Achieving Equalized Explainability Through Data Reconstruction

Shuang Wang
Electrical and Computer Engineering
Clemson University
Clemson, SC, USA
swang8@clemson.edu

Yongkai Wu
Electrical and Computer Engineering
Clemson University
Clemson, SC, USA
yongkaw@clemson.edu

***Abstract*—Recent progress in machine learning has placed a growing emphasis on explainability and fairness. However, many studies have confined efforts to leveraging explanatory techniques to promote model fairness, overlooking the essential fairness of the explanations. This study addresses this gap by proposing a novel principle of equalized explainability, which fulfills fair and uniform explanations across various demographic groups. To this end, we introduce a quantitative measure for assessing the explanation disparity leveraging Explainable Artificial Intelligence (XAI) tools. To achieve equalized explainability, we propose a reconstruction framework including modules for data reconstruction, equalization of explanations, and performance preservation. Experiments using real-world datasets demonstrate this framework's effectiveness in securing equitable and consistent explanations across different groups, as well as achieving trade-offs between fairness and explanation.**

***Index Terms*—machine learning, fairness, XAI**

I. Introduction

Machine learning (ML) has become prevalent in numerous fields and achieved significant advancements recently. Despite its widespread adoption, there is growing concern about the trustworthiness of advanced ML techniques, including neural networks, centered on fairness and explainability. The inherent complexity of these models often results in opaque decision-making processes and obscure underlying mechanisms. This opacity has sparked an increasing demand for better explainability in machine learning, highlighting the need for a deeper understanding of how these models work and make decisions [1]–[3].

To amend the lack of trustworthiness of machine learning models and systems, the research fields of Explainable AI (XAI) and Fairness-aware machine learning (Fair-ML) have emerged in recent years [4]. Fair-ML focuses on creating algorithms that make unbiased decisions, regardless of gender, race, or other sensitive demographic information. It involves developing fairness notions and implementing solutions that treat all groups equitably and do not perpetuate or exacerbate existing societal biases [5]–[10]. Two primary fairness notions often discussed are individual fairness [11], [12], ensuring similar individuals are treated similarly, and group fairness [13], [14], focusing on equal treatment and outcomes for different demographic groups. The Fair-ML solutions include adjusting training data to reduce biases [15], [16], modifying algorithms to prevent discriminatory outcomes [10], [17], or tweaking biased machine-learning decisions [18], [19]. XAI research focuses on developing methods and techniques that make AI decisions understandable to humans, bridging the gap between advanced AI algorithms and human interpretability. This encompasses a range of strategies, from model-agnostic approaches [20] that provide insights across various models to model-specific techniques [21] tailored to particular types of AI systems. Additionally, XAI research delves into global explainability [22], which seeks to provide an overarching understanding of a model's mechanisms, and local explainability [23], which aims to elucidate the reasoning behind specific individual decisions made by the model.

Recently, there has been a notable shift in research exploring the intersections of Fair-ML and XAI. Researchers [24]–[28] are leveraging XAI technologies to interpret and identify biases in ML models, aiming to address and mitigate bias-related issues. However, it is an under-explored concern that XAI might inadvertently introduce new types of biases, especially against under-represented groups. While XAI strives to make AI decisions more transparent and comprehensible, it may inadvertently produce explanations that are not equally accessible or understandable to all users [29]. This under-explored aspect emphasizes the complexity of achieving fairness in explainability and underscores the need for a nuanced understanding of how explainability interacts with different demographic groups, necessitating a more holistic approach to developing AI systems that are both fair and transparent.

In light of this, we propose a new notion of fairness, namely equalized explainability, requiring explanations offered by the same model to be equal across multiple demographic groups. Building on this idea, we develop a data reconstruction approach that aims to fulfill the equalized explainability while simultaneously preserving the data's utility and the model's performance. Specially, we design a general framework with three modules: a) The data reconstruction module preserves data utility by retaining the patterns and structural features of the original data through feature-specific reconstruction; b) The equalization module ensures that demographic groups receive equitable explanations from the XAI approach; c) The performance preservation module maintains predictive performance on the reconstructed data for predictive tasks. To

validate the efficacy of our methods, we conduct experiments with real-world datasets. The results from these experiments demonstrate that our approach effectively achieves equalized explanations by reconstructing the datasets.

II. Related Works

In exploring fairness and explainability within the ML field, existing studies primarily focus on enhancing the transparency and fairness of model decision-making processes, separately. Chakraborty et al. [30] utilize a K-Nearest Neighbors (KNN) algorithm specifically designed to detect and evaluate biases in black-box ML models. Hickey et al. [27] propose a new definition of "Fairness by Explicability", which embeds fairness restrictions by combining adversarial learning with the model explainability technique SHAP (SHapley Additive exPlanations) [31]. This technique exhibits remarkable effectiveness in mitigating the model's bias and strengthening statistical equity. In addition, Begley et al. [28] introduce an approach with the explanatory tool SHAP to promote fairness with ML models. Their attention lies in assigning the cumulative disparity of a model to specific input features, regardless of situations when sensitive attributes are not directly involved. This work effectively displays the capabilities to improve the fairness and explainability of ML models without compromising their performance. These studies above collectively demonstrate a trend of leveraging explainability as a tool or an application to influence and address fairness issues with ML models. With an in-depth exploration of the interaction between fairness and explainability, the concept of fairness should be considered in a broader sense in the context of explanation. Concurrently, the concept of fidelity in explanations becomes increasingly essential. Yeh et al. [32] propose fidelity as a crucial measure to evaluate the accuracy of model explanations. This concept is fundamental when examining the influence of explanations on different groups, as there can be considerable differences in the fidelity of explanations provided to these groups.

Efforts involve the development of algorithms that incorporate fairness and explainability [25], [26], applying tools such as SHAP to get fair results through adversarial learning [27], and improving interpretability by employing SHAP to achieve a fair decision-making process [28]. Also, Aïvodji et al. [33] have identified the concept of *fairwashing* as a means to conceal biases in ML algorithms by providing misleading explanations. While the above studies establish a connection between fairness and explainability, primarily focusing on applying explanatory techniques to assist the model's fairness, they fail to elucidate what a fair explanation is. In view of this, it becomes essential to embed fairness intrinsically within explanations themselves rather than just facilitating fairness through the application of explanation. Although some other studies view *fidelity* as a reliable measure of fair explanation, claiming that similar levels of high explanation fidelity across demographic groups signify their fairness [29], [34], this perspective may inadvertently obscure the full spectrum of what constitutes a fair explanation. It is particularly true when equal levels of high explanation fidelity across groups do not necessarily equate to consistent or equitable explanations for those groups. A recent study corroborates this perspective, indicating that disparities in explanations tend to arise in complex and non-linear models across demographic groups, and the inconsistency is particularly obvious with certain post-hoc explanatory methods [29].

The study accomplished by Dai et al. [29] has attracted considerable attention. This study centers on analyzing the quality of explanations generated via post-hoc explanation techniques among various demographics, illustrating the significance of high-quality explanations for all groups. Accordingly, Balagopalan et al. [34] propose a point of view in which an explanatory model is considered fair if it demonstrates high fidelity for all protected groups. However, this perspective may fail to consider the issue of consistency in explanations among different groups. Associating high fidelity with explanation could ignore the disparities in understanding among demographic groups, thereby indicating the necessity for a more delicate method to establish fairness in explanations.

Inconsistency or disparities in the explanations received by different groups within the same model may lead to biases in understanding the model among these groups, even though the fidelity of those explanations provided to different groups may be similar. Therefore, we believe this bias in comprehension can be seen as the model explanation's prejudices or discrimination against certain groups, resulting in inequality in trust and reliance within demographics. Such biased explanations gradually weaken the model's trustworthiness and eventually trigger certain groups' distrust or even rejection of the conclusions or recommendations derived from the model, impacting its effectiveness and applicability in real-world scenarios. For example, suppose a scenario where the same bank loan model provides distinct explanations for credit approval decisions to male and female users. While this loan model emphasizes marital status for female users and credit history for male users, this disparity in explanations can be considered discrimination against females. If credit history and marital status play an equal impact on the possibility of credit approval for all demographics, then these features should be evaluated to have identical roles in the decision-making process. Therefore, the model's explanations for various groups should reflect this equality, avoiding emphasizing or understating the contribution of certain features for a particular group, guaranteeing equity and fairness in the model's explanations for any demographic.

III. Preliminaries

A. Fairness Notions

In this study, we consider a classification model f trained on a dataset $\mathcal{D}$. The dataset $\mathcal{D}$ is formed by the tuple $< X_i, Y_i >$, and $i \in \{1, ..., m\}$ is a data point index in $\mathcal{D}$. X consists of several demographic groups identified by sensitive attributes such as race or gender. Y is a categorical label vector.

B. Feature Importance Explanation

Feature Importance (FI) is a prevalent metric in interpretative models that serves as an explanation mechanism,

notably in attribution-based methods such as SHAP [31]. SHAP provides a cooperative game theory-based approach to quantify the contribution of each feature to a model's predictions, assigning a methodical allocation of Shapley Values to each feature for each instance. This attribution indicates how the presence or absence of a feature changes the model's prediction, thus elucidating the impact of that specific feature and effectively acting as an explanation by highlighting the relative importance of each feature. Concurrently, the FI metric conforms to the principle of additive feature attribution, which suggests that the total sum of individual feature attributions approximates the model's prediction [31], [35]. Precisely, the definition of FI is the average absolute value of the attribution of a feature across all instances, expressed as $FI_j = \sum_{i=1}^{m} \|\phi_j^{(i)}\|$, where $\phi_j^{(i)}$ represents the Shapley Values or the attribution value of the j-th feature for the i-th instance, and m is the total number of instances. These attributions strengthen comprehension of the model's behavior at a detailed level, thereby delivering an in-depth explanation of the predictive features.

IV. Quantifying Explainablity Disparity

A. A New Notion of Equalized Explainability

Without loss of generality, we divide the dataset $\mathcal{D}$ into k groups, denoting as $\mathcal{D}_{G_1}$, $\mathcal{D}_{G_2}$, . .., and $\mathcal{D}_{G_k}$, based on the sensitive feature(s). Any arbitrary explanatory tool can be employed to generate explanations for the trained $f(\cdot)$ over the dataset $\mathcal{D}$, denoted as $E(f, \mathcal{D})$. To assess the disparity in explanations across different groups within a given dataset, we introduce a non-negative function, denoted as $\Delta(\cdot)$:

$$\Delta_{p,q}(E, f, \mathcal{D}) = \big\|E(f, \mathcal{D}_{G_p}) - E(f, \mathcal{D}_{G_q})\big\|, \tag{1}$$

where p, q are group indexes. This function is designed to quantify the disparity in explanations between any two distinct groups within the dataset.

By measuring the differences in explanations provided to each group pair, i.e., G_p and G_q, $\Delta(\cdot)$ offers a quantitative approach to understanding inequalities in the explanation process. We define the equalized explanation using the following criterion:

$$\max_{\substack{1\le p,q\le n,\\ p\neq q}} \Delta_{p,q}(E, f, \mathcal{D}) \le \epsilon, \tag{2}$$

where ϵ is the pre-defined threshold. If the criterion is met, we consider this model f to be fairly explainable.

B. Equalized Explanation via Feature Importance

In this study, we propose a new metric, Feature Importance Disparity (ΔFI), to evaluate the consistency of feature importance-based explanation across demographic groups. Without loss of generality, we select two arbitrary groups, G_p and G_q. This metric is designed to measure the dissimilarity in the distribution of FI between D_{G_p} and D_{G_q}.

$$\Delta FI = \sum_{j=1}^{n} \|FI_{j,G_p} - FI_{j,G_q}\| \tag{3}$$

Where FI_{j,G_p} represents the j-th feature's FI for Group G_p, calculated as $\sum_{i=1}^{m_{G_p}} \|\phi_j^{(i)}\|$, with m_{G_p} being the number of instances in Group G_p. ΔFI reflects the aggregated explanation disparity across instances and features out of two demographic groups.

V. Achieving Equalized Explanation through Data Reconstruction

Based on the proposed notion and metric for the explanation disparity, we develop a comprehensive framework aimed at mitigating these disparities in AI explanations. This framework is specifically designed to address the unequal distribution of clear and comprehensible explanations across different demographic groups, striving towards achieving a state of equalized explanation. This framework is composed of three integral modules: data reconstruction, equalization of explanations, and preservation of prediction performance. We introduce the modules in the following subsections.

A. Data Reconstruction Module

1) AutoEncoder Reconstruction: The AutoEncoder model is the core component of our data reconstruction module. The AutoEncoder $AE(X)$ encodes the original data X into a latent space Z, accurately representing its fundamental structure and distributional properties, then reconstructs $\hat{X}$ from the latent space Z. The AutoEncoder model is usually built on a multiple-layer neural network which is trained to minimize reconstruction errors. To maintain the consistency data patterns between $\hat{X}$ and X, we employ a trainable model $AE(X)$ within a loss function called $\mathcal{L}_{AE}$, expressed as,

$$\mathcal{L}_{AE} = \|X - AE(X)\|_2^2. \tag{4}$$

Optimizing $\mathcal{L}_{AE}$ is designed to guarantee a minimal discrepancy between $\hat{X}$ and X, delivering an accurate reconstruction of the original data's characteristics and distribution.

2) Prototype Representation Constraints: In addition to the conventional AutoEncoder, which provides a continuous representation of data in a latent space, we adopt prototype representation constraints that epitomize a typical instance of a specific class within a dataset and encapsulate its essential characteristics. Prototypes are vital for enhancing interpretability because they can discern the defining features of a particular class or cluster by pinpointing prototypes for each class [36]. In particular, we incorporate the prototype representation constraints as a loss function. For the samples conditioned by any arbitrary label l out of the original data X, we learn the prototype representation using the following steps.

- Obtain label-specific encoding (E_l). We extract the encoded representations corresponding to label l and denote them as $E_l \leftarrow \{Z_i | Y_i = l, 1 \le i \le m, 0 \le l \le L\}$;
- Calculate the centroid (c_l). The centroid of E_l is calculated as $c_l \leftarrow \frac{1}{|E_l|} \sum_{Z_i \in E_l} Z_i$;
- Computer instance distances to the centroid ($dist_i$). For each Z_i in E_l, the distance to the centroid c_l is computed and denoted as $dist_i \leftarrow ||Z_i - c_l||_2$;

- Get the prototypes (p_l) given label l. E_l is sorted based on the distances $dist_i$, and the top-H samples are selected. The prototype p_l for label l is then the mean of these top-H samples, $p_l \leftarrow \frac{1}{H}\sum_{h=1}^{H} Z_h$.

The summary of prototype calculation with details is presented in Algorithm 1.

Then, we formulate the prototype constraints $\mathcal{L}_{Proto}$ as:

$$\mathcal{L}_{Proto} = \sum_{l=1}^{C} \sum_{i:Y_i=l} \|Z_i - p_l\|_2^2 \tag{5}$$

Incorporating the loss function in Eq. 5 promotes the representation of categorical outcomes during the training and the interpretability of the reconstruction.

Algorithm 1 Prototype Calculation

procedure PROTOTYPECALCULATION($\mathcal{D}, encoder, H$)
  for each pair $< X_i, Y_i >$ in the dataset $\mathcal{D}$ **do**
    $Z_i \leftarrow encoder(X_i)$
  end for
  for each label l in the label domain $\mathcal{Y}$ **do**
    $E_l \leftarrow \{Z_i | Y_i = l; 1 \le i \le m, 0 \le l \le L\}$
    $c_l \leftarrow \frac{1}{|E_l|}\sum_{Z_i \in E_l} Z_i$
    $dist_i \leftarrow \|Z_i - c_l\|_2$ for each $Z_i \in E_l$
    Sort E_l based on $dist_i$ and select top-H samples
    $p_l \leftarrow \frac{1}{H}\sum_{h=1}^{H} Z_h$
  end for
end procedure

3) Data-specific Reconstruction Constraints: In addition to the data-agnostic reconstruction constraints, we deploy two data reconstruction constraints tailored to demographic datasets.

a) Feature-specific Constraints for Real Demographic Datasets: A customized reconstruction strategy provides granular control on each feature, guaranteeing that the reconstructed data remains meaningful and closely reflects real-world scenarios. For example, for features like "age" or "education level", it is logical that their values should only increase; for features like "income" or "working hour", increasing their values might be more challenging than decreasing it; and for features like "race" or "gender", they should remain unchanged to reflect their inherent nature.

b) Asymmetric Quadratic Function with Feature-specific Constraints: To achieve precise control on each feature, we introduce an asymmetric quadratic metric, "deviation", denoted as d. Especially, $d^j = \hat{X}^j - X^j$ is the difference between the reconstructed data $\hat{X}$ and the original data X on the j-th feature. The asymmetric quadratic loss function assigns distinct penalty weights for deviations per feature, ensuring that each feature's movement meets a specific direction. A positive d^j indicates a feature's value in $\hat{X}^j$ exceeds X^j, while a negative d^j signifies the opposite. The square of d^j guarantees differentiality, and any change in d incurs a cost. Mathematically, the asymmetric quadratic loss function is:

$$\begin{cases} a_1^j \times \|d^j\| & \text{if } d^j \ge 0 \quad \text{where } a_1^j, a_2^j > 0, \\ a_2^j \times \|d^j\| & \text{if } d^j < 0 \quad \text{and } j \text{ is the index of feature,} \end{cases}$$

where a_1^j and a_2^j are feature-specific weights.

We demonstrate three scenarios for choosing a_1^j and a_2^j weights, as shown below:

- $a_1^j < a_2^j$: represents feature j that is easier to increase than decrease, like "age" or "education level".
- $a_1^j = a_2^j$: denotes feature j that should remain unchanged, like "gender" or "race".
- $a_1^j > a_2^j$: indicates feature j that is harder to increase than decrease, like "income" or "working hour".

We also define the upper and lower boundaries of each feature in $\hat{X}$ based on the respective maximum and minimum values found in the original dataset X, thus ensuring that $\hat{X}$ captures the feature-wise characteristics.

c) Data-specific Reconstruction Constraints: We propose a customized reconstruction norm, which utilizes the asymmetrical quadratic function with feature-specific boundary restrictions. By applying parameters a_1^j and a_2^j to each feature, we can derive corresponding deviations d_j from the reconstructed $\hat{X}$, thereby establishing the customized loss function $\mathcal{L}_{DS}$:

$$\mathcal{L}_{DS} = \sum_{j=1}^{n} \left(a_1^j \cdot \max(0, d^j) + a_2^j \cdot \min(0, d^j) \right) \tag{6}$$

Here, a_1^j and a_2^j are the positive and negative deviation penalty coefficients for feature j.

B. Equalized Explainability Module

We introduce a module for achieving equalized explanations via $\mathcal{L}_{Equal}$ loss function. This module aims to guarantee that groups have comparable insights or knowledge when making decisions across groups with moderate ΔFI in Eq. (3).

1) Equalized Explainability via Feature Importance: By leveraging the concept of FI, we define

- $\mathcal{L}_{diff}$: the mean of the squared differences in FI for each feature across the groups, formalized as:

$$\mathcal{L}_{diff} = \frac{1}{n}\sum_{j=1}^{n} (FI_{j,G_p} - FI_{j,G_q})^2$$

- $\mathcal{L}_{avg}$: the absolute difference in FI for each feature across the groups, defined as:

$$\mathcal{L}_{avg} = \frac{1}{n}\sum_{j=1}^{n} \left| FI_{j,G_p} - FI_{j,G_q} \right|$$

where FI_{j,G_p} and FI_{j,G_q} represent the FI of the j-th feature in Group G_p and Group G_q, respectively. The final equilized explainability loss, $\mathcal{L}_{Equal}$, is the sum of these two losses.

$$\mathcal{L}_{Equal} = \mathcal{L}_{diff} + \mathcal{L}_{avg} \tag{7}$$

To achieve equalized explanations, $\mathcal{L}_{avg}$ and $\mathcal{L}_{diff}$ evaluate the disparity in FI between groups G_p and G_q through distinct approaches. $\mathcal{L}_{avg}$ calculates the average absolute difference in FI, assigning equal weight to disparities with all features. This loss function provides a straightforward measure of the overall consistency in FI among groups. Nevertheless, $\mathcal{L}_{diff}$ is designed to calculate the sum of squared differences in FI between groups. It inherently assigns a greater weight to large disparities. This loss aims to evaluate the distribution pattern of differences by picking out the most significant disparities in FI, emphasizing the extreme difference in FI between groups. Suppose $\mathcal{L}_{Equal}$ solely relies on $\mathcal{L}_{avg}$ loss. In that case, there is a possibility that there will be disproportionately affected by small disparities that are prevalent across the features, resulting in the model neglecting crucial features that exhibit considerable disparities between groups. Further, it has the potential to diminish the model's interpretive capability. Conversely, a $\mathcal{L}_{Equal}$ composed solely of $\mathcal{L}_{diff}$ may cause the model to excessively emphasize features with significant disparities, disregarding smaller, average disparities. When $\mathcal{L}_{Equal}$ includes $\mathcal{L}_{avg}$ and $\mathcal{L}_{diff}$, $\mathcal{L}_{avg}$ guarantees that the model acknowledges small disparities, while $\mathcal{L}_{diff}$ prevents the model from overreacting to large disparities. Therefore, the model may effectively decrease the average disparities with FI among different groups while simultaneously considering the overall distribution of these disparities, leading to a balanced and comprehensive interpretation.

$\mathcal{L}_{Equal}$ combines $\mathcal{L}_{avg}$ and $\mathcal{L}_{diff}$ to address widespread minor and major disparities across features. By utilizing the Cauchy-Schwarz Inequality, an upper constraint is naturally established on $\mathcal{L}_{avg}$, determined as $\sqrt{\mathcal{L}_{diff}/n}$. This constraint keeps the model does not unnecessarily compress disparities of FI to similar levels, preserving reliable interpretations of important features in the training process. When $\mathcal{L}_{avg}$ approaches its upper bound $\sqrt{\mathcal{L}_{diff}/n}$, it indicates large disparities in FI between groups. This constraint prevents a widespread decrease or fluctuation in FI, guaranteeing consistent focus on the appropriate distribution of FI during the training process.

C. Prediction Performance Preservation Module

To preserve the prediction performance, we introduce a loss function, $\mathcal{L}_{Pred} = f(\hat{X})$. This function guarantees that $\hat{X}$ has comparable predictive performance to the original dataset, which is crucial for preserving the model's predictive capability when promoting equalized explainability.

D. Objective Function of The Proposed Framework

Finally, we have the objective function $\mathcal{L}$ as a combination of all modules to contribute to the reconstruction of $\hat{X}$ and achieve equalized explainability across groups. The objective function as $\mathcal{L}$ is formulated as,

$$\mathcal{L} = \beta_1\mathcal{L}_{AE}+\beta_2\mathcal{L}_{Proto}+\beta_3\mathcal{L}_{DS}+\beta_4\mathcal{L}_{Equal}+\beta_5\mathcal{L}_{Pred} \quad (8)$$

It is important to highlight that the objective function in our framework is differentiable. This differentiability implies the function can be efficiently solved using gradient descent and seamlessly integrated with common deep learning packages, such as PyTorch and TensorFlow. As a result, our framework's compatibility with these packages ensures a more straightforward and accessible implementation process, allowing researchers and practitioners to leverage our approach within the familiar ecosystems of established deep learning tools.

VI. Experiments

In this section, we implement the proposed modular framework based on Captum [37], an interpretability framework designed for PyTorch. We evaluate the proposed methods on several real-world datasets and using various settings. As a comparison, we implement the baseline method using $\mathcal{L}_1$ and $\mathcal{L}_2$ regularization. The baseline objective is formulated as

$$\mathcal{L} = \beta_1\mathcal{L}_{AE}+\beta_2\mathcal{L}_{Proto}+\beta_3\mathcal{L}_{Reg}+\beta_4\mathcal{L}_{Equal}+\beta_5\mathcal{L}_{Pred}, \quad (9)$$

where $\mathcal{L}_{Reg} = \mathcal{L}_1 + \mathcal{L}_2 = \|\hat{X} - X\|_1^1 + \|\hat{X} - X\|_2^2$.

A. Experiment Datasets

1) Adult Dataset: The Adult dataset is extracted from the 1994 Census database, and the task is to predict if an individual's annual income exceeds \$50,000 based on census features. This dataset contains 48,842 samples and 14 attributes, including age, education, sex, occupation, income, marital status, etc. In this work, we consider the feature 'sex' to be a sensitive feature and divide the dataset into two demographic groups. The label is income, i.e., whether the income is larger than 50k.

2) Dutch Dataset: The Dutch dataset consists of 60,420 instances, each described by 12 attributes, including age, education level, economic status, household position, sex, etc. We consider 'sex' as the sensitive attribution. The ML task for the Dutch dataset is to predict occupation levels.

B. Experimental Results

1) Data Reconstruction Evaluation: To evaluate the reconstruction similarity between X and $\hat{X}$, we employ the dimensional reduction and clustering methods. For dimensional reduction, we adopt Uniform Manifold Approximation and Projection (UMAP), an efficient dimensionality reduction approach that preserves the global data structure in a lower-dimensional space [38]. For clustering, we adopt the K-means clustering algorithm to group data points together based on their similarity in feature space [39] and evaluate whether samples from X and $\hat{X}$ can be distinguished in the clustering.

We leverage UMAP to illustrate whether X and $\hat{X}$ exhibit similar distribution patterns within the dimensionality-reduced space, implying the likelihood of them sharing structural characteristics within the high-dimensional space. The original data samples X are denoted as blue dots, and reconstructed samples $\hat{X}$ are denoted by orange dots. Fig. 1 (a) and Fig. 1 (c) show the result of the baseline method, and Fig. 1 (b) and Fig. 1 (d) show the result of our proposed method. The comparison illustrates the proposed method generates fidelitous reconstructed data. This visual similarity is especially pronounced in the Dutch dataset, which indicates that

$\hat{X}$ successfully preserves the structural characteristics of the original data X, demonstrating a structural resemblance to X.

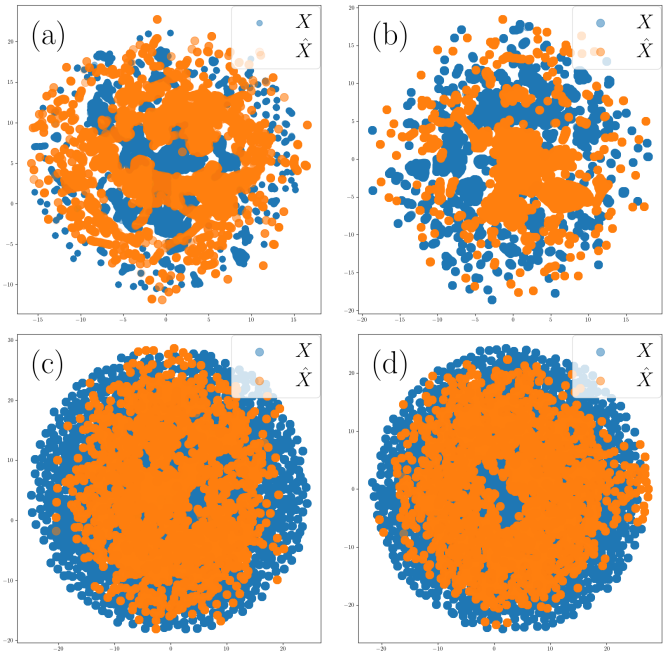

Fig. 1. UMAP Visualizations of Adult (a,b) and Dutch (c,d) Datasets. (a) and (c) Represent the Baseline Approach, While (b) and (d) Denote Our Proposed Approach.

UMAP maps the data samples in a low-dimensional space and visualizes the reconstruction similarity in the low-dimensional space. To compare the original data points and the reconstructed data points in the original space, we adopt K-means clustering and conduct a quantitative evaluation. We apply the K-means algorithm on the mix of original data X and the reconstructed data $\hat{X}$ with 2 clusters. We evaluate the distribution ratios of X and $\hat{X}$ across various cluster counts as a quantitative measure of their similarity. Table I reveals that the distribution ratio of X in most clusters is close to 50% in both the Adult and Dutch datasets, indicating a challenge in distinguishing X from $\hat{X}$ in feature space and suggesting a high degree of similarity between the two datasets, especially in the Adult dataset. While acknowledging the inherent limitations of the K-means, such as assumptions about data distribution and the necessity of choosing cluster count, we have applied it with caution in our analysis. This cautious utilization of K-means allows us to effectively quantify the similarity between X and $\hat{X}$, providing insights despite its inherent constraints.

2) Equalized Explainability: Given the prior statement emphasizing the importance of preserving careful balance in our modular design, we evaluate the performance of the equalized explainability module with a suitable ΔFI. In line with our comprehensive approach, we will not present a quantitative display of ΔFI, as it involves a trade-off between data fidelity and model accuracy alongside explainability.

TABLE I
CLUSTER LABEL IN DIFFERENT CLUSTERS FOR RECONSTRUCTION METHOD APPLIED IN THE ADULT DATASET.

Clusters n	2		3			4	
Cluster Label	0	1	0	1	2	0	1
Baseline(%)	50.00	50.00	49.99	50.00	50.10	49.87	50.00
Ours(%)	50.00	50.00	49.99	50.00	50.10	49.88	50.00
Clusters n	4 (Cont.)		5				
Cluster Label	2	3	0	1	2	3	4
Baseline(%)	50.13	52.59	49.93	50.00	52.97	50.29	50.31
Ours(%)	50.13	52.44	49.94	50.00	52.97	50.12	50.31

TABLE II
CLUSTER LABEL IN DIFFERENT CLUSTERS FOR RECONSTRUCTION METHOD APPLIED IN THE DUTCH DATASET.

Clusters n	2		3		
Cluster Label	0	1	0	1	2
Baseline(%)	51.78	48.36	45.74	48.23	57.58
Ours(%)	51.34	48.76	49.01	53.75	43.84

a) Adult: Feature importance in the original data X (Fig. 2(a)) reveals obvious gender-based explanation disparities between groups, illustrating original explanation differences between genders within the same model. The equalized explainability module efficiently mitigates gender-based explanation disparities in feature importance across groups within the regular baseline, improving equal explanations, as shown in Fig. 2(b). The result of the feature-specific reconstruction strategy is illustrated in Fig. 2(c). Although there are still some disparities in certain features, these differences are explained by intentional constraints implemented through the Asymmetric Quadratic Function with explicitly stated a_1 and a_2 for each feature for reflecting real-life scenarios.

b) Dutch: We utilize the Dutch dataset to evaluate the achievement of equalized explainability. Fig. 3(a) reveals group-level explanation disparities in the Dutch dataset, characterized by unequal feature importance across groups. effectively reduces feature importance disparities as shown in Fig. 3(b). In order to better reflect real-life scenarios, our proposed reconstruction approach prioritizes the orientation of each feature with specific-stated a_1 and a_2 for the Asymmetric Quadratic Function. Although some groups' disparity persists in these features, the overall trend is clearly toward achieving equal explanations, as shown in Fig. 3(c).

Both the baseline and our proposed reconstruction strategies effectively promote balanced explainability, providing consistent and fair explanations across groups. The baseline reconstruction approach excels in mitigating gender-based explanation disparities, whereas the feature-specific reconstruction module finely adjusts specific features while reflecting real-life scenarios, accomplishing balanced explanations.

3) Model Accuracy: We evaluate the trade-off between achieving data similarity, equalized explainability, and model predictive performance by comparing the accuracy of methods, as shown in Table III.

a) Adult: The accuracy of $f(X)$ is 84.50% using the original dataset. With the baseline method, we observe a

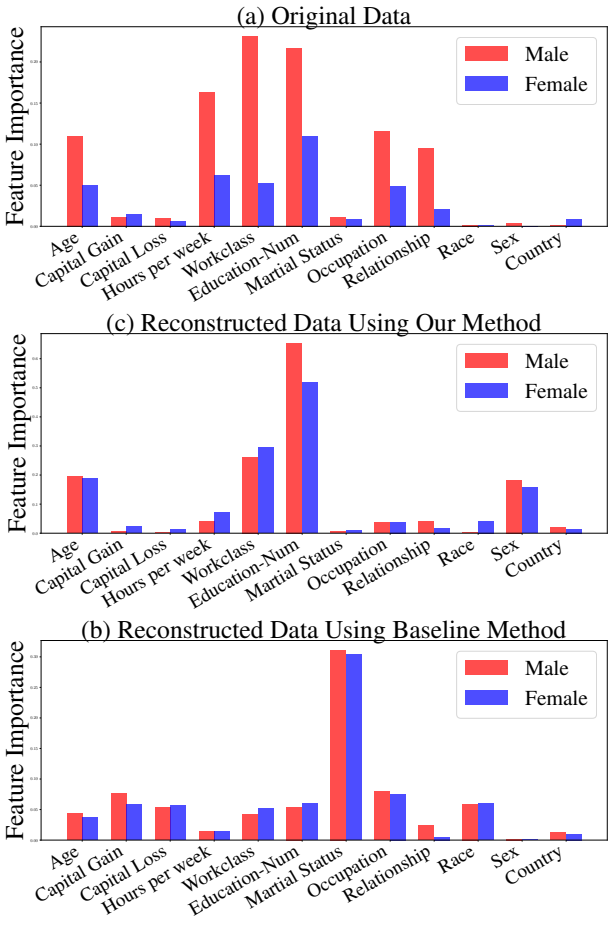

Fig. 2. Visualizations Equalized Explanation of Reconstructed Adult Dataset.

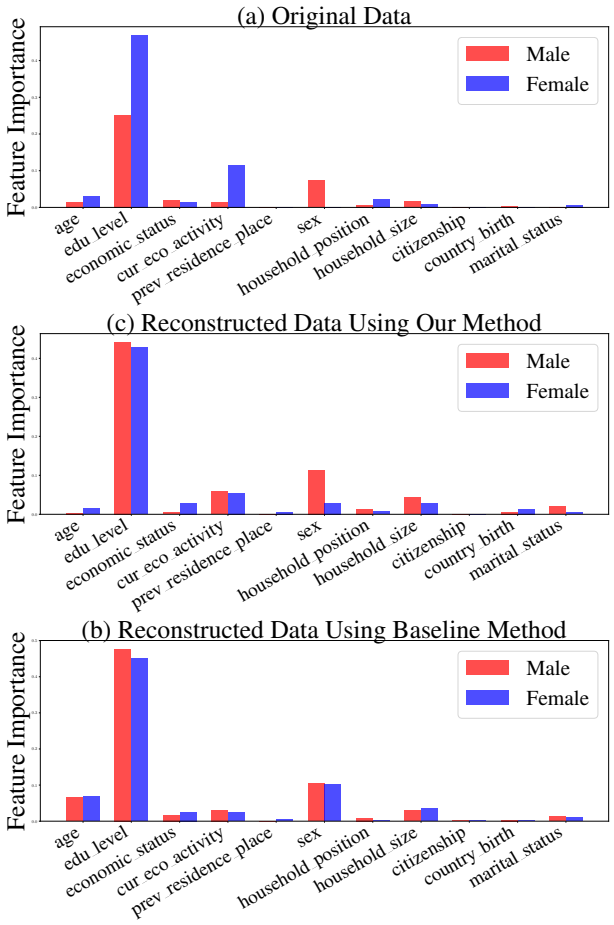

Fig. 3. Visualizations Equalized Explanation of Reconstructed Dutch Dataset.

reduction in accuracy to 80.10%. This reduction indicates that while equalizing explanations are compelling, it comes at the cost of some predictive accuracy. Our proposed method achieves an accuracy of 83.26%, closer to the binary model's performance, implying a better balance between preserving predictive performance and achieving equalized explainability.

b) Dutch: The accuracy of $f(X)$ is 83.40% on the original dataset. When we apply the baseline, there is a slight decrease in test accuracy to 82.14%, a trend also noted in the Adult dataset, which suggests that the process of equalizing explanations has a minor impact on model accuracy. However, in our method, the test accuracy drops slightly to 79.66%. The specific constraints applied in the customized reconstruction process considerably influence the model's predictive ability.

TABLE III
MODEL ACCURACY COMPARISON ON ADULT AND DUTCH DATASETS.

Dataset Accuracy	Classic	Baseline	Ours
Adult	84.50%	80.10%	83.26%
Dutch	83.40%	82.14%	79.66%

VII. CONCLUSION

This research centers on advancing fair explanations in machine learning, emphasizing the delivery of uniform and unbiased explanations across various demographic groups. Our proposed framework comprises three distinct modules, each targeting a crucial aspect: ensuring equality in explanations, maintaining data similarity, and preserving model accuracy. The design of our framework mirrors real-world complexities, offering significant insights and contributions to the field of fairness-aware machine learning and explainable artificial intelligence. Experimental evidence confirms that our approach effectively ensures equitable and consistent explanations across different groups. This achievement marks a substantial stride towards bolstering trust and transparency in the realm of fairness-aware machine learning systems.

REFERENCES

[1] C. Chang, S. Tan, B. J. Lengerich, A. Goldenberg, and R. Caruana, "How interpretable and trustworthy are gams?" in *The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2021, pp. 95–105.

[2] A. Holzinger, "From machine learning to explainable AI and beyond," in *Proceedings of the 11th International Joint Conference on Computational Intelligence, IJCCI 2019, Vienna, Austria, September 17-19, 2019*, 2019.

[3] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *IEEE Access*, pp. 42 200–42 216, 2020.

[4] Y. Zhao, Y. Wang, and T. Derr, "Fairness and explainability: Bridging the gap towards fair model explanations," in *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Washington, DC, USA, February 7-14, 2023*. AAAI Press, 2023, pp. 11 363–11 371.

[5] S. Bird, B. Hutchinson, K. Kenthapadi, E. Kiciman, and M. Mitchell, "Fairness-aware machine learning: Practical challenges and lessons learned," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*. ACM, 2019, pp. 3205–3206.

[6] S. Caton and C. Haas, "Fairness in machine learning: A survey," *CoRR*, 2020.

[7] L. Zhang, Y. Wu, and X. Wu, "A causal framework for discovering and removing direct and indirect discrimination," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia, August 19-25, 2017*. ijcai.org, 2017, pp. 3929–3935.

[8] Y. Wu, L. Zhang, and X. Wu, "On discrimination discovery and removal in ranked data using causal graph," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK*. ACM, 2018, pp. 2536–2544.

[9] Y. Wu, L. Zhang, X. Wu, and H. Tong, "PC-Fairness: A unified framework for measuring causality-based fairness," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, 2019, pp. 3399–3409.

[10] Y. Wu, L. Zhang, and X. Wu, "On convexity and bounds of fairness-aware classification," in *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*. ACM, 2019, pp. 3356–3362.

[11] L. Zhang, Y. Wu, and X. Wu, "Situation testing-based discrimination discovery: A causal inference approach," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*. IJCAI/AAAI Press, 2016, pp. 2718–2724. [Online]. Available: http://www.ijcai.org/Abstract/16/386

[12] W. Fleisher, "What's fair about individual fairness?" in *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*. ACM, 2021, pp. 480–490.

[13] Y. Wu and X. Wu, "Using loglinear model for discrimination discovery and prevention," in *2016 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016, Montreal, QC, Canada, October 17-19, 2016*. IEEE, 2016, pp. 110–119.

[14] T. Räz, "Group fairness: Independence revisited," in *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*. ACM, 2021, pp. 129–137.

[15] L. Zhang, Y. Wu, and X. Wu, "Achieving non-discrimination in data release," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*. ACM, 2017, pp. 1335–1344.

[16] S. Biswas and H. Rajan, "Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline," in *ESEC/FSE '21: 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Athens, Greece, August 23-28, 2021*. ACM, 2021, pp. 981–993.

[17] M. Wan, D. Zha, N. Liu, and N. Zou, "In-processing modeling techniques for machine learning fairness: A survey," *ACM Trans. Knowl. Discov. Data*, no. 3, pp. 35:1–35:27, 2023.

[18] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016, pp. 3315–3323.

[19] Y. Wu, L. Zhang, and X. Wu, "Counterfactual fairness: Unidentification, bound and algorithm," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, Macao, China, August 10-16, 2019*. ijcai.org, 2019, pp. 1438–1444.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *CoRR*, 2016.

[21] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022.

[22] R. Saleem, B. Yuan, F. Kurugollu, A. Anjum, and L. Liu, "Explaining deep neural networks: A survey on the global interpretation methods," *Neurocomputing*, pp. 165–180, 2022.

[23] M. R. Zafar and N. Khan, "Deterministic local interpretable model-agnostic explanations for stable explainability," *Mach. Learn. Knowl. Extr.*, no. 3, pp. 525–541, 2021.

[24] J. Dai, S. Upadhyay, S. H. Bach, and H. Lakkaraju, "What will it take to generate fairness-preserving explanations?" *CoRR*, 2021.

[25] A. Ortega, J. Fiérrez, A. Morales, Z. Wang, M. de la Cruz, C. L. Alonso, and T. Ribeiro, "Symbolic AI for XAI: evaluating LFIT inductive programming for explaining biases in machine learning," *Comput.*, no. 11, p. 154, 2021.

[26] A. Stevens, P. Deruyck, Z. V. Veldhoven, and J. Vanthienen, "Explainability and fairness in machine learning: Improve fair end-to-end lending for kiva," in *2020 IEEE Symposium Series on Computational Intelligence, SSCI 2020*. IEEE, 2020, pp. 1241–1248.

[27] J. M. Hickey, P. G. D. Stefano, and V. Vasileiou, "Fairness by explicability and adversarial SHAP learning," in *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2020, Ghent, Belgium, September 14-18, 2020, Proceedings, Part III*, ser. Lecture Notes in Computer Science. Springer, 2020, pp. 174–190.

[28] T. Begley, T. Schwedes, C. Frye, and I. Feige, "Explainability for fair machine learning," *CoRR*, 2020.

[29] J. Dai, S. Upadhyay, U. Aïvodji, S. H. Bach, and H. Lakkaraju, "Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations," in *AIES '22: AAAI/ACM Conference on AI, Ethics, and Society, Oxford, United Kingdom, May 19 - 21, 2021*. ACM, 2022, pp. 203–214.

[30] J. Chakraborty, K. Peng, and T. Menzies, "Making fair ML software using trustworthy explanation," in *35th IEEE/ACM International Conference on Automated Software Engineering, ASE 2020, Melbourne, Australia, September 21-25, 2020*. IEEE, 2020, pp. 1229–1233.

[31] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017, pp. 4765–4774.

[32] C. Yeh, C. Hsieh, A. S. Suggala, D. I. Inouye, and P. Ravikumar, "On the (in)fidelity and sensitivity of explanations," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019, pp. 10 965–10 976.

[33] U. Aïvodji, H. Arai, S. Gambs, and S. Hara, "Characterizing the risk of fairwashing," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021, pp. 14 822–14 834.

[34] A. Balagopalan, H. Zhang, K. Hamidieh, T. Hartvigsen, F. Rudzicz, and M. Ghassemi, "The road to explainability is paved with bias: Measuring the fairness of explanations," in *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*. ACM, 2022, pp. 1194–1206.

[35] Y. Zhou, S. Booth, M. T. Ribeiro, and J. Shah, "Do feature attribution methods correctly attribute features?" in *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 2022, pp. 9623–9633.

[36] A. V. Looveren and J. Klaise, "Interpretable counterfactual explanations guided by prototypes," in *Machine Learning and Knowledge Discovery in Databases. Research Track - European Conference, ECML PKDD 2021, Bilbao, Spain, September 13-17, 2021, Proceedings, Part II*, ser. Lecture Notes in Computer Science. Springer, 2021, pp. 650–665.

[37] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for pytorch," *CoRR*, 2020.

[38] L. McInnes and J. Healy, "UMAP: uniform manifold approximation and projection for dimension reduction," *CoRR*, 2018.

[39] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, no. 2, pp. 129–136, 1982.