric as a comparative metric, enabling us to identify which fairness-aware practice provided the greatest fairness improvements with the least performance compromise. A value *lower than one* indicated a more cost-effective technique, as they yielded higher fairness gains per unit of performance cost. Finally, for each dataset and fairness-aware practice, we computed the cost-effectiveness ratio for all the possible combinations of fairness and performance metrics. We then aggregated these data by averaging the cost, effectiveness, and cost-effectiveness ratios for each combination of dataset and practice. This approach enabled a high-level comparison, identifying which practices consistently balanced fairness and performance.

IV. Analysis and Discussion of the Results

In the following sections, we present and discuss the results of the empirical study for each dataset, followed by recommendations based on overall cost-effectiveness. The discussion is arranged according to the corresponding **RQ**.

TABLE I
Wilcoxon Signed-Rank Test p-values for Fairness Metrics Across Contexts. Significant results (p <0.05) are colored and marked with an asterisk (*)

Context (Dataset)	AOD	FDRD	DI
Recidivism (COMPAS)	0.0391*	0.7422	0.1953
Economic (Adult)	0.0547	0.4609	0.1484
Marketing (Bank Marketing)	0.1670	0.0156*	0.0156*
Finance (German Credit)	0.0078*	0.0156*	0.0078*
Crime (Communities and Crime)	0.0422*	0.1077	1.0000

A. *RQ_1 — Contextual Fairness Evaluation*

In our **RQ_1**, we assessed the significance of fairness improvements achieved through fairness-aware practices applied to ML models across various domains. Table I presents the results of our statistical evaluation. More details are available in our online appendix [41].

A key observation is the variation in fairness improvements depending on the context and the specific fairness metrics. For example, in the **Recidivism** and **Crime** contexts, significant improvements were observed only in the *AOD* metric. Since these contexts fall under the broader Social Sciences domain, our findings indicate that fairness-aware practices may not consistently yield high equity improvements in these areas.

Conversely, the **Marketing** domain showed significant improvements in the *FDRD* and *DI* metrics, while **Finance** demonstrated the most robust results, with significant improvements across all metrics. In contrast, the **Economic** domain showed the weakest outcomes, with no metric achieving a statistically significant improvement in fairness. These results suggest that the effectiveness of fairness-aware practices is context-sensitive, with some domains more responsive to fairness interventions than others.

The varied significance levels across metrics and domains underscore the complexity of achieving fairness in ML. The effectiveness of fairness-aware practices appears to depend heavily on the context of the application. This variability highlights the need for comprehensive fairness evaluations that account for multiple fairness metrics and domain-specific factors. A universal approach to fairness may be insufficient; researchers and practitioners should analyze the unique characteristics of each domain to design tailored fairness-aware solutions. Contextual factors must be carefully considered to

ensure that fairness-aware ML models address the specific challenges of different real-world scenarios.

RQ$_1$ — Summary of the Results.

The results of our evaluation show that the effectiveness of fairness-aware practices in ML models varies across application domains and fairness metrics. Domains like *Finance* demonstrated significant improvements in all metrics, while *Economic* showed no improvement. Fairness improvements were more limited in domains like *Recidivism* and *Crime*, suggesting that the success of these practices depends on the specific characteristics of each domain. This highlights the importance of context-specific fairness evaluations in ML.

B. *RQ$_2$ — Cost-Effective Evaluation*

In the context of **RQ$_2$**, we assessed the cost-effectiveness of applying fairness-aware practices, where fairness gains were treated as benefits and performance loss as costs. This allowed us to recommend practices based on their relative trade-offs. The results of this analysis are presented in Table II. For each fairness-aware practice and dataset context, we report the average *cost-effectiveness ratio* for each combination of performance and fairness metrics computed. All the specific results used for these evaluations, alongside data and code, are available in our online appendix [41]. These values should be interpreted as follows: if the cost-effectiveness value exceeds one or falls below minus one for a specific practice on a specific dataset, the performance loss incurred by that practice outweighs its fairness improvement [18]. In Table II, we have highlighted with colors and marked with an asterisk the values indicating positive cost-effectiveness, meaning the specific practice is *recommended* for that dataset. Values marked as "NA" result from errors in computing the cost-effectiveness ratios, such as divisions by zero.

For the **Economic context**, analyzing the Adult dataset, both *Iterative Imputer* and *Simple Imputer* achieved the lowest cost-effectiveness ratio (0.60), suggesting that these data transformation techniques [39] offer the best balance between fairness and performance. *Label Encoder* (0.8648) and *Oversampling* (0.1338) had cost-effectiveness ratios below one, indicating that they improve fairness with minimal performance loss. Similarly, *Undersampling* (0.0072) also demonstrated a favorable trade-off, with high fairness gains relative to a small performance cost. In contrast, *SelectBest* (3.3612) and *Regularization* (1.5822) exceeded the threshold of one, meaning they provide fairness improvements but with more significant performance costs.

In the **Marketing context**, *Mutation Testing* (30.0711) showed an extremely high cost-effectiveness ratio, indicating a substantial performance loss with minimal fairness improvement, making it highly unfavorable, alongside *Undersampling* (-13.9254) and *Regularization* (-9.4770). Practices like *Label Encoder* (-0.1725), *Iterative Imputer* (-0.6635), and *SelectBest* (0.7065) displayed ratios close to zero, suggesting that their fairness benefits outweigh their performance costs.

For the **Crime context**, *Oversampling* (0.3125) and *Undersampling* (0.5744) exhibited low cost-effectiveness ratios, making them effective choices as they provide significant fairness improvements with minimal performance trade-offs. *Iterative Imputer* (0.1914) and *Simple Imputer* (0.2162) also showed favorable ratios, indicating they strike a good balance between fairness gains and minimal performance loss. On the other hand, *Regularization* (-0.0239) had a negative cost-effectiveness, reflecting less desirable trade-offs. As a result, *Oversampling*, *Undersampling*, *Iterative Imputer*, and *Simple Imputer* are the most cost-effective choices for this dataset.

In the **Recidivism prediction context**, using the COMPAS dataset, all practices showed cost-effectiveness ratios below one, indicating they all provide fairness improvements with relatively low performance costs, making them recommended options. Particularly, *Mutation Testing* (0.0914) stands out for offering one of the highest fairness improvements per unit of performance loss, followed by *Label Encoder* (-0.1250).

Finally, for the **Finance context** represented by the German Credit dataset, all practices demonstrated good cost-effectiveness, except for *Mutation Testing* (-1.8315). The most effective practices were *Iterative Imputer*, *Simple Imputer*, and *Label Encoder* (-0.0158), which effectively increase fairness without significantly harming performance. *Oversampling* (-0.0837) and *SelectBest* (-0.4439) also showed good results.

This dataset-specific analysis highlights that fairness-aware practices are not universally effective across different contexts. Practices applied at the *Data Preparation* stage tend to perform well, offering fairness improvements with minimal performance costs, while *Mutation Testing* is notably less effective. By focusing on practices with a low cost-effectiveness ratio, we can better guide the selection of fairness-aware techniques that optimize the balance between fairness and performance.

RQ$_2$ — Summary of the Results.

The cost-effectiveness analysis of fairness-aware practices across different datasets revealed that practices applied at the *Data Preparation* stage generally offer good fairness improvements with minimal performance costs. On the one hand, techniques like *Iterative Imputer*, *Simple Imputer*, *Label Encoder*, *Oversampling*, and *Undersampling* were consistently effective, providing a favorable balance between fairness and performance. On the other hand, *Mutation Testing* was generally less effective, showing high-performance costs with minimal fairness gains. Each dataset showed distinct results, highlighting the importance of context in selecting the most cost-effective practices.

V. Threats To Validity

This section outlines potential threats to the validity of our empirical study and the mitigation strategies applied.

TABLE II
AVERAGE COST-EFFECTIVENESS OF FAIRNESS-AWARE PRACTICES ACROSS CONTEXTS. AN ASTERISK HIGHLIGHTS RECOMMENDED PRACTICES.

Practice	Economic (Adult)	Marketing (Bank)	Crime (Communities&Crime)	Recidivism (COMPAS)	Finance (German)
IterativeImputer	0.0060*	-0.6635*	0.1914*	0.2873*	-0.0158*
LabelEncoder	0.8648*	-0.1725*	NA	-0.1250*	-0.0158*
Mutation Testing	-0.1423*	30.0711	2.8356	0.0914*	-1.8315
Oversample	0.1338*	NA	0.3125*	0.2464*	-0.0837*
Regularization	1.5822	-9.4770	-0.0239*	-0.5797*	-0.7214*
SelectBest	3.3612	0.7065*	-0.2439*	0.4941*	-0.4439*
SimpleImputer	0.0060*	NA	0.2162*	0.3241*	-0.0158*
Undersample	0.0072*	-13.9254	0.5744*	0.2477*	-0.2180*

Internal Validity. Internal validity concerns whether our results genuinely reflect the factors under study. A primary threat is the specific implementation choices made for fairness-aware practices. For instance, choosing a regular oversampling technique over SMOTE [35] could impact results. To mitigate this, we closely examined the definitions of these practices [14] and based our implementation decisions on the original design of the cataloged practices. However, alternative implementations could produce different outcomes, influencing both performance and fairness results.

External Validity. External validity pertains to the generalizability of our findings beyond the study's specific setup. We selected contexts for experimentation grounded in recent research [11]. Furthermore, the datasets chosen for each context are frequently used in fairness studies [8], [12], [46]. Nonetheless, our experimentation may not cover all possible contexts, and further studies are needed to validate the broader applicability of our findings. To support replication and further research, all data and scripts are publicly accessible [41].

Construct Validity. Construct validity reflects how well the study's measurements align with the constructs being evaluated. A potential threat is the choice of datasets to represent different contexts. To address this, we selected widely used datasets [11] that are pertinent to our focus on fairness-performance trade-offs [8], [12], [24], [46]. Another consideration is our selection of fairness metrics (AOD, FDRD, DI) and performance metrics, which, though not exhaustive, are widely recognized in the literature as robust fairness measures [8], [46]. The ML model used could also have influenced results; we, therefore, employed a well-established model common in fairness research [8], [11]. For $\mathbf{RQ_2}$, we based our conclusions on the established cost-effectiveness framework [18].

Conclusion Validity. Conclusion validity addresses the reliability of our conclusions. A key threat lies in the statistical test applied to $\mathbf{RQ_1}$, namely, the Wilcoxon signed-rank test [48]. This test assumes certain data distribution characteristics, and violating these assumptions could impact results. To mitigate this, we assessed the data distribution using the Shapiro-Wilk test [47] to check for normality, ensuring we selected the most appropriate test for reliable conclusions.

VI. CONCLUSION

This paper empirically examines fairness-aware practices—established ML engineering practices known to impact fairness positively. To achieve this, we conducted a contextual evaluation to assess the significance of fairness improvements achieved by these practices. In this evaluation, we carefully selected high-stakes application domains to explore whether specific areas benefit from particular fairness-aware practices. Second, we performed a cost-effectiveness evaluation, treating performance loss as the cost of applying these practices and fairness improvement as the benefit, assessing this trade-off empirically. Our findings indicate that different contexts may require tailored fairness adjustments, as not all practices proved effective across all domains. Furthermore, the cost-effectiveness evaluation revealed that some practices may not be justifiable in specific contexts due to their performance costs, providing practitioners with a preliminary recommendation of which practice they should apply.

The insights gained from our study set the foundation for our future research agenda. First, we plan to broaden our work by including additional tasks, protected attributes, and context-specific metrics to deepen the evaluation of selected practices. Additionally, we aim to develop a recommender system to guide practitioners, based on our extensive experiments, in selecting optimal combinations of fairness-aware practices to achieve fairness in ML systems.

DATA AVAILABILITY

The data that support the findings of this study are openly available in our online appendix [41].

ACKNOWLEDGMENT

We acknowledge the use of ChatGPT-4 to ensure linguistic accuracy and enhance the readability of this article. We acknowledge the support of the European Union - NextGenerationEU through the Italian Ministry of University and Research, Project PRIN 2022 PNRR "FRINGE: context-aware FaiRness engineerING in complex software systEms" (grant n. P2022553SL, CUP: D53D23017340001) and Project PRIN 2022 "QualAI: Continuous Quality Improvement of AI-based Systems" (grant n. 2022B3BP5S, CUP: H53D23003510006).

REFERENCES

[1] J. Zhou and F. Chen, *Human and Machine Learning*. Springer, 2018.
[2] P. Wang, E. Fan, and P. Wang, "Comparative analysis of image classification algorithms based on traditional machine learning and deep learning," *Pattern recognition letters*, vol. 141, pp. 61–67, 2021.

[3] J. Ni, Y. Chen, Y. Chen, J. Zhu, D. Ali, and W. Cao, "A survey on theories and applications for self-driving cars based on deep learning methods," *Applied Sciences*, vol. 10, no. 8, p. 2749, 2020.

[4] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[5] T. P. Pagano, R. B. Loureiro, F. V. Lisboa, R. M. Peixoto, G. A. Guimarães, G. O. Cruz, M. M. Araujo, L. L. Santos, M. A. Cruz, E. L. Oliveira *et al.*, "Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods," *Big data and cognitive computing*, vol. 7.

[6] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Computing Surveys (CSUR)*.

[7] M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro, "Bias mitigation for machine learning classifiers: A comprehensive survey," *ACM Journal on Responsible Computing*, vol. 1, no. 2.

[8] Z. Chen, J. M. Zhang, F. Sarro, and M. Harman, "Fairness improvement with multiple protected attributes: How far are we?" in *Proceedings of the IEEE/ACM 46th ICSE*.

[9] M. Zhang and J. Sun, "Adaptive fairness improvement based on causality analysis," in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2022. New York, NY, USA: Association for Computing Machinery, 2022, p. 6–17. [Online]. Available: https://doi.org/10.1145/3540250.3549103

[10] M. Hort, J. M. Zhang, F. Sarro, and M. Harman, "Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods," in *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2021. New York, NY, USA: Association for Computing Machinery, 2021, p. 994–1006. [Online]. Available: https://doi.org/10.1145/3468264.3468565

[11] A. Fabris, S. Messina, G. Silvello, and G. A. Susto, "Algorithmic fairness datasets: the story so far," *Data Mining and Knowledge Discovery*, vol. 36, no. 6. [Online]. Available: http://dx.doi.org/10.1007/s10618-022-00854-z

[12] V. De Martino, G. Voria, C. Troiano, G. Catolino, and F. Palomba, "Examining the impact of bias mitigation algorithms on the sustainability of ml-enabled systems: A benchmark study," *Available at SSRN 4966447*.

[13] S. Martínez-Fernández, J. Bogner, X. Franch, M. Oriol, J. Siebert, A. Trendowicz, A. M. Vollmer, and S. Wagner, "Software engineering for ai-based systems: a survey," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 31, no. 2, pp. 1–59, 2022.

[14] G. Voria, G. Sellitto, C. Ferrara, F. Abate, A. De Lucia, F. Ferrucci, G. Catolino, and F. Palomba, "A catalog of fairness-aware practices in machine learning engineering," *arXiv preprint arXiv:2408.16683*, 2024.

[15] A. Burkov, *Machine learning engineering*. True Positive Incorporated, 2020, vol. 1.

[16] G. Voria, G. Sellitto, C. Ferrara, F. Abate, A. De Lucia, F. Ferrucci, G. Catolino, and F. Palomba, "Fairness-aware practices from developers' perspective: A survey," *Available at SSRN 4949224*.

[17] V. Raina and S. Krishnamurthy, *Data Preparation*, 2022. [Online]. Available: https://doi.org/10.1007/978-1-4842-7419-4_14

[18] S. Riegg Cellini and J. Edwin Kee, "Cost-effectiveness and cost-benefit analysis," *Handbook of practical program evaluation*, pp. 636–672, 2015.

[19] L. Pascarella, F. Palomba, and A. Bacchelli, "Fine-grained just-in-time defect prediction," *Journal of Systems and Software*, vol. 150, pp. 22–36, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0164121218302656

[20] H. Hofmann, "Statlog (German Credit Data)," UCI Machine Learning Repository, 1994, DOI: https://doi.org/10.24432/C5NC77.

[21] C. Starke, J. Baleis, B. Keller, and F. Marcinkowski, "Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature."

[22] 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*.

[23] M. Wei and Z. Zhou, "Ai ethics issues in real world: Evidence from ai incident database," 2022.

[24] J. Chakraborty, S. Majumder, and T. Menzies, "Bias in machine learning software: why? how? what to do?" in *Proceedings of the 29th ACM ESEC/FSE*.

[25] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and information systems*, 2012.

[26] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*.

[27] J. Chakraborty, S. Majumder, Z. Yu, and T. Menzies, "Fairway: a way to build fair ml software," in *Proceedings of the 28th ACM ESEC/FSE*, 2020, pp. 654–665.

[28] S. Galhotra, Y. Brun, and A. Meliou, "Fairness testing: testing software for discrimination," in *Proceedings of the 2017 11th FSE*.

[29] S. Udeshi, P. Arora, and S. Chattopadhyay, "Automated directed fairness testing," in *Proceedings of the 33rd ACM/IEEE international conference on automated software engineering*.

[30] A. Aggarwal, P. Lohia, S. Nagar, K. Dey, and D. Saha, "Black box fairness testing of machine learning models," in *Proceedings of the 2019 27th ACM ESEC/FSE*.

[31] P. Zhang, J. Wang, J. Sun, G. Dong, X. Wang, X. Wang, J. S. Dong, and T. Dai, "White-box fairness testing through adversarial sampling," in *Proceedings of the ACM/IEEE 42nd ICSE*.

[32] Z. Chen, J. M. Zhang, F. Sarro, and M. Harman, "A comprehensive empirical study of bias mitigation methods for machine learning classifiers," *ACM Transactions on Software Engineering and Methodology*, vol. 32, no. 4, pp. 1–30, 2023.

[33] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, A. Wesslén *et al.*, *Experimentation in software engineering*. Springer, 2012, vol. 236.

[34] P. Ralph, S. Baltes, D. Bianculli, Y. Dittrich, M. Felderer, R. Feldt, A. Filieri, C. A. Furia, D. Graziotin, P. He, R. Hoda, N. Juristo, B. A. Kitchenham, R. Robbes, D. Méndez, J. S. Molléri, D. Spinellis, M. Staron, K. Stol, D. A. Tamburri, M. Torchiano, C. Treude, B. Turhan, and S. Vegas, "ACM SIGSOFT empirical standards," *CoRR*, vol. abs/2010.03525, 2020. [Online]. Available: https://arxiv.org/abs/2010.03525

[35] I. Valentim, N. Lourenco, and N. Antunes, "The impact of data preparation on the fairness of software systems."

[36] M. L. C. Lauron and J. P. Pabico, "Improved sampling techniques for learning an imbalanced data set," *arXiv preprint arXiv:1601.04756*, 2016.

[37] M. Vega-Gonzalo and P. Christidis, "Fair models for impartial policies: Controlling algorithmic bias in transport behavioural modelling," *Sustainability (Switzerland)*, vol. 14, no. 14, 2022, cited By 0.

[38] S. Raza, D. Reji, and C. Ding, "Dbias: detecting biases and ensuring fairness in news articles," *International Journal of Data Science and Analytics*, 2022, cited By 0.

[39] S. Biswas and H. Rajan, "Fair preprocessing: Towards understanding compositional fairness of data transformers in machine learning pipeline," S. D., Ed. Association for Computing Machinery, Inc.

[40] P. Ma, S. Wang, and J. Liu, "Metamorphic testing and certified mitigation of fairness violations in nlp models," B. C., Ed., vol. 2021-January. International Joint Conferences on Artificial Intelligence, 2020, pp. 458–465, cited By 31.

[41] A. Authors, "Online appendix." [Online]. Available: https://figshare.com/s/7d42fe4fbedd483b482e

[42] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi, "A survey on datasets for fairness-aware machine learning," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 12, no. 3, p. e1452, 2022.

[43] B. Becker and R. Kohavi, "Adult," UCI Machine Learning Repository, 1996, DOI: https://doi.org/10.24432/C5XW20.

[44] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decision Support Systems*, vol. 62, pp. 22–31, 2014.

[45] M. Redmond, "Communities and Crime," UCI Machine Learning Repository, 2002, DOI: https://doi.org/10.24432/C53W3X.

[46] S. Majumder, J. Chakraborty, G. R. Bai, K. T. Stolee, and T. Menzies, "Fair enough: Searching for sufficient measures of fairness," *ACM Transactions on Software Engineering and Methodology*.

[47] E. González-Estrada and W. Cosmes, "Shapiro–wilk test for skew normal distributions based on data transformations," *Journal of Statistical Computation and Simulation*, vol. 89, no. 17, pp. 3258–3272, 2019.

[48] R. F. Woolson, "Wilcoxon signed-rank test," *Encyclopedia of Biostatistics*, vol. 8, 2005.

[49] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*. [Online]. Available: https://doi.org/10.1145/2347736.2347755