Multi-feature Procedural Fairness in Machine Learning

Tianze Zhu
School of Data Science
Lingnan University
Hong Kong SAR, China
tianzezhu@ln.hk

Ziming Wang
Guangdong Provincial Key Laboratory of Brain-inspired Intelligent Computation
Department of Computer Science and Engineering
Southern University of Science and Technology
Shenzhen, China
wangzm2021@mail.sustech.edu.cn

Hao Tong
School of Data Science
Lingnan University
Hong Kong SAR, China
haotong@ln.edu.hk

Changwu Huang
School of AI and Liberal Arts
Beijing Normal-Hong Kong Baptist University
Zhuhai, China
changwuhuang@bnbu.edu.cn

Jialin Liu
School of Data Science
Lingnan University
Hong Kong SAR, China
jialin.liu@ln.edu.hk

Xin Yao
School of Data Science
Lingnan University
Hong Kong SAR, China
xinyao@ln.edu.hk

***Abstract*—Machine learning models may reinforce existing social biases if fairness is not properly considered. Prior research on fair machine learning has largely focused on distributive fairness, which requires models to produce fair prediction outcomes across different demographic groups. By contrast, procedural fairness, which focuses on model decision logic, requiring consistent treatment of similar individuals across demographic groups, has been relatively understudied. Existing limited studies examine procedural fairness under a single sensitive feature with two groups (e.g., female and male), leaving more complex settings with multiple sensitive features and multiple groups unexplored. To address this limitation, we propose a novel metric for evaluating overall procedural fairness across multiple sensitive features and multiple groups. We further propose a learning method that incorporates procedural fairness objectives while preserving predictive performance. Experimental results show that our method achieves high overall procedural fairness across all sensitive features and groups while maintaining a competitive accuracy.**

***Index Terms*—Procedural Fairness, AI Ethics, Fair Machine Learning, Feature Attribution Explanation, Fairness Accuracy Trade-off**

I. Introduction

Several real-world cases have demonstrated that machine learning (ML) systems can produce biased or discriminatory behavior when not carefully designed or monitored [1] [2]. Even models with strong predictive performance may rely on opaque decision logic, making it difficult to detect or prevent unfair treatment of certain demographic groups. These challenges have made fairness a central concern in modern machine learning research [1].

Most existing work on ML fairness has focused on distributive fairness [3], which aims to equalize prediction outcomes across demographic groups. Various fairness metrics and mitigation techniques, including pre-process , in-process , and post-process methods , have been proposed [4]. While these approaches are important, outcome-level parity alone does not guarantee that a model's decision-making process is fair [3]. Models may rely on different features, or assign different importance to the same features, for individuals from different groups, even when their final predictions appear similar. Such inconsistencies can introduce procedural bias, potentially reinforcing structural inequalities despite satisfying distributive fairness criteria [3] .

In this context, procedural fairness [3] shifts the focus from model outputs to the internal decision logic of the model, emphasizing consistent treatment of similarly situated individuals across protected groups. Although neural networks and other black-box models are inherently difficult to interpret, advances in explainable artificial intelligence (XAI), particularly feature attribution explanation (FAE) methods [5] , provide tools for examining how models use input features for individual predictions [6]. These explanations make it possible to analyze whether decision processes differ systematically across demographic groups.

Despite its importance, procedural fairness remains underexplored in the ML literature. Only a small number of studies formalize metrics to measure procedural fairness, and they are limited to settings with only two demographic groups (e.g., female and male) [3] [7] [8]. This restriction limits their applicability in real-world scenarios, where decisions often involve multiple sensitive features and a diverse set of demographic subgroups.

Beyond these challenges, class imbalance can further affect predictive performance and model behavior [9]. In procedural fairness settings, it may distort feature importance across demographic groups, biasing optimization. Mitigating class imbalance is therefore important to ensure fair and reliable model decisions.

Motivated by these limitations, this paper focuses on measuring and optimizing overall procedural fairness in multi-feature multi-group settings. By moving beyond pairwise or

two-group assumptions, the proposed framework better reflects real-world decision-making scenarios involving diverse populations. Our approach aims to ensure that ML models are both accurate and procedurally fair, with consistent decision logic for similar individuals across demographic groups.

The main contributions of our paper include: (i) A new metric to quantify the overall procedural fairness of ML models across all sensitive features, each comprising two or more demographic groups, in settings where multiple sensitive features are present; (ii) A new learning method for constructing models that achieve overall procedural fairness across all sensitive features, each comprising two or more demographic groups, while maintaining competitive accuracy; and (iii) Comprehensive experimental studies explain why the proposed learning method improves overall procedural fairness, investigate the impact of class imbalance on procedural fairness optimization, and examine the relationship between optimizing overall procedural fairness and accuracy.

II. Related Work

A. Measuring Procedural Fairness

Early work by Grgić-Hlača et al. [10] evaluated procedural fairness based on whether input features were perceived as fair by users. More recent studies have examined fairness in model explanations, including protected groups and explanation-based bias detection [11] [12].

Wang et al. [3] further formalized procedural fairness by distinguishing between individual-level and group-level notions. Individual procedural fairness ensures consistent decision logic for similar individuals, while group procedural fairness ensures consistency across demographic groups. To quantify group procedural fairness, they proposed the GPF_{FAE} metric, which computes FAEs for two groups and applies the maximum mean discrepancy (MMD) test to measure their differences, with the resulting p-value serving as the metric. The FAEs are obtained using $SHAP$ [13], which reflects feature contributions to model outputs.

Germino et al. [8] proposed another group procedural fairness metric, $EDiff$, which relies on counterfactual samples generated by flipping sensitive features for each instance. $EDiff$ measures the total absolute difference between the FAEs of original and counterfactual samples and averages this quantity across the dataset, where the FAEs are computed using $FastSHAP$ [14].

However, both GPF_{FAE} [3] and $EDiff$ [8] share a key limitation: they only consider a single sensitive feature and measure procedural fairness between two demographic groups defined by that feature. In real-world settings, multiple sensitive features with multiple demographic groups are often involved. In such cases, existing metrics are unable to measure overall procedural fairness across all sensitive features and demographic groups. This limitation motivates the development of our new metric that captures overall procedural fairness.

In distributive fairness, the concept of intersectional groups or subgroups [15] has been introduced to handle multiple sensitive features by evaluating fairness over combinations of sensitive features. This line of work offers a promising direction for procedural fairness, motivating methods that assess fairness across multiple groups and sensitive features to better capture real-world complexity.

B. Achieving Procedural Fairness

Besides methods for measuring procedural fairness, several optimization approaches have been proposed to achieve procedural fairness during model training. Wang et al. [7] introduced a learning method that jointly optimizes procedural fairness and accuracy. Their experimental results demonstrate that models trained with this loss function can achieve high procedural fairness with only a small decrease in accuracy.

Germino et al. [8] also proposed a learning-based approach that simultaneously optimizes distributive fairness, procedural fairness, and predictive performance. By directly incorporating Statistical Parity and $EDiff$ into the training objective, their method is able to produce models that achieve both strong distributive fairness and procedural fairness.

However, both approaches optimize procedural fairness only between two demographic groups under a single sensitive feature, making them unsuitable for settings with multiple sensitive features, where overall procedural fairness across all features cannot be guaranteed. This limitation highlights an important research gap: the lack of learning methods that can optimize procedural fairness across multiple sensitive features and multiple demographic groups, motivating the development of a new learning method that optimizes overall procedural fairness in this multi-feature, multi-group setting.

III. Multi-feature Procedural Fairness

A. New Multi-feature Multi-group Procedural Fairness Metric

We first partition samples into subgroups defined by unique combinations of sensitive feature values, and then measure procedural fairness across all subgroups. To quantify overall procedural fairness, we compute procedural fairness for all subgroup pairs and aggregate the results into our proposed metric, Subgroup Procedural Fairness ($SGPF$). For each subgroup pair, we employ GPF_{FAE} [3] to measure pairwise procedural fairness, and aggregate all pairwise GPF_{FAE} as defined in Eq. (1):

$$SGPF = \left(\frac{\sum\limits_{i,j \in \{1,\ldots,N_G\}, i \neq j} (GPF_{i,j})^{0.5}}{\binom{N_G}{2}} \right)^2, \qquad (1)$$

where N_G denotes the total number of subgroups and $GPF_{i,j}$ represents the GPF_{FAE} value between subgroups i and j.

When aggregating all pairwise GPF_{FAE}, we employ the power mean with a power of 0.5. This aggregation assigns greater weight to smaller GPF_{FAE} values in the final result, which means placing greater emphasis on subgroup pairs with greater unfairness, while still accounting for all pairwise GPF_{FAE}. $SGPF$ is bounded within $[0, 1]$, where larger values indicate greater procedural fairness, and $SGPF = 1$ corresponds to perfect overall procedural fairness.

B. Learning with Multi-feature Multi-group Procedural Fairness

We propose a new learning method, named as Multi-feature Multi-group Procedural Fairness Learning ($MMPFL$), to train models that jointly optimize predictive performance and $SGPF$ by formulating a loss function with two components. The first component is the binary cross-entropy loss, denoted as L_{CE}. The second component targets $SGPF$. Since $SGPF$ employs GPF_{FAE}, which is non-differentiable [3], $SGPF$ cannot be directly embedded into the loss function. Therefore, we optimize $SGPF$ by minimizing differences among FAEs across subgroups, resulting in a differentiable surrogate loss denoted as L_{SGPF}. The $SHAP$ explanation method [13] used to compute GPF_{FAE} is notably time-consuming and thus unsuitable for training. Consequently, following Wang et al. [7], we use gradient-based explanations ($Grad$ [16]) to obtain FAEs during model training.

To obtain L_{SGPF}, we first define the FAE difference in Eq. (2):

$$dist(k, n, i, j) = \|FAE_{k,i} - FAE_{n,j}\|_1, \tag{2}$$

which represents the FAE difference between sample k and sample n, where k belongs to subgroup i and n belongs to subgroup j. This difference is measured by the L_1 distance between the FAEs of k and n. We then compute L_{SGPF} as shown in Eq. (3):

$$L_{SGPF} = \left(\frac{\sum_{i=1}^{N_G} \sum_{k=1}^{size(G_i^{'})} \left[\sum_{\substack{j=1 \\ j \neq i}}^{N_G} dist(k, n, i, j) \right]^2}{\sum_{i=1}^{N_G} size(G_i^{'})} \right)^{1/2}, \tag{3}$$

where $FAE_{k,i}$ denotes the FAE of sample k from subgroup i, and $FAE_{n,j}$ denotes the FAE of sample n from subgroup j, with n being the closest neighbor of k in subgroup j in terms of Euclidean distance. Here, $G_i^{'}$ denotes a set of 100 samples randomly selected from subgroup i, and N_G denotes the total number of subgroups.

Finally, we combine L_{CE} and L_{SGPF} to form the total loss in Eq. (4):

$$L_{total} = L_{CE} + \alpha \times CI \times L_{SGPF}, \tag{4}$$

where α controls the weight of L_{SGPF} in the loss function. CI denotes the minority class ratio, defined as the minority class size divided by the total sample size. This term ensures that the model maintains predictive performance when the classes are highly imbalanced. The complete training process is summarized in Algorithm 1.

Algorithm 1 MMPFL: Optimizing $SGPF$ to Achieve Overall Procedural Fairness across Multiple Sensitive Features and Groups

Require: Number of subgroups N_G, training data $\mathcal{D}$, number of training epochs T

1: Randomly initialize model
2: **for** $t = 1$ to T **do**
3: Compute binary cross-entropy loss L_{CE} via training data $\mathcal{D}$
4: Compute SGPF loss L_{SGPF} according to Eq. (3) via training data $\mathcal{D}$ and N_G
5: Compute total loss L_{total} according to Eq. (4)
6: Update model using L_{total}
7: **end for**

IV. Experimental Studies

To validate the effectiveness of $MMPFL$, we evaluate it on multiple real-world datasets and examine its performance in terms of accuracy and overall procedural fairness. In this section, we first describe the experimental settings, then present the evaluation results of $MMPFL$, and finally provide further analysis of the results.

A. Experiment Settings

In our experiments, we use nine publicly available datasets for binary classification: $Adult$ [17], $Bank$ [18], $COMPAS$ [19], $Default$ [20], $Diabetes$ [21], $Drug$ [22], $Dutch$ [23], $LSAT$ [24], and $OULAD$ [25]. Among these, $Adult$, $Bank$, $COMPAS$, $Drug$, $LSAT$, and $OULAD$ each include two sensitive features, both with two demographic groups. In contrast, $Default$, $Diabetes$, and $Dutch$ each have two sensitive features, with one feature having two groups and the other three. These datasets also exhibit varying levels of class imbalance, with seven having a minority-class proportion below 30%, two below 20%, and $LSAT$ being the most imbalanced at 9.82%. During preprocessing, certain features are dropped or re-encoded as needed following Le Quy et al. [26], samples with missing values are removed, each dataset is randomly split into training and testing sets with a 5:5 ratio, ensuring that all subgroups are also split proportionally. Numerical features are normalized, and nominal features are encoded using one-hot encoding.

In our experiments, following Wang et al. [7], we train a two-layer multi-layer perceptron (MLP) with ReLU activation and the Adam optimizer on all datasets. The loss function is computed as described in Algorithm 1. All models share the same configuration, with a hidden layer of 16 nodes, a learning rate of 0.01, and a maximum of 300 training iterations. The hyperparameter α in our loss function is set to 0.2, and each experiment is independently repeated 20 times using different random seeds.

B. Experimental Evaluation of $MMPFL$

To evaluate our proposed $MMPFL$, we compare it with the state-of-the-art methods by Wang et al. [7] and Germino et al. [8]. For both baseline models, we adopt their proposed loss functions while keeping all other MLP settings identical for consistency. The two baselines are denoted as MLP_{GPF} and MLP_{EDiff}, respectively.

Since our datasets contain multiple subgroups across multiple sensitive features and both baselines are designed to optimize procedural fairness between only two demographic groups, we create multiple models for each baseline, each

targeting a specific pair of subgroups. For example, $Adult$, $Bank$, $COMPAS$, $Drug$, $LSAT$, and $OULAD$ each contain two sensitive features, each with two demographic groups. Therefore, they contain four subgroups in total, resulting in six unique subgroup pairs. Consequently, we create six models for both MLP_{GPF} and MLP_{EDiff} baselines, yielding 12 baseline models in total. For $Default$, $Diabetes$, and $Dutch$, each contains two sensitive features, one with two demographic groups and the other with three. This results in six subgroups, which correspond to 15 unique subgroup pairs. Thus, we obtain 30 baseline models in total, consisting of 15 MLP_{GPF} models and 15 MLP_{EDiff} models.

For each dataset, we compare $MMPFL$ with each baseline model in terms of $SGPF$ and accuracy using the Wilcoxon rank-sum test. Whenever $MMPFL$ is statistically better, similar, or worse than a baseline model, we record it as a Win, Draw, or Loss, respectively. The results are summarized in the 2nd and 3rd columns of Table I.

TABLE I
WIN/DRAW/LOSS OUTCOMES OF $MMPFL$ AND $MMPFL_{noCI}$ VERSUS MLP_{GPF} AND MLP_{EDiff} BASELINES, BASED ON WILCOXON RANK-SUM TESTS FOR $SGPF$ AND ACCURACY OVER 20 TRIALS.

Dataset	$MMPFL$ $SGPF$	$MMPFL$ Accuracy	$MMPFL_{noCI}$ $SGPF$	$MMPFL_{noCI}$ Accuracy
Adult	10/2/0	5/6/1	12/0/0	0/2/10
Bank	12/0/0	2/4/6	12/0/0	0/5/7
COMPAS	11/1/0	6/4/2	11/1/0	6/1/5
Drug	12/0/0	7/5/0	12/0/0	6/5/1
LSAT	6/2/4	10/0/2	6/3/3	6/4/2
OULAD	12/0/0	6/6/0	12/0/0	6/1/5
Default	21/9/0	15/0/15	30/0/0	15/0/15
Diabetes	30/0/0	15/15/0	30/0/0	15/15/0
Dutch	29/1/0	0/1/29	30/0/0	0/0/30

From Table I, we observe that on eight out of nine datasets $MMPFL$ achieves higher or comparable $SGPF$ than all baseline models. In $Adult$, $COMPAS$, $Drug$, $OULAD$, and $Diabetes$, $MMPFL$ achieves $SGPF$ no worse than any baseline, while maintaining a competitive accuracy.

However, in $Bank$, $Default$, and $Dutch$, although $MMPFL$ achieves high $SGPF$, its accuracy is lower than that of at least half of the baseline models. We further evaluated $MMPFL$ on these datasets with smaller values of α in our loss function and found that it can achieve improved accuracy while maintaining $SGPF$ at least as high as any baseline model on all three datasets. The experimental results on $Bank$, $Default$, and $Dutch$ using smaller values of α are reported in Table II.

On $LSAT$, $MMPFL$ is unable to simultaneously achieve an $SGPF$ comparable to all baseline models and a competitive level of accuracy. $LSAT$ exhibits the most severe class imbalance, with a minority-class proportion of 9.82%, suggesting that class imbalance may be a key factor in its suboptimal performance, and more effective methods for addressing class imbalance are thus still needed.

TABLE II
WIN/DRAW/LOSS OUTCOMES OF $MMPFL$ VERSUS MLP_{GPF} AND MLP_{EDiff} BASELINES ON THE $Bank$, $Default$, AND $Dutch$ DATASETS, WHERE α IS ADJUSTED. RESULTS ARE BASED ON WILCOXON RANK-SUM TESTS FOR $SGPF$ AND ACCURACY OVER 20 TRIALS.

Dataset	$SGPF$	Accuracy	α
Bank	12/0/0	6/5/1	0.1
Default	17/13/0	15/15/0	0.05
Dutch	22/8/0	12/7/11	0.04

C. Further Analysis

1) Why $MMPFL$ Works: Besides comparing $MMPFL$ with baseline models in terms of $SGPF$ and accuracy, we further analyze the source of $MMPFL$'s performance gains by examining decision logic differences, specifically feature contributions to model predictions. Before this analysis, it is important to note that overall group procedural fairness is defined as having similar feature contributions when an ML model makes predictions for different subgroups. Feature contribution is determined by two factors: the feature values and how the model uses these feature values to make predictions.

At the same time, different subgroups often exhibit differences in feature values, especially for features that are highly correlated with sensitive features. Figure 1 presents 2D visualizations of subgroup distributions for the nine datasets used in our experiments, from which we observe that different subgroups follow distinct distributions. As a result, to achieve overall procedural fairness, an ML model can reduce feature contribution differences across subgroups by assigning lower weights to features that are highly correlated with sensitive features during prediction.

With this understanding, we can now examine the key difference between $MMPFL$ and the baseline models. The baseline models minimize feature contribution differences between only two specific subgroups while ignoring the remaining subgroups, whereas $MMPFL$ minimizes feature contribution differences across all subgroups. When a model minimizes feature contribution differences between only two subgroups, other subgroups with different distributions are ignored, and procedural fairness is optimized in only a limited direction. In contrast, $MMPFL$ optimizes procedural fairness across all directions by minimizing feature contribution differences among all subgroups.

In addition, we further analyze the contributions of features that are highly correlated with sensitive features, which account for most of the variation across subgroups. The results are shown in Table III. For each dataset, we first identify the ten features most correlated with the sensitive features. For $MMPFL$, we then compute the approximate contributions of these features by averaging their Shapley values [13] across all testing samples. In Table III, the second column reports the mean feature contribution of these ten features in $MMPFL$, while the third column reports the mean contribution of the same features averaged across all baseline models.

We observe that, across all datasets, the average contribution of these ten features in $MMPFL$ is statistically smaller than

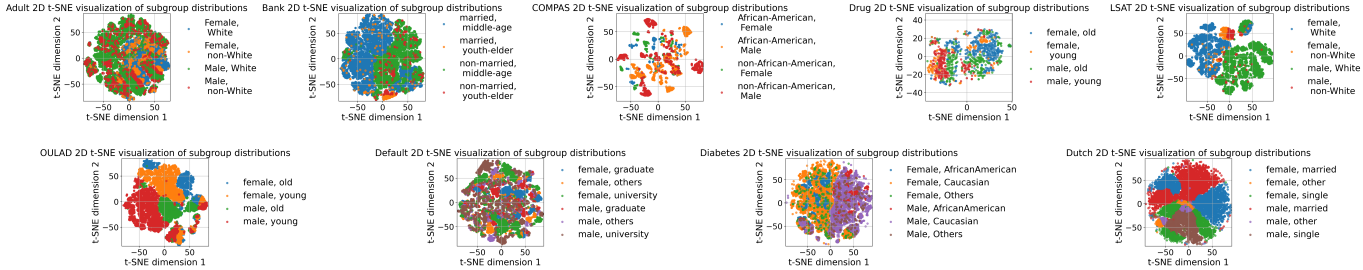

Fig. 1. t-SNE visualizations of subgroups distributions for the nine datasets.

or at least comparable to that in the baseline models. This indicates that, compared to the baselines, $MMPFL$ relies less on highly sensitive features when making predictions, resulting in smaller feature contribution differences across subgroups. In contrast, since the baseline models optimize feature contribution differences for only a specific pair of subgroups while ignoring others, they may still differentiate between other subgroups based on these highly sensitive features. This leads to larger average contributions of sensitive features and, consequently, lower overall procedural fairness. This analysis explains why $MMPFL$ achieves better or at least comparable overall procedural fairness than all baseline models on most datasets.

TABLE III
CONTRIBUTIONS OF THE TEN FEATURES HIGHLY CORRELATED WITH SENSITIVE FEATURES IN $MMPFL$ AND BASELINE MODELS, AVERAGED OVER 20 TRIALS. FOR EACH DATASET, THE SMALLER AVERAGE FEATURE CONTRIBUTION IS HIGHLIGHTED.

Dataset	$MMPFL$	Baselines
Adult	1.537e-02 (4.327e-02)	1.546e-02 (4.306e-02)
Bank	1.326e-04 (2.282e-04)	2.001e-03 (3.341e-03)
COMPAS	2.116e-02 (3.541e-02)	2.156e-02 (3.368e-02)
Drug	2.148e-02 (3.491e-02)	2.452e-02 (3.284e-02)
LSAT	8.754e-03 (1.358e-02)	9.456e-03 (1.452e-02)
OULAD	1.560e-03 (2.268e-03)	3.951e-03 (5.545e-03)
Default	4.078e-03 (6.828e-03)	4.182e-03 (7.206e-03)
Diabetes	3.670e-06 (5.319e-06)	7.292e-04 (1.344e-03)
Dutch	1.402e-03 (2.587e-03)	1.338e-02 (1.843e-02)

2) Impact of Class Imbalance: Inside $MMPFL$'s loss function, as shown in Eq. (4), in addition to L_{CE} and L_{SGPF}, we introduce CI, which represents the class imbalance ratio and is computed as the minority class size divided by the total sample size. $MMPFL$ is designed for binary classification tasks, where class imbalance is a well-known issue. Therefore, the purpose of incorporating CI into the loss function is to ensure that when the training data is highly imbalanced, $MMPFL$ can place greater emphasis on predictive performance by reducing the influence of L_{SGPF} on the total loss.

To validate the effectiveness of CI, we modify $MMPFL$ by removing CI, resulting in the loss function shown in Eq. (5):

$$L_{total}^{'} = L_{CE} + \alpha \times L_{SGPF}, \tag{5}$$

while keeping all other model settings and parameters unchanged. We then compare this modified model, denoted as $MMPFL_{noCI}$, with the baseline models on $SGPF$ and accuracy. The results are shown in columns 4 and 5 of Table I.

We observe that after removing CI, $MMPFL_{noCI}$ achieves higher $SGPF$ on most datasets, but its accuracy decreases. This occurs because $CI \in (0, 1)$ reduces the relative contribution of L_{SGPF} in the total loss. Without CI, the loss function places greater relative emphasis on overall procedural fairness and less on accuracy, leading to higher $SGPF$ but lower accuracy.

However, the effect of removing CI goes beyond this trade-off. We further compare $MMPFL_{noCI}$ with $MMPFL$ by examining changes in average $SGPF$, accuracy, and $F1$ score, where $F1$ reflects the model's ability to handle class imbalance. The results are shown in Table IV, where the datasets are sorted by CI in descending order, and $\delta SGPF$, $\delta Accuracy$, and $\delta F1$ represent the average changes in $SGPF$, accuracy, and $F1$ score, respectively, across 20 trials after removing CI from L_{total}.

TABLE IV
IMPACT OF REMOVING CI ON $SGPF$, ACCURACY, AND $F1$ SCORE. "$\delta SGPF$", "$\delta Accuracy$", AND "$\delta F1$" DENOTE $SGPF$, ACCURACY, AND $F1$ SCORE CHANGES AFTER REMOVING CI, AVERAGED OVER 20 TRIALS.

Dataset	CI	$\delta SGPF$	$\delta Accuracy$	$\delta F1$
Dutch	0.476	0.099	-0.005	-0.005
COMPAS	0.451	0.001	-0.002	-0.002
OULAD	0.314	0.103	-0.002	-0.011
Diabetes	0.242	0.022	-0.002	-0.000
Adult	0.240	0.010	-0.008	-0.022
Default	0.221	0.055	-0.009	-0.041
Drug	0.219	0.067	-0.002	-0.032
Bank	0.117	0.059	-0.002	-0.045
LSAT	0.098	0.018	-0.005	-0.084

We observe that removing CI leads to increases in $SGPF$ and decreases in accuracy and $F1$ score on most datasets. Notably, unlike $SGPF$ and accuracy, the decrease in $F1$ score tends to be larger for datasets with more severe class imbalance. These results indicate that incorporating CI into the loss function helps mitigate the impact of class imbalance, especially when classes are highly imbalanced, by preventing the model from favoring the majority class.

3) Parameter Sensitivity Analysis: In addition, we test the sensitivity of α in our loss function shown in Eq. (4), which represents the weight of overall procedural fairness. Specifically, we vary α from 0.2 to different values and observe

the resulting changes in $SGPF$ and accuracy of $MMPFL$. Figure 2 shows the average $SGPF$ and accuracy across all datasets for each α. As can be seen, increasing α helps to improve $SGPF$, with relatively minor impact on the accuracy.

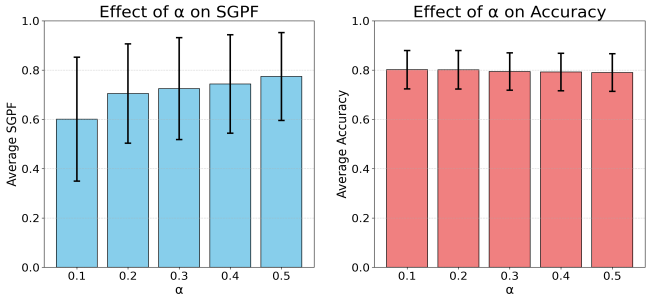

Fig. 2. Effect of α on $SGPF$ (left) and accuracy (right). Points denote mean values across datasets (error bars indicate standard deviation). Each dataset is run for 20 trials.

V. Conclusion

We propose $SGPF$, a new metric measuring overall procedural fairness across all sensitive features and demographic groups, and $MMPFL$, a novel learning method that optimizes both overall procedural fairness and model accuracy. $SGPF$ is the first metric to measure procedural fairness over multiple sensitive features and groups. Although an MLP is used in this paper, $MMPFL$ is generic and applicable to other machine learning models. Extensive experiments show that $MMPFL$ achieves high overall procedural fairness while maintaining strong accuracy. On most datasets, it improves both procedural fairness and accuracy simultaneously, though a trade-off between the two was observed.

In the future, we plan to investigate more efficient alternatives to $SHAP$, used in FAE, since $SHAP$ computation is very costly. We will also study multi-objective learning approaches that treat fairness and accuracy as separate learning objectives (loss functions) [2] [27].

Acknowledgment

This work was supported in part by the Hong Kong Phd Fellowship to Tianze Zhu funded by the Research Grants Council of the Hong Kong SAR Government, and in part by the Internal Grants of Lingnan University (F106118, SDS24A3, SUG-008/2425).

References

[1] C. Huang, Z. Zhang, B. Mao, and X. Yao, "An overview of artificial intelligence ethics," *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 4, pp. 799–819, 2022.

[2] Q. Zhang, J. Liu, Z. Zhang, J. Wen, B. Mao, and X. Yao, "Mitigating unfairness via evolutionary multiobjective ensemble learning," *IEEE Transactions on Evolutionary Computation*, vol. 27, no. 4, pp. 848–862, 2022.

[3] Z. Wang, C. Huang, and X. Yao, "Procedural fairness in machine learning," *Journal of Artificial Intelligence Research*, vol. 85, p. Article 20, February 2026, 30 pages.

[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] Z. Wang, C. Huang, Y. Li, and X. Yao, "Multi-objective feature attribution explanation for explainable machine learning," *ACM Transactions on Evolutionary Learning and Optimization*, vol. 4, no. 1, pp. 1–32, 2024.

[6] Z. Wang, C. Huang, and X. Yao, "A roadmap of explainable artificial intelligence: Explain to whom, when, what and how?" *ACM Transactions on Autonomous and Adaptive Systems*, vol. 19, no. 4, pp. 1–40, 2024.

[7] Z. Wang, C. Huang, K. Tang, and X. Yao, "Procedural fairness and its relationship with distributive fairness in machine learning," *arXiv preprint arXiv:2501.06753*, 2025.

[8] J. Germino, Y. Zhao, T. Derr, N. Moniz, and N. V. Chawla, "Explanation difference: Bridging procedural and distributional fairness," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 8, no. 2, 2025, pp. 1078–1090.

[9] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[10] N. Grgić-Hlača, M. B. Zafar, K. P. Gummadi, and A. Weller, "Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[11] A. V. Menon, Z. A. Omar, N. Nahar, X. Papademetris, L. E. Fiellin, and C. Kästner, "Lessons from clinical communications for explainable AI," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, 2024, pp. 958–970.

[12] M. Waller, O. Rodrigues, and O. Cocarascu, "Identifying reasons for bias: An argumentation-based approach," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 19, 2024, pp. 21 664–21 672.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[14] N. Jethani, M. Sudarshan, I. Covert, S.-I. Lee, and R. Ranganath, "FastSHAP: Real-time Shapley value estimation," *ICLR 2022*, 2022.

[15] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2564–2572.

[16] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[17] R. Kohavi *et al.*, "Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid," in *KDD*, vol. 96, 1996, pp. 202–207.

[18] 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.

[19] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," in *Ethics of Data and Analytics*. Auerbach Publications, 2022, pp. 254–264.

[20] I.-C. Yeh and C.-h. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2473–2480, 2009.

[21] B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, and J. N. Clore, "Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records," *BioMed Research International*, vol. 2014, no. 1, p. 781670, 2014.

[22] E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan, and A. N. Gorban, "The five factor model of personality and evaluation of drug consumption risk," in *Data Science: Innovative Developments in Data Analysis and Clustering*. Springer, 2017, pp. 231–242.

[23] P. Van der Laan, "The 2001 census in the Netherlands: Integration of registers and surveys," in *Conference the Census of Population*, 2000.

[24] L. F. Wightman, "LSAC national longitudinal bar passage study. LSAC research report series." 1998.

[25] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," *Scientific Data*, vol. 4, no. 1, pp. 1–8, 2017.

[26] 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.

[27] Q. Zhang, J. Liu, and X. Yao, "Fairness-aware multiobjective evolutionary learning," *IEEE Transactions on Evolutionary Computation*, vol. 29, no. 6, pp. 2372–2385, Dec. 2025.