A Canonical Data Transformation for Achieving Inter- and Within-group Fairness

Zachary McBride Lazri, *Graduate Student Member, IEEE,* Ivan Brugere, Xin Tian, Dana Dachman-Soled, Antigoni Polychroniadou, Danial Dervovic, and Min Wu, *Fellow, IEEE*

***Abstract*—Increases in the deployment of machine learning algorithms for applications that deal with sensitive data have brought attention to the issue of fairness in machine learning. Many works have been devoted to applications that require different demographic groups to be treated fairly. However, algorithms that aim to satisfy inter-group fairness (also called group fairness) may inadvertently treat individuals within the same demographic group unfairly. To address this issue, this paper introduces a formal definition of *within-group fairness* that maintains fairness among individuals from within the same group. A pre-processing framework is proposed to meet both inter- and within-group fairness criteria with little compromise in performance. The framework maps the feature vectors of members from different groups to an inter-group fair canonical domain before feeding them into a scoring function. The mapping is constructed to preserve the relative relationship between the scores obtained from the unprocessed feature vectors of individuals from the same demographic group, guaranteeing within-group fairness. This framework has been applied to the Adult, COMPAS risk assessment, and Law School datasets, and its performance is demonstrated and compared with two regularization-based methods in achieving inter-group and within-group fairness.**

I. Introduction

The deployment of machine learning (ML) models in sensitive domains—including criminal justice, healthcare, advertising, and finance— is increasingly raising potential legal and fairness concerns [1]. In a number of studies involving sensitive information, bias was detected in ML models that were used to make decisions. For example, racial discrimination was detected in the COMPAS recidivism assessment model [2] and Google's Ads model [3]; gender discrimination was found in Amazon's recruiting model [4]; and skin tone bias was found in models used to detect melanoma in images [5]. Because the results produced by ML models rely on the data on which they are trained, any pre-existing societal biases could be embedded in data and may be inherited by these models unless proper mitigants are applied. There is a growing body of research focused on combating unfairness in ML models from which a number of definitions of fairness have been advanced, and many tools have been developed to satisfy them. Two of the most prominent categories of fairness definitions include: (1) group fairness [6]–[12] and (2) individual fairness [13]–[19]. The output of an ML model is a distribution of scores, between 0 and 1, to which a threshold is applied to produce decisions. Group fairness definitions focus on ensuring that different demographic groups are treated equally, meaning that these decisions are unbiased based on group affiliation. Individual fairness definitions aim to guarantee that individuals with similar feature values are treated similarly, meaning that an ML model assigns them similar scores.

As multiple works have shown [7], [20], [21], achieving a universal notion of fairness is impossible. Compromises must be made when it is desirable to satisfy multiple conflicting notions of fairness. How to balance these trade-offs depends on the particular application. For example, unless a discrete exception applies, it is prohibited under U.S. fair lending laws[1] to take into consideration in any aspect of a credit transaction protected attributes, including, but not limited to, race, sex, and religion [22]. Thus, financial institutions undertake qualitative and quantitative fair lending analyses to identify, assess, and mitigate any associated fair lending risks [23]. The European Union also prohibits discrimination on the basis of a similar set of protected attributes under Title III (Equality), Article 21 of the European Union Charter of Fundamental Rights [24].

A variety of approaches have been proposed to preserve group-based fairness [6], [8], [25]–[29]. To offset biases reflected in the score distributions produced by a model for different demographic groups, Chen et al. construct an optimization problem to solve for separate scoring thresholds for each group to produce fair decisions [26]. Hardt et al. [6] and Hsu et al. [25] approach this problem by equalizing the output scores of different groups by constructing post-processing transformations that are applied to the original distribution of scores produced by an ML model. However, these approaches assume that sensitive group affiliation information is directly available to their decision-making models for processing in the testing stage of the ML process, which is often not the case and, under some circumstances, may be prohibited by applicable law. That is to say, given the restrictive, highly sensitive nature of protected attributes, it is common to restrict the model development team's access to those attributes. Instead, that team

Manuscript received 21 October 2023; revised 20 February 2024 and 02 May 2024; accepted 21 May 2024. This article was recommended for publication by Associate Editor Edgar Weippl upon evaluation of the reviewers' comments. *(Corresponding author: Zachary McBride Lazri)*

Zachary Lazri, Dana Dachman-Soled, and Min Wu are with the University of Maryland, College Park (e-mail: zlazri@umd.edu; danadach@umd.edu; minwu@umd.edu).

Ivan Brugere, Antigoni Polychroniadou, and Danial Dervovic are with JPMorgan A.I. Research (e-mail: ivan.brugere@jpmchase.com; antigoni.polychroniadou@jpmorgan.com; danial.dervovic@jpmchase.com).

Xin Tian was with the University of Maryland, College Park when this work was started and is now with Meta, Menlo Park, CA 94025 USA (e-mail: xtian17@terpmail.umd.edu)

[1]The Equal Credit Opportunity Act and its implementing regulation, Regulation B, prohibit creditors from considering any protected characteristic (such as race, gender, or age, etc.) in any aspect of a credit transaction, unless an express exception applies.

develops the models that will be integrated into the institution's decision-making process, while a separate compliance team with access to the sensitive data evaluates the models for fairness. As such, the models do not make direct use of the sensitive attributes, but those attributes are available to the compliance team that tests the models for fairness.

Regularization-based methods provide a potential solution to this bottleneck [8], [26]–[29]. Regularization is a common method used in many AI/ML applications [30]–[32] to improve model generalization or teach models to satisfy additional constraints. In fairness-aware AI/ML applications, regularization is often used to teach a model to enforce group fairness by decorrelating its decisions with respect to the sensitive attribute during the training process. Yet these methods do not directly include the sensitive attribute as an input to the model, meaning that it is only used in the testing phase of the ML process, not the training phase. However, a lack of heterogeneity in the feature values of individuals in different demographic groups may impede a model from learning how to equalize different demographic groups' score distributions without disparately impacting similar individuals within each group. Hence, asking a classifier to approximate statistical parity between different groups may cause it to badly violate parity among the individuals within a group [33].

This highlights an important consideration that must be addressed in dealing with problems in which group fairness is a hard constraint—solutions that satisfy group fairness must treat individuals *within* each group fairly with respect to each other. Two issues must be addressed to satisfy this requirement: (1) a definition must be constructed to articulate what it means for individuals within the same group to be treated fairly (within-group fairness) and (2) a solution must be developed to satisfy within-group fairness without violating group fairness. This paper addresses these two issues, particularly under the constraining scenario in which the sensitive attribute is not allowed to be directly provided to a decision-making model in the testing phase of the ML process. In the remainder of this paper, we will use the terminology *inter-group fairness* in place of group fairness to avoid any potential confusion between this notion and the notion of *within-group fairness*.

To achieve our two aforementioned goals, we take advantage of a key insight: though an ML model is prohibited from directly using a sensitive attribute in the testing phase of the ML process, it may be available to *pre-process the data* prior to providing it to the ML model for decision-making. Referring to our example on fair lending, this means that a compliance team could pre-process the data it oversees before providing it to the machine learning team for use. Thus, in this paper, we introduce a pre-processing framework for simultaneously satisfying inter- and within-group fairness. The core idea of our pre-processing framework is to devise a mapping that (1) removes the correlation between different groups' feature distributions while (2) ensuring that the ordering of scores produced by an ML model for individuals within the same group is preserved for the pre-processed feature distribution. We adopt a stringent notion of fairness—the threshold invariant form of demographic parity [10]—for our inter-group fairness definition [8] and introduce a new definition of *within-group fairness*, which is used to devise a measure for fairness assessment. Furthermore, we compare this approach against regularization-based training methods to verify its utility. To summarize, this paper provides the following contributions:

- We introduce a new definition of fairness, termed *within-group fairness*, which ensures that individuals within the same group are treated fairly with respect to each other.
- We provide a metric for measuring within-group fairness.
- We introduce a novel framework that uses pre-processing to achieve both inter- and within-group fairness.
- We experimentally verify that our pre-processing framework can achieve both inter- and within-group fairness with little compromise in accuracy, while empirically demonstrating that regularization methods struggle to satisfy both without compromising model performance.

The remainder of this paper is organized as follows. In Section II, a summary of basic fairness-related concepts relevant to this work is provided. In Section III, the problem setup is formulated, while in Section IV, the pre-processing framework is described along with the regularized training approach used for comparison. Experimental results are presented in Section V to quantify each method's accuracy and ability to achieve inter-group fairness and within-group fairness. A discussion of the limitations associated with this paper is provided in Section VI. Finally, the paper is concluded in Section VII.

II. Background and Related Work

A. Fairness definitions

Fairness definitions may be grouped into three main categories, which include inter-group, individual, and subgroup fairness [34]. Inter-group fairness is measured by prediction performance parity across different demographic groups. To deal with multifaceted issues of bias and discrimination, various inter-group fairness notions have been proposed in the literature, including demographic parity [13], equalized odds [6], test fairness [7], and threshold invariant fairness [8]. However, solutions that only enforce inter-group fairness may produce outcomes that are blatantly unfair from the perspective of an individual [35]. Individual fairness aims to ensure that similar prediction performance is achieved for similar individuals with respect to the same task [13]. To bridge the gap between inter-group and individual fairness, Kearns et al. [33] propose the notion of subgroup fairness. This definition enforces inter-group fairness on a large collection of subgroups defined over combinations of protected attribute values.

In this paper, we define a new notion of fairness in Section III-C, termed *within-group* fairness, which aims to preserve the scoring hierarchy of individuals from the same demographic group before and after accounting for inter-group fairness. This idea is related to the sub-group fairness definition proposed by Kearns et al. [33] in the special case where each individual in a group is considered a sub-group.

B. Design of fairness-aware algorithms

Many methods have been studied to achieve some notion of fairness in machine learning, which can be categorized

accordingly: (1) pre-processing, (2) in-processing, and (3) post-processing [34]. Pre-processing approaches remove the underlying biases in the raw feature data prior to providing it to an ML model for training or decision-making. Kamiran et al. [10] propose a pre-processing method that "massages" a set of training labels to remove bias in it prior to training a model. In their method, a ranker is used to select which labels to alter while minimizing deterioration in prediction accuracy. In-processing methods introduce fairness-aware regularizers into the objective function during the training process, which aim to strike a balance between maximizing accuracy and minimizing unfairness [12]. Calders et al. [11] devised a regularizer that requires a trained classifier to make decisions independent of the sensitive attribute, while Zafar et al. [27], [28] enforce demographic parity and equalized odds constraints in the classifier training process. Chen et al. [8] designed a fairness loss function that aims to equalize the score distributions of different demographic groups. Post-processing achieves fairness by reassigning the scores produced by the initial black-box model by applying a function to the original output scores [9]. Other post-processing methods [36] can mitigate bias by identifying unfair features via attention mechanisms and manipulating the corresponding attention weights.

In this paper, we propose a novel pre-processing framework that maps the feature vectors of different demographic groups to a *canonical feature distribution* in which inter- and within-group fairness can simultaneously be achieved with low cost to a model's performance. For a comprehensive review of algorithmic fairness that provides a discussion of the prominent causes of unfairness, algorithms, definitions, and datasets available in this field of study, we refer the reader to Pessach et al. [37].

III. Problem Setup

In this section, we outline the problem that we aim to solve. Our end goal is to devise an algorithm that can achieve both inter- and within-group fairness without compromising predictiveness. To achieve this goal, we now introduce the inter- and within-group fairness notions that we want our model to satisfy and the measures, Δ_{TIDP} and Δ_{WGF}, that quantify how well a method is able to satisfy inter- and within-group fairness, respectively.

A. Notation

Following Dwork et al., we refer to *individuals* as the objects that we aim to classify [13]. We refer to the set of individuals as $\mathcal{I}$. Each individual, $i \in \mathcal{I}$, has a corresponding *feature vector*, $\mathbf{f}_i \in \mathbb{R}^k$, the elements of which represent values associated with features from the *feature set*, $\mathcal{F} \triangleq \{f_1, \ldots, f_k\}$. We refer to a subset of features, $\mathcal{A} \subseteq \mathcal{F}$ as *sensitive* if we are prohibited from discriminating against individuals based on the values of these features. Thus, no sensitive feature values are included in feature vectors. Each individual also has a *label*, $y \in \{0, 1\}$, associated with him or her. A *dataset* of N individuals, $X \in \mathbb{R}^{N \times k}$, is a matrix in which each row represents an individual's feature vector. The primary task of binary classification is to predict the labels of individuals using a scoring function. A scoring function, $s_{\boldsymbol{\theta}} : \mathbb{R}^k \to [0, 1]$, is defined as a mapping of a feature vector to a score between 0 and 1, inclusive. Given a threshold, t, the estimator associated with a scoring function, $s_{\boldsymbol{\theta}}$, is represented by the function:

$$\hat{Y} = \begin{cases} 0, & s_{\boldsymbol{\theta}}(\mathbf{f}) \leq t \\ 1, & s_{\boldsymbol{\theta}}(\mathbf{f}) > t, \end{cases} \tag{1}$$

and is used to predict the label associated with a given individual. A scoring function is trained on data to optimize its parameters to satisfy an objective and tested on separate data to ensure the results it produces are generalizable. We refer to the training dataset as $\mathbf{X}_{tr}$ and the testing dataset as $\mathbf{X}_{ts}$. A *loss function*, $L_{\boldsymbol{\theta}}$, is used to quantify the ability of the scoring function to produce scores that satisfy the objective. A *learning algorithm* optimizes the weights of the scoring function, $\boldsymbol{\theta}$, according to $L_{\boldsymbol{\theta}}$.

A *group*, $g \subseteq \mathcal{I}$, is a collection of individuals that share the same values for the sensitive features in $\mathcal{A}$. The set of all groups is denoted $\mathcal{G}$. We use numerical superscripts to specify group membership. For example, $X^{(i)}$ refers to the subset of the dataset X belonging to group i. The omission of superscripts indicates that we are referring to population information. Lastly, $d_{\mathcal{X}}$ represents a distance measure over the set $\mathcal{X}$.

B. Inter-group Fairness

A variety of definitions have been proposed to capture inter-group fairness. Two of the most prominent definitions include demographic parity (DP) and equalized odds (EO), and we focus on DP in this paper.

Definition 1. *(Demographic Parity) An estimator, $\hat{Y}$, satisfies demographic parity for a binary feature, $A \in \{0, 1\}$, if $P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$.*

An important observation to make about this definition is that it is threshold dependent. That is, a particular classifier may achieve DP for a given threshold, but not for all thresholds in the range $[0, 1]$ when the group score distributions produced by a scoring function deviate from each other. In many applications, this may lead to fairness issues. As described in Section I, for example, model development teams may not have access to sensitive demographic information associated with the data. Without such information, they may produce models with unequal output score distributions across groups, raising potential fairness concerns. Thus, we wish to ensure that a model produces score distributions that are equal across groups without consideration of any protected attributes. This intuition is encapsulated in the more strict definition of *threshold-invariant fairness* defined by Chen et al. [8] (also coined *strong demographic parity* by Jiang et al. [29]), and has been employed in a variety of works to enforce inter-group-fair classification [8], [26], [29], [38], [39] and regression [38], [40]. Thus, we use this definition for the inter-group fairness notion in our analysis:

Definition 2. *(Threshold Invariant Fairness) Threshold Invariant Demographic Parity (TIDP) is achieved when DP is satisfied, independent of the decision threshold t.*

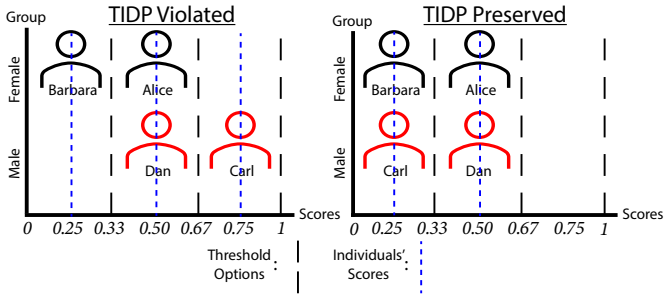


Fig. 1: Illustration of Fairness in Example 1.

To measure how well a classifier preserves TIDP, we can take the average of the Calder-Verwer (CV) score [11] for every possible threshold value $t \in [0, 1]$, where the CV scores for DP are given by: $|P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1)|$. We use Δ_{TIDP} to refer to this resulting average value. The framework that we propose in this paper aims to achieve TIDP while maintaining within-group fairness.

C. Within-group Fairness

In the previous subsection, we discussed measures of inter-group fairness, which are blind to the treatment of individuals. To understand the negative side-effects that may arise from only taking inter-group fairness into account, we provide the following simple motivating example.

Example 1. *Consider four individuals, Alice and Barbara, who are female, and Carl and Dan, who are male. Assume that all other sensitive attributes are the same for all individuals. According to the information in their loan applications, they receive the following loan approval scores, respectively: 0.50, 0.25; 0.75, 0.50. Since the distribution of scores between the male and female groups is different, a loan officer adjusts Carl's score to a value of 0.25 so that TIDP-based inter-group fairness is satisfied since for any threshold, the same number of individuals from each sex score above it. (see Figure 1). However, assuming all applicants were scored according only to non-sensitive features, we see that Carl's outcome is unfair since he has a stronger application than Dan.*

This simple example highlights an important consideration that must be taken into account in the fairness process: individual fairness must still be preserved within each group. Clearly, a more individually fair way to achieve TIDP in Example 1 would be to provide Carl with a score of 0.50 and Dan with a score of 0.25. While the specific example we provide here deals with finance, such considerations could potentially impact recidivism assessment [41], hiring decisions [42], credit approvals [43], and any other real-world classification applications in which bias has been detected in the score distributions produced by an ML model. Thus, we introduce a new fairness definition, *within-group fairness*, which takes advantage of the insight provided in this example: the most accurate scores are the fairest for individuals that belong to the same group. This suggests that the scores provided by a *baseline* model, which does not explicitly account for any notion of fairness, yield the fairest scores when comparing individuals within the same group.

The inspiration for our definition comes from Dwork et al.'s definition of individual fairness, which states that similar individuals should be treated similarly [13]. Definition 3, below, provides the formal definition of individual fairness.

Definition 3. *(Individual Fairness) A scoring function $s : \mathbb{R}^k \to [0, 1]$ satisfies individual fairness if for any two individuals $i, j \in \mathcal{I}$, $|P(\hat{Y}_i = y) - P(\hat{Y}_j = y)| \leq \epsilon; \quad if \;\; d_{\mathcal{I}}(i, j) \approx 0.$*

With this in mind, we provide a few definitions that build toward our definition of within-group fairness.

Definition 4. *(Signed Distance Function) The signed distance function for $x, y \in \mathcal{X}$ is given by:*

$$\phi_{\mathcal{X}}(x, y) = \begin{cases} d_{\mathcal{X}}(x, y), & x \leq y \\ -d_{\mathcal{X}}(x, y), & x > y \end{cases} \tag{2}$$

Definition 5. *(Individual Fairness Across Mappings) A mapping $Q : \mathbb{R}^k \to [0, 1]$ satisfies individual fairness with respect to a different mapping $R : \mathbb{R}^k \to [0, 1]$ if for any two individuals, $i, j \in \mathcal{I}$,*

$$|\phi_{[0,1]}(Q(\mathbf{f}_i), Q(\mathbf{f}_j)) - \phi_{[0,1]}(R(\mathbf{f}_i), R(\mathbf{f}_j))| \leq \epsilon. \tag{3}$$

Definition 6. *(Within-group Fairness Across Mappings) A mapping $Q : \mathbb{R}^k \to [0, 1]$ is said to satisfy within-group fairness with respect to a different mapping $R : \mathbb{R}^k \to [0, 1]$ if for any group $g \in \mathcal{G}$, individual fairness across models is satisfied between any two individuals, $i, j \in g$. That is,*

$$|\phi_{[0,1]}(Q(\mathbf{f}_i), Q(\mathbf{f}_j)) - \phi_{[0,1]}(R(\mathbf{f}_i), R(\mathbf{f}_j))| \leq \epsilon, \quad \forall g \in \mathcal{G}; \forall i, j \in g \tag{4}$$

Intuitively, our within-group fairness definition states that the scores provided by one mapping are fair with respect to another if the distance between any two individuals' scores is relatively preserved and the ordering of their scores does not change across models. In the algorithm proposed in Section IV, we consider one of the mappings to be the baseline scoring function. The second mapping is considered to be a pre-processing transformation $T : \mathbb{R}^k \to \mathbb{R}^k$ composed with the baseline scoring function. Since we design T to remove the bias between the distribution of feature vectors for different demographic groups, if Definition 6 is satisfied, then our framework achieves both inter- and within-group fairness. To quantify how well a mapping, Q, is able to preserve within-group fairness with respect to another mapping, R, we use the following equation:

$$\Delta_{WGF} = \frac{2 \sum_{g \in \mathcal{G}} \sum_{\substack{i,j \in g \\ i<j}} \mathbf{1}\{|\phi_{[0,1]}(Q(\mathbf{f}_i), Q(\mathbf{f}_j)) - \phi_{[0,1]}(R(\mathbf{f}_i), R(\mathbf{f}_j))| > \epsilon\}}{\sum_{g \in \mathcal{G}} N_g(N_g - 1)}, \tag{5}$$

where $\mathbf{1}$ represents the indicator function, ϵ may be a user-specified parameter quantifying how much within-group unfairness we are willing to tolerate, and N_g is the number of individuals in group g. Δ_{WGF} accounts for the change in signed distance between every pair of scores within each demographic group that results from using mapping R in place of mapping Q. If the signed distance between the scores of

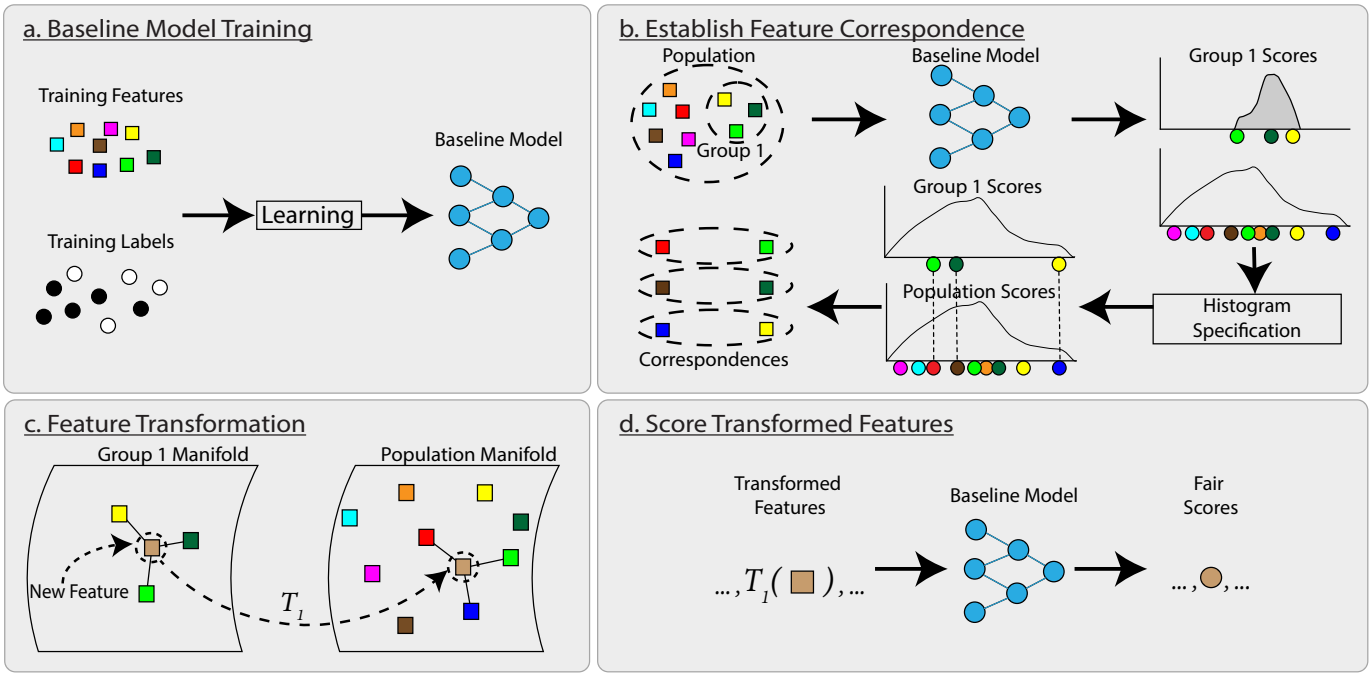


Fig. 2: Overview of the pre-processing framework for achieving inter- and within-group fairness.

any pair of individuals produced by mapping Q is drastically different from the baseline model (i.e. greater than ϵ), this means these individuals are being treated drastically differently. That is, within-group fairness is violated for this pair of individuals. Hence, the numerator of Equation 5 counts all the pairs of individuals for which an inter-group-fair model violates within-group fairness, while the denominator counts how many pairs of individuals exist in each group. Thus, Δ_{WGF} provides the percentage of individuals for which within-group fairness is violated.

IV. Framework for Achieving Group and Within-Group Fairness

In this section, we outline the proposed pre-processing framework for achieving inter- and within-group fairness. We also describe an in-processing algorithm in which fairness is encouraged through the loss function during training with the use of regularizers, which we use to compare against our framework.

A. Proposed Pre-processing Framework

Figure 2 illustrates the proposed pre-processing algorithm. In Phase I, we train an ML scoring function on training data. This scoring function may produce disparate score distributions for different groups, violating inter-group fairness. To deal with this issue, we construct transformations that map the feature vectors of individuals from different groups to a domain in which the bias between these groups is removed, which we refer to as the *canonical domain*. The goal is to design these mappings such that when the baseline scoring function is applied to the transformed feature vectors of any group, the output score distribution resembles the original un-transformed *population* score distribution. Each group's transformation is constructed by creating a *correspondence* between the feature vectors within the group and the feature vectors in the entire population. The details of each phase of the system are presented in the ensuing subsections.

Phase I: Baseline Model Training: Assume we have a population training dataset, $\mathbf{X}_{tr} \in \mathbb{R}^{N_{tr} \times k}$ with associated labels $\mathbf{y} \in \{0,1\}^{N_{tr}}$ consisting of N_{tr} individuals. We construct a scoring function of the form $s_{\boldsymbol{\theta}}(x) = \sigma(m_{\boldsymbol{\theta}}(x))$, where $m_{\boldsymbol{\theta}} : \mathbb{R}^k \to \mathbb{R}$ is an ML model (e.g. logistic regression, support vector machine, neural network, etc.) and $\sigma : \mathbb{R} \to [0,1]$ is the sigmoid function. We call this scoring function the *baseline model*. The baseline model is designed to maximize *accuracy* according to the training data using a standard (sub-) differentiable loss function, $L_{\boldsymbol{\theta}}$, (e.g., cross-entropy, hinge, etc.). Finally, we use a learning algorithm (e.g., gradient descent, stochastic gradient descent, etc.) to optimize the model weights, $\boldsymbol{\theta}$.

Phase II: Establish Feature Correspondence: Once the baseline model, $s_{\boldsymbol{\theta}}$, is obtained from Phase I, we split the population training data into groups using the sensitive attribute, where we assume these groups are mutually exclusive. Specifically, we decompose $\mathbf{X}_{tr}$ into group training datasets $\mathbf{X}_{tr}^{(i)}, i = 1, \ldots, |\mathcal{G}|$. The baseline model is applied to each group's training dataset and the population training dataset to obtain the score vectors $\mathbf{s}_{tr}^{(i)}, i = 1, \ldots, |\mathcal{G}|$ and $\mathbf{s}_{tr}$, respectively. We then map the distribution of scores associated with each group to the population score distribution. To do this, we apply histogram specification [44], which is a distribution matching technique commonly applied in image processing to transform the histogram of the pixel values in a given image to match a specified histogram. In our case, the histograms of the group and population score distributions are computed and histogram

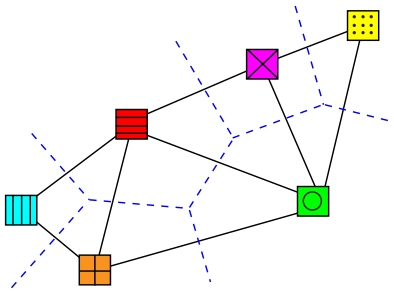


Fig. 3: Dual relationship between Voronoi diagram and Delaunay Triangulation.

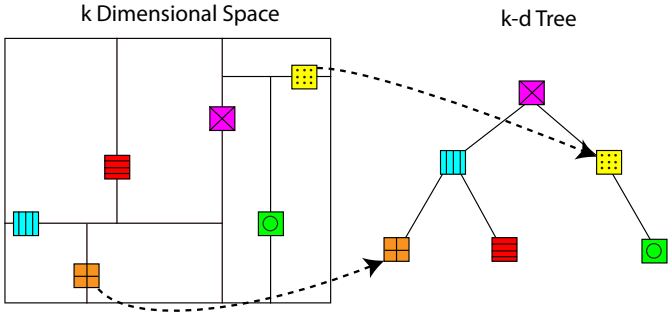


Fig. 4: Representation of a k-d tree.

matching is applied to each group's distribution to match it to the population distribution. We denote these score vectors by $\hat{\mathbf{s}}_{tr}^{(i)}, i = 1, \ldots, |\mathcal{G}|$. Because the scores in $\hat{\mathbf{s}}_{tr}^{(i)}$ and $\mathbf{s}_{tr}$ are approximately equally distributed, each score in $\hat{\mathbf{s}}_{tr}^{(i)}$ should be close in distance to some score in $\mathbf{s}_{tr}$. We pair group and population feature vectors if their scores are sufficiently close, calling such pairs *feature correspondences*. Below, we provide the formal definition of a feature correspondence.

Definition 7. *(Feature Correspondence) Let $a \in \hat{\mathbf{s}}_{tr}^{(i)}$ and $b \in \mathbf{s}_{tr}$ be the scores associated with the n^{th} feature vector of group i, $\mathbf{f}_n^{(i)}$, and the l^{th} feature vector of the population, $\mathbf{f}_l$, respectively. We say $\mathbf{f}_n^{(i)}$ and $\mathbf{f}_l$ form a correspondence, denoted $\mathbf{f}_n^{(i)} \sim \mathbf{f}_l$, if for any $c \in \mathbf{s}_{tr}$, such that $c \neq b$, the following inequality holds $|a - b| \leq |a - c|$.*

Our goal is to exploit these feature correspondences to create a transformation that maps feature vectors from each group to the canonical population domain. This idea is illustrated in Figure 2b. Critically, histogram specification is a monotonically increasing transform on the scores. As a result, when we use these correspondences to construct our group transformations, it should follow that the scores that the baseline model produces for the transformed features preserve the ordering of the scores that the baseline model produced for the *original* feature vectors. As a result, our notion of within-group fairness should be preserved.

Phase III: Mapping Construction: Phase III uses the feature correspondences found in Phase II to construct a mapping between each group domain and the population domain for individuals not seen in the training data. Our approach assumes that each group's data lie on a manifold in $\mathbb{R}^k$ and that the training data sample this manifold densely enough to provide a faithful representation of it. Hence our goal is to construct a more general correspondence between each group's manifold to the population manifold. To achieve this goal, we note that the field of computational geometry provides a variety of candidate data structures that may be used for manifold representation to construct these mappings. Two notable candidates include: (1) the Delaunay Triangulation (DT), which serves as the dual counterpart of the Voronoi Diagram (VD), and (2) the k-d tree.

Given a set of points in $\mathbb{R}^k$ called *cell cites*—which are given by the training feature vectors in our context, a VD induces a cell decomposition of space where each cell represents all of the points closest to a particular cell site. A DT is a planar subdivision of $\mathbb{R}^k$, constructed by drawing edges connecting any two cell sites that share an edge between their VD cells, and as a result, the closest pair of sites in a VD is represented as neighbors in the DT. In this way, each cell in the VD corresponds to a vertex in the dual complex. Figure 3 illustrates this primal-dual relationship.

The k-d tree is a space-partitioning data structure used to organize points in k-dimensional space. In this data structure, each leaf node represents one data point, while each non-leaf node represents a splitting hyperplane used to generate two half-spaces. The points in each half-space correspond to the leaves of a given node's subtree. This structure is illustrated in Figure 4.

The utility of each of these data structures lies in their ability to represent new feature vectors as a weighted combination of the closest neighboring training feature vectors in $\mathbb{R}^k$ used to construct each data structure. This can be done by constructing $|\mathcal{G}|$ DT or k-d tree data structures using the feature vectors in the training sets $\mathbf{X}_{tr}^{(i)}, i = 1, \ldots, |\mathcal{G}|$ to represent each group's individual manifold. When a new feature vector for group i, $\mathbf{f}_{new}^{(i)}$, is to be pre-processed, we can represent it by a weighted average of feature vectors to which it is closest, as illustrated in Figure 2c, which are obtained by traversing the given data structure to find the closest neighbors to represent the new feature. In particular, mapping new feature vectors from the group domain to the population domain can be done as follows. Without loss of generality (WLOG), let $\{\mathbf{f}_n^{(i)}\}_{n=1}^k$ represent the k nearest neighbors of $\mathbf{f}_{new}^{(i)}$, belonging to group i. The Euclidean distances between $\mathbf{f}_{new}^{(i)}$ and the elements of $\{\mathbf{f}_n^{(i)}\}_{n=1}^k$ are represented by $\{d_n^{(i)}\}_{n=1}^k$. Inspired by the algorithm presented by Dongsheng et al. [45], we construct a set of weights $\{d_n^{(i)-1} / \sum_{l=1}^k d_l^{(i)-1}\}_{n=1}^k = \{w_n^{(i)}\}_{n=1}^k$ associated with each element in $\{\mathbf{f}_n^{(i)}\}_{n=1}^k$ that reflect how closely they resemble the new feature vector in the group domain. In the population domain, WLOG, we obtain the corresponding feature vectors $\{\mathbf{f}_n\}_{n=1}^k$, where $\mathbf{f}_n^{(i)} \sim \mathbf{f}_n$, and use these weights to construct their new feature vectors in the population domain $\mathbf{f}_{new} = \sum_{l=1}^k w_l f_l$. We use the notation $T^{(i)} : \mathbb{R}^k \to \mathbb{R}^k$ to describe the mapping from a feature vector from group i's domain to the population domain. Hence, given a test dataset, $\mathbf{X}_{ts} \in \mathbb{R}^{N_{ts} \times k}$ of N_{ts} individuals, we can decompose it in group test sets $\mathbf{X}_{ts}^{(i)}, i, ..., |\mathcal{G}|$, obtain the canonical group test

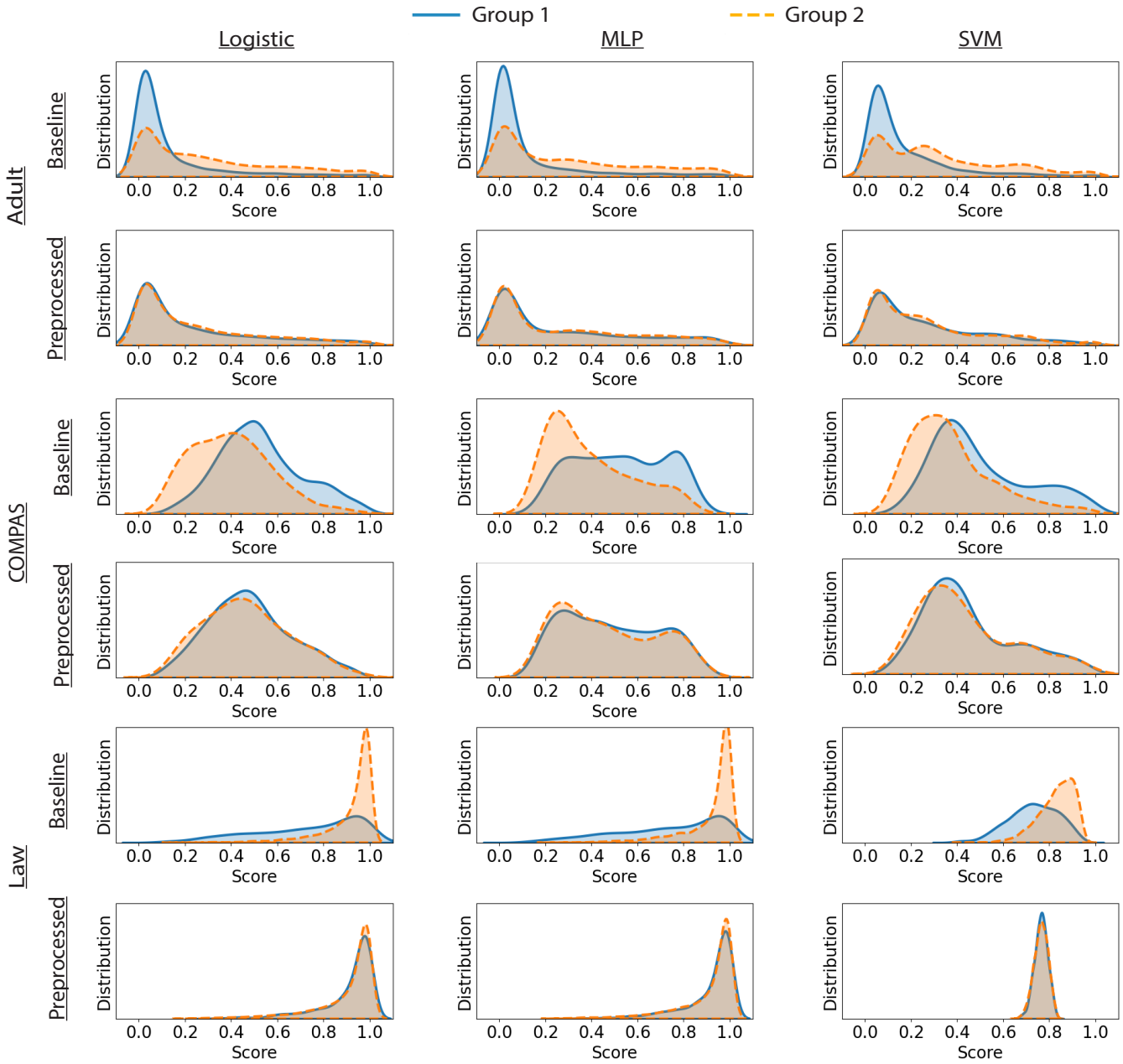


Fig. 5: The distribution of scores produced for each dataset by the baseline scoring function and the pre-processing framework for using three different ML models.

dataset, $T^{(i)}(\mathbf{X}_{ts}^{(i)}), i, ..., |\mathcal{G}|$, and apply the baseline model, $s_{\boldsymbol{\theta}}$, to it. A user-specified threshold, t, can then be used to make the classification decision for these test data.

In our pre-processing framework, we utilize the k-d tree instead of the DT because it possesses one important property: it can be used to *efficiently* perform nearest neighbor searches for data in *high dimensions*. Conversely, performing searches using DTs in higher dimensions can be extremely expensive. We elaborate on the significance of this property in Section V-G and provide runtime analyses for each data structure to illustrate the importance of this property.

B. In-processing Framework

In this section, we describe an in-processing alternative to our pre-processing framework for comparison. In this approach, a scoring function is trained to internalize inter-group fairness in the training process. This approach is similar to that of the algorithm described by Chen et al. [8].

We start by training a baseline scoring function, $s_{\boldsymbol{\theta}}(x)$, as in Phase I of Section IV-A. The weights $\boldsymbol{\theta}$ serve as a starting point from which we will *fine-tune* to improve the fairness of the scoring function. Specifically, once we obtain

$$\boldsymbol{\theta}^{base} = \arg\min_{\boldsymbol{\theta}} L_{\boldsymbol{\theta}}(\mathbf{X}_{tr}), \tag{6}$$

We learn another set of weights, $\boldsymbol{\theta}^{eq}$, for $s_{\boldsymbol{\theta}}(x)$ through

$$\boldsymbol{\theta}^{eq} = \arg\min_{\theta} \; L_{\boldsymbol{\theta}}(\mathbf{X}_{tr}) + \lambda E(\mathbf{X}_{tr}), \tag{7}$$

by initializing $\boldsymbol{\theta} = \boldsymbol{\theta}^{base}$ before applying the learning algorithm. Here, $E(\mathbf{X}_{tr})$, is a regularizer that uses a distance notion to calculate the distance between the scores from two different groups, $i, j \in \mathcal{G}$. We take this two-stage fine-tuning approach instead of directly optimizing Equation 7 starting from random weights for two reasons. First, by using $\boldsymbol{\theta}^{base}$ as our starting point, we are able to start at the most accurate solution and observe the trade-off we incur between inter-group fairness and accuracy during the training process in 7. Second, training the scoring model using Equation 7 with a randomized initialization of $\boldsymbol{\theta}$ can lead to instability in the resulting output of $\boldsymbol{\theta}^{eq}$ reflected by the random seed from which $\boldsymbol{\theta}$ is initialized.

Two common regularizers are considered to compare with our pre-processing framework: (1) an Earth Mover's distance (EMD) regularizer [46] and (2) a Kullback–Leibler (KL) divergence regularizer [8]. The EMD regularizer is of the form:

$$EMD(\mathbf{X}_{tr}^{(i)}, \mathbf{X}_{tr}^{(j)}) = \sum_{n=1}^{C} (CDF_n(\hat{\mathbf{h}}_i) - CDF_n(\hat{\mathbf{h}}_j))^2, \tag{8}$$

where $\hat{\mathbf{h}}_i$ and $\hat{\mathbf{h}}_j$ are Gaussian approximated histogram bins calculated from $\mathbf{X}_{tr}^{(i)}$ and $\mathbf{X}_{tr}^{(j)}$, CDF_n represents the energy in the n^{th} bin of the cumulative distribution function, and n represents the bin number. Because rectangular histogram bins are non-differentiable at the bin edges, we use Gaussian approximations of the rectangular histogram bins as done by Chen et al. [8] so that we can apply a (sub-)gradient-based learning algorithm for optimizing $\boldsymbol{\theta}$.

For the second regularizer tested, we use the symmetric KL divergence with an added Gaussian assumption that Chen et al. [8] formulate. Since the KL divergence is less tractable than the EMD, using the aforementioned Gaussian approximated histogram binning approach leads to convergence issues when trying to equalize group distributions. Adding the Gaussian assumption instead leads to a simple loss that is only dependent on the second-order statistics of each group dataset, alleviating these tractability issues. Thus, letting $\mathcal{N}_i$ and $\mathcal{N}_j$ represent normal distributions with the same mean and variance as group i and group j, respectively. Then, the regularizer is given as:

$$KL(\mathbf{X}_{tr}^{(i)}, \mathbf{X}_{tr}^{(j)}) = \frac{1}{2}\left(\frac{(\hat{\mu}_i - \hat{\mu}_j) - \hat{\sigma}_i}{\hat{\sigma}_j} + \frac{(\hat{\mu}_j - \hat{\mu}_i) - \hat{\sigma}_j}{\hat{\sigma}_i}\right) - 2. \tag{9}$$

V. Experimental Results

A. Datasets for Experimental Studies

In this section, the proposed pre-processing framework is analyzed on the Adult [47], COMPAS risk assessment [41] and Law School [48] datasets—three benchmark datasets with known biases with respect to a given sensitive attribute. For brevity, we simply refer to these datasets as the Adult, COMPAS, and Law datasets. The Adult dataset contains a variety of features correlated with an individual's financial status. The ML task for this dataset is to predict whether individuals in the dataset make above or below $50,000 annually, with bias existing in the features on the basis of sex, with females being less likely than males to make over $50,000. The COMPAS dataset is composed of the records of criminal defendants screened by the COMPAS model in Broward County. It includes information on a defendant's demographics, case details, and criminal histories, with the ML task being to use this information to predict whether a person recidivated within two years of being screened. Biases have been found to exist in the score distributions of white and black demographic groups produced by ML models trained on this dataset, with the white demographic group having lower scores than the black demographic group on average. The Law dataset contains law school admissions records from a survey conducted by the Law School Admission Council (LSAC) across 163 law schools in the United States in 1991. Using the variables provided in this dataset, the ML task is to predict whether someone is likely to pass the bar exam or not. ML models trained on this dataset have been shown to produce biased score distributions when predicting whether white and non-white demographic groups will pass the bar examination, with the white demographic group having much higher scores on average.

TABLE I: Dataset Compositions

Characteristic	Adult	COMPAS	Law
Features	age workclass fnlwgt education education-num marital-status occupation relationship race capital-gain capital-loss hours-per-week native-country	age sex juv_fel_count juv_misd_count juv_other_count priors_count c_charge_degree	decile1b decile3 lsat ugpa zfygpa zgpa fulltime family income sex tier
Class Label	Income >50k (0/1)	Recividate (0/1)	Pass Bar (0/1)
Sensitive attribute	Sex (Male/Female)	Race (White/Black)	Race (White/Non-white)
# Samples (Group 1/Group 2)	30527 / 14695	2103 / 3175	17491 / 3307
# 0/1 Class Labels (Group 1)	20988 / 9539	822 / 1281	1377 / 16114
# 0/1 Class Labels (Group 2)	13026 / 1669	1514 / 1661	916 / 2391

In Table I we provide details on the composition of each dataset used for training and testing the models for our experiments. Full descriptions of the features in each dataset are provided in Table IV of the appendix. Group 1 and Group 2 respectively refer to the advantaged and disadvantaged groups of each dataset. It can clearly be observed that bias exists in the class label distributions associated with each group. We also note that 0 and 1 class labels associated with each dataset may have opposing connotations. For the Adult and Law datasets, a 1 class label respectively indicates that a person makes over $50,000 annually or passed the bar examination (which is good), while for the COMPAS dataset, a 1 class label indicates that a person did recidivate (which is bad). Moreover, in the

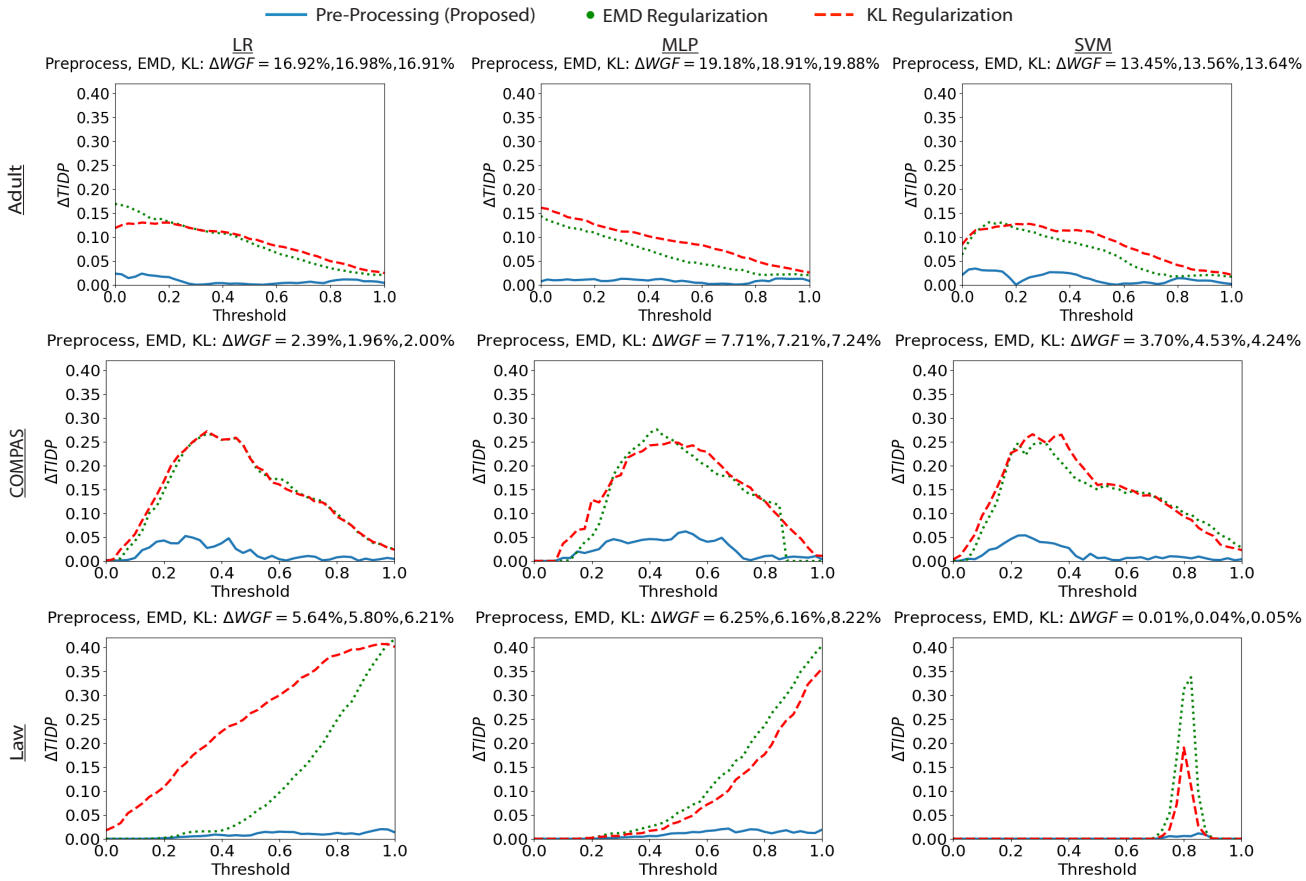


Fig. 6: Plots of the value of Δ_{TIDP} at different thresholds when an average Δ_{WGF} value has been achieved for three different models: (a) LR, (b)MLP, and (c) SVM.

COMPAS dataset, the minority group (Group 1) is advantaged, while the opposite is the case in the Adult and Law dataset. We can further observe that there is a large variation in the sample size of each dataset. Thus, there is diversity in the distributions associated with each dataset, which is good for drawing generalized insights from our results.

B. Experimental Details

The ensuing subsections are devoted to testing our pre-processing framework's ability to achieve both inter- and within-group fairness on all three datasets. We compare its performance with the regularization approaches described in Section IV-B. As done by Chen et al. [8], we split all data into 70% training and 30% testing for the COMPAS and Law datasets and make the simplifying assumption that race is the only binary sensitive attribute that creates inter-group biases in the data. The Adult dataset was originally partitioned into a 75%-25% training-testing split by its creators. For this dataset, we treat sex as the sensitive attribute. All binary and categorical variables were one-hot-encoded and all ML models were trained in TensorFlow using gradient descent with step sizes between 0.1 and 0.0001. All programs were conducted using Python 3.8 and run on a MacBook Pro (1.7 GHz Quad-Core Intel Core i7) with no GPU support. We used the SciPy library for k-d tree construction and set $k = 10$ for the number of nearest neighbors used in constructing the mappings from the group to population domain for all experiments. We also use this library to construct Delaunay triangulations for the experimental analysis provided in Section V-G.

C. Baseline vs. Pre-processing Framework: Risk Distribution Comparison

In this section, the risk score distributions produced by the baseline scoring function and the pre-processing framework are visually compared for three different scoring functions: a logistic regression (LR), three-layer perception (MLP), and support vector machine (SVM). These distributions are presented in Figure 5, with the first and second rows corresponding with the Adult dataset, the third and fourth rows corresponding with the COMPAS dataset, and the fifth and sixth rows corresponding with the Law dataset. The first, third, and fifth rows of distributions capture the baseline scoring function results applied to the raw feature vectors. The second, fourth, and sixth rows of distributions use the pre-processing framework to map the raw feature vectors to the canonical population domains for each dataset before applying their baseline scoring functions to generate their score distributions. The dark, overlapped regions in these plots capture the similarity between the distributions of Groups 1 and 2. A classifier that perfectly captures the threshold invariant demographic parity group fairness definition, TIDP,

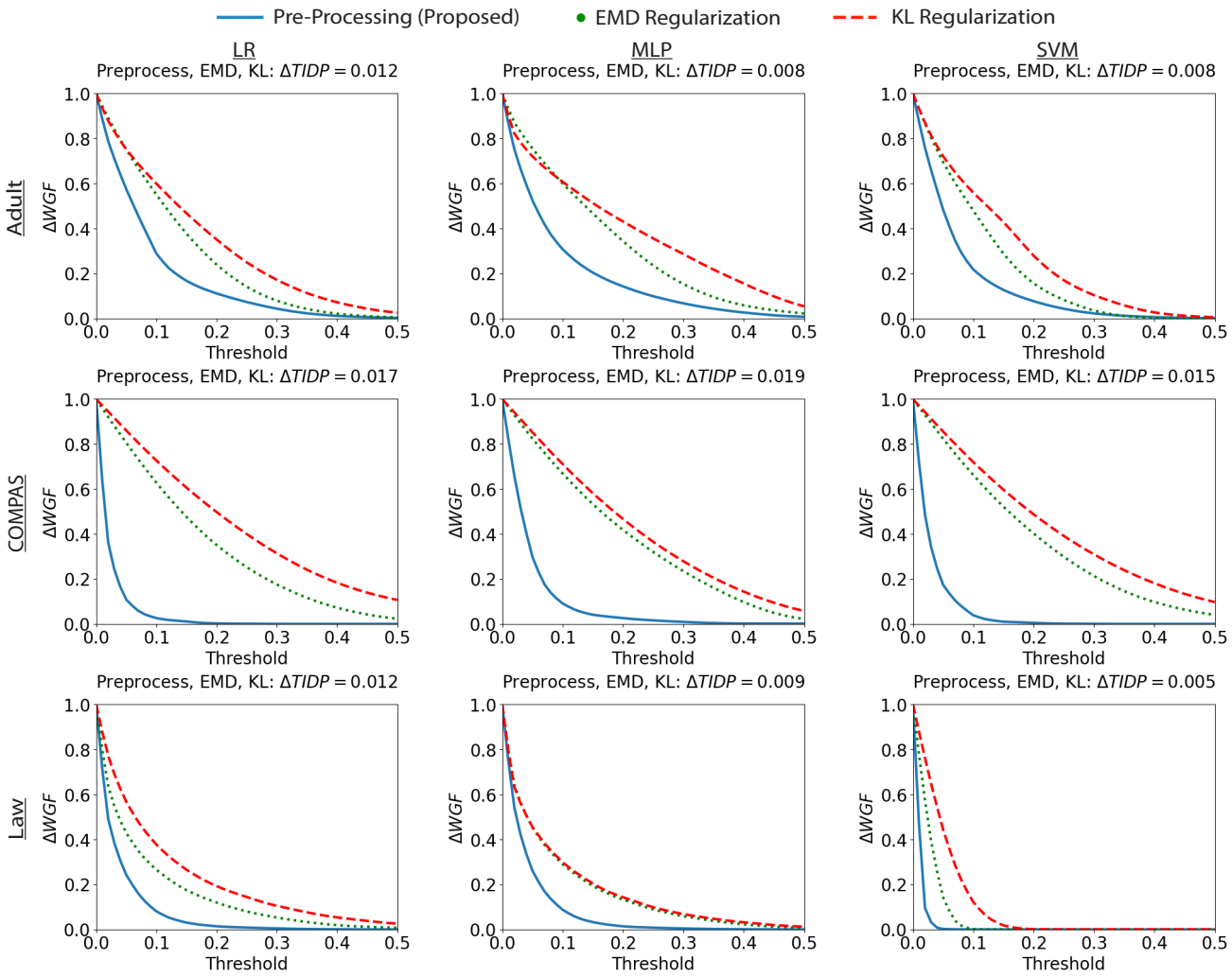


Fig. 7: Plots of Δ_{WGF} at different thresholds when an average Δ_{TIDP} value has been achieved for three different models: (a) LR, (b)MLP, and (c) SVM.

would produce entirely overlapped distributions. It can clearly be seen that distributions associated with the pre-processed features show significantly more overlap than those associated with the raw features for each model for all datasets. Notably, the biases seen in the risk distributions for all sensitive groups is removed when the canonical features are used in place of the raw features for each dataset. Hence, these illustrations capture the improvements made in inter-group fairness by applying the baseline scoring function to the canonical features instead of to raw features.

D. Pre-processing Framework vs. Regularization Methods

In this section, our pre-processing framework's ability to achieve inter- and within-group fairness is compared with the two regularization methods described in Section IV-B respectively using the measures Δ_{TIDP} and Δ_{WGF} from Sections III-B and III-C. Δ_{TIDP} represents the average Calder-Verwer (CV) score obtained for all thresholds $t \in [0, 1]$. The CV score measures the absolute difference between the percentage of individuals in each demographic group that lie on one side of the threshold decision boundary, t. The lower this value, the better a model is at satisfying inter-group fairness for a given value of t. By averaging over all values of $t \in [0, 1]$, Δ_{TIDP} ensures that inter-group fairness is not sensitive to an arbitrary choice of t. Hence smaller values of Δ_{TIDP} indicate that a model is able to satisfy inter-group fairness regardless of our choice in t, meaning that the distributions of scores for different demographic groups are equalized.

As described in Section III-C, Δ_{WGF} measures the percentage of individuals that are treated differently by two different mappings (i.e. the change in the signed distance between pairs of scores within a group is greater than ϵ). This means that a lower value of this metric leads to better preservation of the treatment of individuals across two mappings. In our case, we require any proposed method to treat pairs of individuals within the same demographic group similarly to the baseline model to ensure that within-group fairness is preserved.

Intuitively, obtaining low values of Δ_{TIDP} may require a model to adjust two groups' score distributions to overlap. In situations in which a model applies large morphological changes to each group's score distribution (meaning the ordering of scores or shapes of the group distributions is

significantly changed), this will cause tension between Δ_{TIDP} and Δ_{WGF}. This is because Δ_{WGF} aims to preserve the structure of each group's baseline score distribution. Thus, a model that can successfully achieve low values of Δ_{TIDP} and Δ_{WGF} will minimize the structural changes required to equalize each group's score distributions.

Regularization methods achieve different levels of Δ_{TIDP} and Δ_{WGF} through tuning the hyper-parameter, λ, in Equation 7. In contrast, our pre-processing framework achieves a single level of fairness since it involves no hyper-parameters. Therefore, we aim to compare the ability of these approaches to simultaneously satisfy inter- and within-group fairness by answering the following questions: (1) How do the values of Δ_{TIDP} compare when the regularization methods are required to achieve comparable values of Δ_{WGF} to our pre-processing method? (2) How do the values of Δ_{WGF} compare when the regularization methods are required to achieve comparable values of Δ_{TIDP} to our pre-processing method? If the regularization methods are able to perform as well as our pre-processing framework, the answer to question (1) should be that the differences between groups, measured by Δ_{TIDP}, across all models are comparable, and the answer to question (2) should be that the mapping differences, measured by Δ_{WGF}, across all models are comparable.

We approach the first question as follows. For each pair of models and regularizers, we perform multiple rounds of training over a range of values for λ in Equation 7. Then, from all the training sessions for each pair, we select the model for which the average value (taken over $\epsilon \in [0, 0.5]$) of Δ_{WGF} is closest to equaling the average value of Δ_{WGF} produced by our pre-processing method. Given that the pre-processing framework and regularization methods achieve approximately equal values of Δ_{WGF}, the method that produces a lower value of Δ_{TIDP} is better at achieving overall fairness. The second question is approached in the reverse direction. That is, we again perform multiple rounds of training for each model-regularizer pairing. Then, from all the training sessions for each pair, we select the model for which the average value (taken over $t \in [0, 1]$) of Δ_{TIDP} is closest to the average value of Δ_{TIDP} produced by our pre-processing method. We refer the reader to Section V-F for a detailed analysis of λ.

The results of our comparative analysis are summarized in Figure 6 and Figure 7 for the Adult, COMPAS, and Law testing data. Figure 6 displays plots of the inter-group fairness measure, Δ_{TIDP}, for the pre-processing framework and both regularization methods for three different ML models. The values of Δ_{WGF} above each plot represent the average value of the within-group fairness measure associated with a particular method. Thus, we can see that for approximately equal values of Δ_{WGF}, the pre-processing framework consistently provides significantly lower Δ_{TIDP} values for all models and thresholds for all datasets.

Figure 7 displays plots of the within-group fairness measure, Δ_{WGF}, for the pre-processing framework and both regularization methods for three different ML models. The value of Δ_{TIDP} above each plot represents the average value of the inter-group fairness measure associated with a particular method. Again, regardless of the ML model, our pre-processing framework can achieve lower values Δ_{WGF} for a wide range of thresholds for all datasets, indicating that fewer pairs of individuals in the pre-processing framework violate within-group fairness. Thus, the results in Figures 6 and 7 indicate that the regularization methods must make significantly larger compromises between inter-group and within-group fairness than our pre-processing framework.

E. Performance Evaluation of Pre-processing Framework

In this section, we compare the performances of our pre-processing framework to the regularization methods when all methods achieve comparable levels of inter-group fairness (i.e. Δ_{TIDP} is approximately the same for each approach). The baseline classifier should produce superior performance to all of these methods. However, the performance of a fair model should not significantly deteriorate if it is to be considered useful in practice. Thus, we use two metrics—accuracy and area under the curve (AUC) for the receiver operating characteristic (ROC) curve—to compare the performances of our pre-processing framework and regularization methods to the baseline classifier's performance.

To analyze the accuracy of a method, we use a value of 0.5 as the threshold for all classifiers since this is the most common value used in practice. The resulting performances of these methods on the test data of the Adult, COMPAS, and Law datasets are shown in Table II. To verify that each approach achieves a comparable level of inter-group fairness, we provide the values of Δ_{TIDP} associated with each model in this table. The results show that the pre-processing framework outperforms the regularization-based methods in AUC for all datasets and accuracy for all datasets except the Adult dataset, for which the EMD regularizer produces slightly higher accuracy. Compared with the baseline model, no more than a 2.78% drop in accuracy and 0.022 drop in AUC is incurred by any of the three ML models when applying the pre-processing framework to the COMPAS dataset. In contrast, when applying the regularization methods to the COMPAS dataset, we see accuracy drop-offs of over 8.91% and AUC drop-offs of over 0.100 in all cases.

Although the pre-processing framework also outperforms the regularization approaches on the Law dataset, the difference is less pronounced. That is, all methods incur a little drop in accuracy from the baseline model. This is because there is large class imbalance for this dataset, with 88.86% of the class labels in the test set and 88.97% of the class labels in the entire dataset having a value of 1. Thus, simply assigning everyone the same class label will achieve a high accuracy. This explains why the accuracies of all regularization methods for the Law dataset are identical—each of the regularized models assigns a class label of 1 to every person in the test set. On the other hand, the accuracies displayed for our pre-processing method for the Law dataset show slightly higher variability. The higher accuracies produced by the LR and MLP models result from our pre-processing method's ability to preserve the left tails of the score distributions produced by their baseline models. This is illustrated in the first and second plots in the bottom row of Figure 5 in which the tails of the score distributions produced by the pre-processing method stretch below the 0.5

TABLE II: Performance comparisons on the Adult, COMPAS, and Law Datasets. Results are presented across five trials.

		Baseline			Pre-processing			EMD Regularizer			KL Regularizer		
		Acc.	AUC	$\Delta TIDP$	Acc.	AUC	$\Delta TIDP$	Acc.	AUC	$\Delta TIDP$	Acc.	AUC	$\Delta TIDP$
Adult	LR	$84.55_{\pm 0.20}$	$0.901_{\pm 0.001}$	$0.186_{\pm 0.003}$	$81.84_{\pm 0.12}$	$\mathbf{0.860}_{\pm 0.001}$	$0.008_{\pm 0.001}$	$\mathbf{82.16}_{\pm 0.01}$	$0.847_{\pm 0.001}$	$0.007_{\pm 0.001}$	$81.13_{\pm 0.26}$	$0.830_{\pm 0.005}$	$0.014_{\pm 0.005}$
	MLP	$85.19_{\pm 0.10}$	$0.908_{\pm 0.001}$	$0.186_{\pm 0.001}$	$81.48_{\pm 0.09}$	$\mathbf{0.868}_{\pm 0.001}$	$0.010_{\pm 0.001}$	$\mathbf{81.85}_{\pm 0.17}$	$0.842_{\pm 0.002}$	$0.013_{\pm 0.001}$	$80.29_{\pm 0.39}$	$0.840_{\pm 0.005}$	$0.017_{\pm 0.002}$
	SVM	$84.40_{\pm 0.16}$	$0.898_{\pm 0.001}$	$0.173_{\pm 0.001}$	$81.91_{\pm 0.17}$	$\mathbf{0.859}_{\pm 0.002}$	$0.014_{\pm 0.002}$	$\mathbf{82.33}_{\pm 0.05}$	$0.847_{\pm 0.001}$	$0.009_{\pm 0.002}$	$81.27_{\pm 0.20}$	$0.826_{\pm 0.002}$	$0.014_{\pm 0.005}$
	Avg. Drop				2.97	**0.040**		**2.60**	0.570		3.81	0.070	
COMPAS	LR	$68.41_{\pm 0.00}$	$0.740_{\pm 0.000}$	$0.150_{\pm 0.001}$	$\mathbf{67.09}_{\pm 0.00}$	$\mathbf{0.718}_{\pm 0.000}$	$0.017_{\pm 0.000}$	$56.89_{\pm 0.16}$	$0.588_{\pm 0.002}$	$0.017_{\pm 0.010}$	$55.25_{\pm 1.07}$	$0.554_{\pm 0.001}$	$0.022_{\pm 0.006}$
	MLP	$67.95_{\pm 0.50}$	$0.743_{\pm 0.000}$	$0.152_{\pm 0.005}$	$\mathbf{66.60}_{\pm 0.15}$	$\mathbf{0.722}_{\pm 0.001}$	$0.021_{\pm 0.002}$	$60.31_{\pm 1.01}$	$0.641_{\pm 0.006}$	$0.019_{\pm 0.004}$	$59.68_{\pm 2.78}$	$0.595_{\pm 0.049}$	$0.041_{\pm 0.020}$
	SVM	$67.53_{\pm 0.00}$	$0.737_{\pm 0.000}$	$0.148_{\pm 0.000}$	$\mathbf{64.75}_{\pm 0.00}$	$\mathbf{0.717}_{\pm 0.000}$	$0.015_{\pm 0.000}$	$56.18_{\pm 1.45}$	$0.556_{\pm 0.001}$	$0.015_{\pm 0.003}$	$52.03_{\pm 1.77}$	$0.530_{\pm 0.007}$	$0.017_{\pm 0.009}$
	Avg. Drop				**1.33**	**0.021**		11.52	0.145		13.16	0.180	
Law	LR	$89.74_{\pm 0.05}$	$0.869_{\pm 0.000}$	$0.197_{\pm 0.001}$	$\mathbf{89.14}_{\pm 0.03}$	$\mathbf{0.818}_{\pm 0.000}$	$0.007_{\pm 0.001}$	$88.86_{\pm 0.00}$	$0.712_{\pm 0.001}$	$0.007_{\pm 0.001}$	$88.86_{\pm 0.00}$	$0.630_{\pm 0.018}$	$0.005_{\pm 0.002}$
	MLP	$89.75_{\pm 0.04}$	$0.872_{\pm 0.001}$	$0.188_{\pm 0.002}$	$\mathbf{89.19}_{\pm 0.02}$	$\mathbf{0.819}_{\pm 0.001}$	$0.007_{\pm 0.004}$	$88.86_{\pm 0.00}$	$0.714_{\pm 0.013}$	$0.006_{\pm 0.001}$	$88.86_{\pm 0.00}$	$0.701_{\pm 0.013}$	$0.002_{\pm 0.002}$
	SVM	$88.86_{\pm 0.00}$	$0.862_{\pm 0.002}$	$0.047_{\pm 0.012}$	$88.86_{\pm 0.00}$	$\mathbf{0.819}_{\pm 0.001}$	$0.002_{\pm 0.001}$	$88.86_{\pm 0.00}$	$0.745_{\pm 0.033}$	$0.001_{\pm 0.000}$	$88.86_{\pm 0.00}$	$0.651_{\pm 0.021}$	$0.001_{\pm 0.000}$
	Avg. Drop				**0.60**	**0.049**		0.88	0.144		0.88	0.204	

threshold. As a result, some of the test samples are assigned 0 class labels. Conversely, the AUC for the ROC curve shows much more significant drop-offs in the results obtained for the regularization methods than the pre-processing method. This is because the ordering of scores produced by the regularization methods has become shuffled with respect to the score ordering of the baseline model. This causes the results to be sensitive to the placement of the decision threshold over the support of the score distribution.

As previously mentioned, our pre-processing approach achieves slightly lower accuracy on the Adult dataset compared to the EMD regularizer when using a 0.5 classification threshold. This is not unexpected since accuracy is determined by applying a single threshold to the score distributions to perform classification. Though 0.5 is a common choice of threshold, it may not be universally optimal for every dataset. This observation is reflected in the fact that our pre-processing framework's AUC performance is higher than the EMD regularizer's for every dataset, including the Adult dataset. Since AUC captures the dynamics associated with selecting any threshold in the range $[0, 1]$, it is threshold invariant.

F. λ Sensitivity Analysis

In Section V-D, we analyzed inter- and within-group fairness for the EMD and KL regularization methods by tuning λ in Equation 7. In this section, we provide experimental results that illustrate the effect that λ has on the values of Δ_{TIDP} and Δ_{WGF} for these methods. Figure 8 provides plots that illustrate the progression of the values of Δ_{TIDP} and Δ_{WGF} as λ is tuned over a range of values to train the LR, MLP, and SVM models. For all datasets, each of these models was trained over five sessions for each value of λ. Each plot contains four curves (with standard deviation error bars from the five trials) which provide the test results for the values of Δ_{TIDP} and Δ_{WGF} for the EMD and KL regularizers. The left and right y-axes respectively provide the scales for the Δ_{TIDP} (blue and red) and Δ_{WGF} (black and green) curves. We personalize the tuning of λ for each model and regularizer pairing since the strength of λ may have different effects on each pairing. Thus, the top x-axis provides the scale of λ for all EMD regularizer results (the blue and black curves), while the bottom x-axis provides the scale of λ for all KL regularizer results (the red and green curves).

As the strength of λ increases, we can see that the red and blue curves decrease in value, meaning that the values of Δ_{TIDP} associated with the two regularizers drop, and group fairness is improved. However, in every case, drops in Δ_{TIDP} lead to increases in Δ_{WGF}, indicated by the green and black curves, which means that there is an increase in the violation of within-group fairness. The size of this effect varies by dataset and model. The COMPAS dataset, for example, tends to lead to the largest violation of Δ_{WGF} and also produces the least stable curves, indicated by the large standard deviations associated with its curves. The features in this dataset likely lack enough heterogeneity to allow a model to learn to distinguish between its two sensitive groups, causing it to make extreme adjustments to its baseline score distribution to satisfy inter-group fairness. These results highlight the contribution of our pre-processing framework, which does not suffer this inter- and within-group fairness trade-off. The Adult and Law datasets also suffer from this issue, though the extent of these violations is less pronounced. This is in part because these contain a larger and more fruitful set of features.

G. Time Complexity Comparison: k-d Tree vs Delaunay Triangulation

An important and non-trivial consideration that must be taken into account for representing the manifold used in our pre-processing framework is the time required to construct the data structure used to represent it. While we have found that using the Delaunay Triangulation (DT) in place of the k-d tree in our pre-processing framework produces comparable performance results for the COMPAS dataset, the time complexity for computing a DT from N feature vectors in dimension k is $\mathcal{O}(N^{\lceil k \rceil})$, which quickly becomes intractable as the feature vector dimension grows to double digits. In comparison, the time complexity of constructing a k-d tree is $\mathcal{O}(kNlog(N))$, which is far less expensive and linear in the feature vector dimension, k. To illustrate this computational difference, we present the time complexities in seconds associated with computing these data structures for Groups 1 and 2 for the Adult, COMPAS, and Law datasets in Table III. A clear and stark contrast between the costs associated with computing the DTs and k-d trees can clearly be observed in this table. In particular, the k-d tree is always able to be computed in under 20 seconds for each group for all datasets. Conversely, no DT is able to be constructed in a relatively similar amount of time. Notably, for the Adult and Law datasets, each group's associated DT is unable to be computed in under 12 hours. This is because both of these datasets contain far more samples than

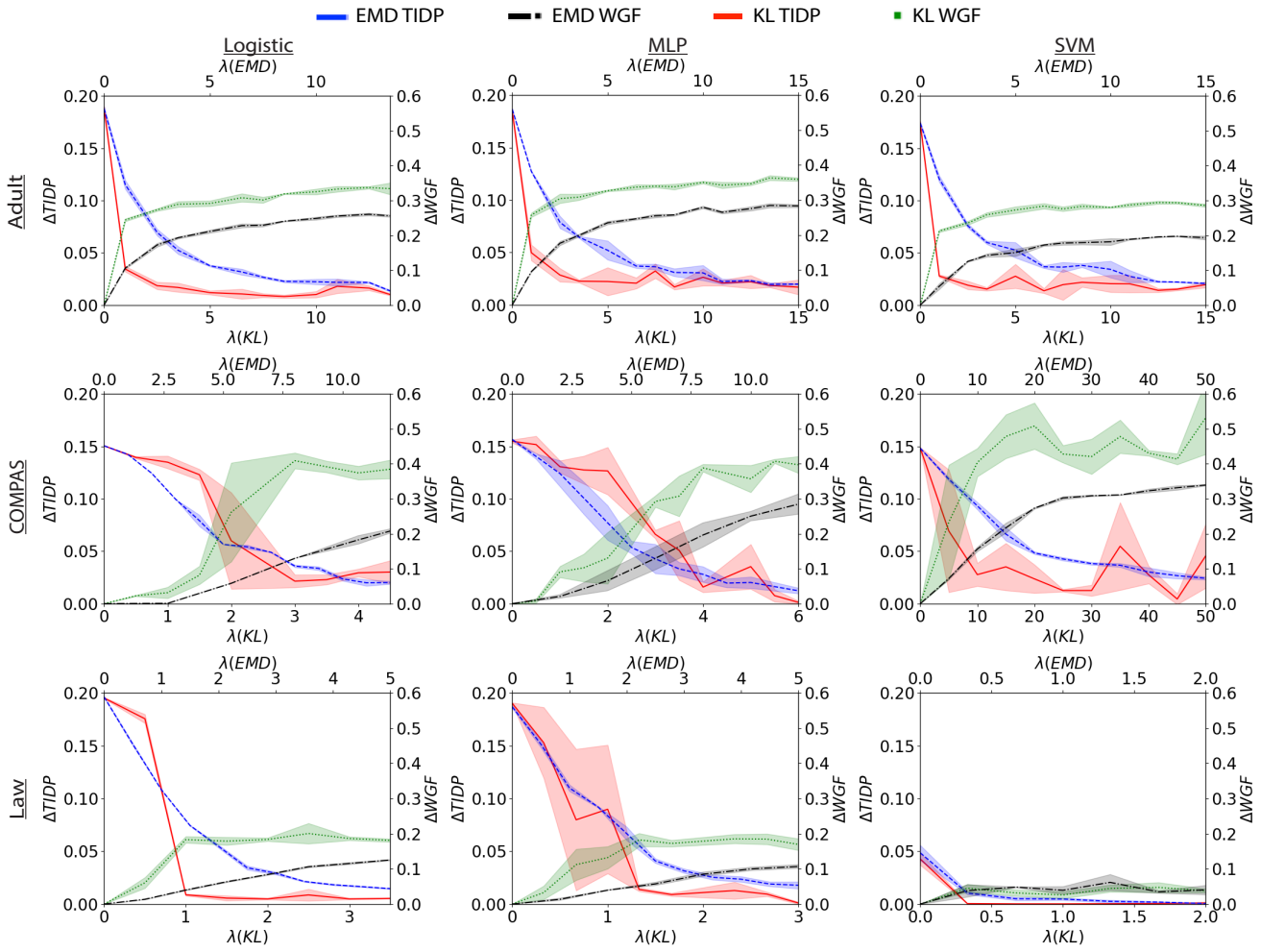


Fig. 8: Plots of Δ_{TIDP} and Δ_{WGF} for the EMD and KL regularizers for different values of λ. Results provided for training LR, MLP, and SVM models over 5 sessions for the Adult, Compas, and Law datasets.

TABLE III: Time complexity comparison for constructing data structures. Results are presented in seconds across five trials. Processes not terminated within 12 hours are reported as "—."

	Adult		COMPAS		Law	
	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2
DT	—	—	$172.59_{\pm 1.73}$	$905.46_{\pm 21.88}$	—	—
k-d Tree	$16.12_{\pm 0.56}$	$4.78_{\pm 0.04}$	$0.05_{\pm 0.01}$	$0.10_{\pm 0.01}$	$1.07_{\pm 0.03}$	$0.16_{\pm 0.01}$

the COMPAS dataset. This emphasizes the scalability issues associated with using the DT for manifold representation in higher dimensions. It is also notable that it costs more time to construct the k-d tree for the Adult dataset than for the Law dataset. The reasons for this are: (1) the Adult dataset contains more samples and (2) the Adult dataset contains more categorical features than the Law dataset, meaning that when they are one-hot encoded, the dimensions of the associated feature vectors increase. Nevertheless, the time complexities associated with constructing the k-d tree for the Adult dataset are still quite efficient.

VI. Limitations and Discussions

In this section, we list and discuss the limitations associated with our proposed framework. We first note that for the proposed algorithm to succeed in achieving both inter- and within-group fairness, the following assumption must approximately hold: if the distance between the scores of two feature vectors is close, then the distance between these feature vectors in the input space should be close. This assumption holds well on for the Adult, COMPAS, and Law datasets, but may not hold in general. If this situation fails, feature vectors that are close in the group domain may map to feature vectors that are far from each other in the population domain. Thus, and convex sum of these population feature vectors may lie outside the population manifold. To resolve this issue in our ongoing research, we are investigating new approaches for constructing feature correspondences between the group and population manifolds.

Second, we note that the mapping constructed in the configuration of the pre-processing framework proposed in this paper requires storing the entire training dataset in the k-d tree data structure. While this approach allowed us to construct an interpretable mapping between a group and the population manifold, more compact representations may potentially be constructed by incorporating the mapping construction into the learning process. We plan to further investigate such solutions in our future work.

We also would like to observe that the focus of this work has been analyzing the effectiveness of our framework on tabular data. Nevertheless, our framework has the potential to be generalized to other forms of data, such as image data. While such data in its raw form may exist in much higher dimensions, which may produce a computational burden when constructing the k-d tree, our framework may be applied to the features in the latent space of intermediate layers in an ML model or to feature vectors to which dimensionality reduction has first be applied.

Finally, we would like to mention two comments with regard to the results of our study. First, while we have observed that $TIDP$ and WGF may conflict for the Adult, COMPAS, and Law datasets, this conflict is distributional dependent and should not be present in datasets for which the baseline classifier does to produce a major violation of demographic parity. For reference, Le Quy et al. [49] provide a survey of the primary benchmark datasets used for analyzing fairness in ML and list common fairness definitions that conflict with each. Second, we would like to acknowledge that, while we have compared our fairness framework against two regularization methods using three ML models, this list does not include an exhaustive analysis of all potential regularizers and ML models, which may influence our results.

VII. Conclusion

In this paper, we study the importance of maintaining within-group fairness in situations where satisfying inter-group fairness is required. We introduce a notion of within-group fairness and a metric for measuring it. We have adopted the stricter threshold invariant form of demographic parity for analyzing inter-group fairness, which requires the scores generated for different groups to be equally distributed. Furthermore, we introduce a novel pre-processing framework that maps raw features to a canonical domain before applying a classifier and performs well in achieving inter-group and within-group fairness. This framework requires sensitive group attributes to be used only for pre-processing the raw features in the testing stage of the machine learning process and never explicitly provides the sensitive attributes to a classifier to make decisions. To verify its effectiveness, we compare the performance of our framework to models in which fairness is embedded in the training process through regularization. Experimental results demonstrate that our pre-processing framework can achieve both inter-group and within-group fairness with little penalty on accuracy.

VIII. Acknowledgement

We thank the reviewers for their insightful comments and feedback. This work was supported in part by the NSF grant #IIS-2147276. We gratefully acknowledge support from the JP Morgan Chase AI Faculty Research Award.

References

[1] E. Mishraky, A. B. Arie, Y. Horesh, and S. M. Lador, "Bias detection by using name disparity tables across protected groups," *Journal of Responsible Technology*, vol. 9, p. 100020, 2022.

[2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," in *Ethics of Data and Analytics*, pp. 254–264, Auerbach Publications, 2016.

[3] L. Sweeney, "Discrimination in online ad delivery," *Communications of the ACM*, vol. 56, no. 5, pp. 44–54, 2013.

[4] J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," in *Ethics of Data and Analytics*, pp. 296–299, Auerbach Publications, 2018.

[5] A. S. Adamson and A. Smith, "Machine learning and health care disparities in dermatology," *JAMA Dermatology*, vol. 154, no. 11, pp. 1247–1248, 2018.

[6] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[7] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.

[8] M. Chen and M. Wu, "Towards threshold invariant fair classification," in *Conference on Uncertainty in Artificial Intelligence*, pp. 560–569, PMLR, 2020.

[9] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268, 2015.

[10] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012.

[11] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010.

[12] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50, Springer, 2012.

[13] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012.

[14] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[15] M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth, "Fairness in learning: Classic and contextual bandits," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[16] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making," *Communications of the ACM*, vol. 64, no. 4, pp. 136–143, 2021.

[17] C. Jung, M. J. Kearns, S. Neel, A. Roth, L. Stapleton, and Z. S. Wu, "Eliciting and enforcing subjective individual fairness," *CoRR*, 2019.

[18] C. Ilvento, "Metric learning for individual fairness," in *1st Symposium on Foundations of Responsible Computing*, 2020.

[19] D. Mukherjee, M. Yurochkin, M. Banerjee, and Y. Sun, "Two simple ways to learn individual fairness metrics from data," in *International Conference on Machine Learning*, pp. 7097–7107, PMLR, 2020.

[20] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," 09 2016.

[21] H. Zhao and G. J. Gordon, "Inherent tradeoffs in learning fair representations," *The Journal of Machine Learning Research*, vol. 23, no. 1, pp. 2527–2552, 2022.

[22] "FDIC Consumer Compliance Examination Manual," 2021.

[23] J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell, "Fairness under unawareness: Assessing disparity when protected class is unobserved," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 339–348, 2019.

[24] C. Kilpatrick and H. Eklund, "Article 21–non-discrimination," in *The EU Charter of Fundamental Rights*, pp. 613–638, Nomos Verlagsgesellschaft mbH & Co. KG, 2022.

[25] B. Hsu, R. Mazumder, P. Nandy, and K. Basu, "Pushing the limits of fairness impossibility: Who's the fairest of them all?," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 32749–32761, Curran Associates, Inc., 2022.

[26] M. Chen, A. Shahverdi, S. Anderson, S. Y. Park, J. Zhang, D. Dachman-Soled, K. Lauter, and M. Wu, "Transparency tools for fairness in ai (luskin)," *Research in Mathematics and Public Policy*, pp. 47–80, 2020.

[27] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial intelligence and statistics*, pp. 962–970, PMLR, 2017.

[28] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017.

[29] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa, "Wasserstein fair classification," in *Uncertainty in artificial intelligence*, pp. 862–872, PMLR, 2020.

[30] Y. Tian and Y. Zhang, "A comprehensive survey on regularization strategies in machine learning," *Information Fusion*, vol. 80, pp. 146–166, 2022.

[31] H. Xiong, R. Wan, J. Zhao, Z. Chen, X. Li, Z. Zhu, and J. Huan, "Grod: Deep learning with gradients orthogonal decomposition for knowledge transfer, distillation, and adversarial training," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 16, no. 6, pp. 1–25, 2022.

[32] J. Zhao, J. Li, F. Zhao, S. Yan, and J. Feng, "Marginalized cnn: Learning deep invariant representations," 2017.

[33] 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*, pp. 2564–2572, PMLR, 2018.

[34] 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.

[35] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, pp. 325–333, PMLR, 2013.

[36] N. Mehrabi, U. Gupta, F. Morstatter, G. Ver Steeg, and A. Galstyan, "Attributing fair decisions with attention interventions," in *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pp. 12–25, 2022.

[37] D. Pessach and E. Shmueli, "Algorithmic fairness," in *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*, pp. 867–886, Springer, 2023.

[38] C. Silvia, J. Ray, S. Tom, P. Aldo, J. Heinrich, and A. John, "A general approach to fairness with optimal transport," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 3633–3640, 2020.

[39] V. V. Ramaswamy, S. S. Kim, and O. Russakovsky, "Fair attribute classification through latent space de-biasing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9301–9310, 2021.

[40] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil, "Fair regression with wasserstein barycenters," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7321–7331, 2020.

[41] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, "How we analyzed the COMPAS recidivism algorithm," *ProPublica (5 2016)*, vol. 9, no. 1, pp. 3–3, 2016.

[42] M. Langenkamp, A. Costa, and C. Cheung, "Hiring fairly in the age of algorithms." *Available at https://ssrn.com/abstract=3723046 or http://dx.doi.org/10.2139/ssrn.3723046* , 2019.

[43] C. Intahchomphoo and O. E. Gundersen, "Artificial intelligence and race: A systematic review," *Legal Information Management*, vol. 20, no. 2, pp. 74–84, 2020.

[44] R. C. González and R. E. Woods, *Digital Image Processing, 3rd Edition*. Pearson Education, 2008.

[45] D. An, Y. Guo, N. Lei, Z. Luo, S.-T. Yau, and X. Gu, "Ae-Ot: A new generative model based on extended semi-discrete optimal transport," *International Conference on Learning Representations*, 2020.

[46] L. Hou, C.-P. Yu, and D. Samaras, "Squared earth movers distance loss for training deep neural networks on ordered-classes," in *NIPS workshop*, vol. 5, 2017.

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

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

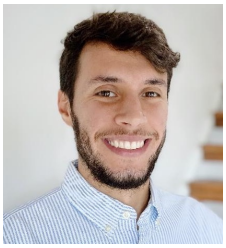
[49] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi, "A survey on datasets for fairness-aware machine learning," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 12, no. 3, p. e1452, 2022.

APPENDIX

In this appendix section, we provide details on the meaning of each of the features used for classification in the Adult [47], COMPAS [41], and Law [48] datasets. These descriptions are respectively provided in Table IV.

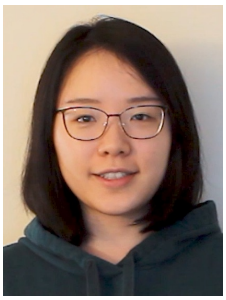
TABLE IV: Feature descriptions for the Adult, COMPAS, and Law datasets

Dataset	Feature	Description
Adult	age	Individual's age
	workclass	Employment status
	fnlwgt	Weight variable
	education	Highest education level attained
	education-num	Numerical form of education variable
	marital-status	Individual's age
	occupation	Individual's general occupation category
	relationship	Individual's relationship to others
	race	Individual's race
	capital-gain	Amount of capital gains
	capital-loss	Amount of capital losses
	hours-per-week	Number of hours worked per week
	native-country	Individual's country of origin
COMPAS	age	Defendant age
	sex	Defendant sex
	juv_fel_count	# juvenile felonies committed
	juv_misd_count	# juvenile misdemeanors committed
	juv_other_count	# of other crimes committed
	priors_count	# prior criminal records
	c_charge_degree	Charge degree (felony/ misdemeanor)
Law	decile1b	Decile in the school given student's grades in Year 1
	decile3b	Decile in the school given student's grades in Year 3
	lsat	LSAT score
	ugpa	Undergraduate GPA
	zfygpa	First year law school GPA
	zgpa	Cumulative law school GPA
	fulltime	Whether student will work full- or part-time
	family income	Student's family income bracket
	sex	Student's sex
	tier	Student's tier



Zachary McBride Lazri (Student Member, IEEE) is a Ph.D candidate in the Department of Electrical and Computer Engineering at the University of Maryland, College Park, MD, USA. He received his B.S. degree in mathematics with a double major in economics from the University of Maryland, College Park, MD in Spring 2017. He is the recipient of the 2024-2025 Ann G. Wylie Dissertation Fellowship and won the Three-Minute Thesis (3MT) competition in 2023. In Summer 2021 and Summer 2022, he was with Dolby Laboratories as a Video Processing Research Intern. He was also with the Center for Devices and Radiological Health of the U.S. Food and Drug Administration (FDA) as an ORISE research fellow from Fall 2019 to Summer 2020.

Ivan Brugere is an AI Research Scientist at J.P. Morgan-Chase since 2021 as part of the Explainable AI Center of Excellence. His work focuses on AI fairness, model robustness, predictive multiplicity, and multi-objective optimization. Prior to this, Ivan was a Research Scientist at Salesforce Research. Ivan completed his PhD from the University of Illinois at Chicago in Computer Science in 2020, working on graph representation learning, focused in biological applications.



Xin Tian (Member, IEEE) received the B.E. degree in optoelectronic information science and engineering from the Huazhong University of Science and Technology of China, Wuhan, China, in 2017, and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA, in 2022. She has been a Research Scientist with Meta Inc., Menlo Park, CA, USA, since 2022. Her research interests are signal processing, data science, and machine learning in smart health. Dr. Tian was awarded as an Outstanding Graduate Assistant by the University of Maryland in 2020 and 2021, respectively, and was selected as a Future Faculty Fellow.

Dana Dachman-Soled is an associate professor in the Department of Electrical and Computer Engineering and UMIACS at the University of Maryland, College Park and is a core faculty member of the Maryland Cybersecurity Center. She is the recipient of an NSF CAREER award, a Ralph E. Powe Junior Faculty Enhancement award, and a JP Morgan AI Faculty Research Award. In addition, her research has been funded by NSF, NIST, Cisco, Amazon, and Intel. Prior to joining the University of Maryland, Dana spent two years as a postdoc at Microsoft Research New England. She received her Ph.D in computer science at Columbia University in 2011.

Her research interests are in cryptography, complexity theory, and machine learning. She is currently working in the broad areas of non-malleable codes, post-quantum cryptography, secure multiparty computation, and fairness and privacy in machine learning.

Antigoni Polychroniadou, Executive Director at J.P. Morgan AI Research, heads the J.P. Morgan AlgoCRYPT Center of Excellence. Holding a Ph.D. from Arhus University under the supervision of Ivan Damgaard and post-doctoral training from Cornell University, she has received several awards, including the junior Simons fellowship, and is named Prolific JPMorgan inventor. Her team's cryptographic tools, leveraging new privacy-preserving algorithms, are reshaping the landscape of data utilization in finance.

Danial Dervovic received the Ph.D. degree in Computer Science from University College London, United Kingdom, in 2020. Since 2019, he has worked at J.P. Morgan as an AI Research Scientist, authoring multiple papers and patents. His research interests include AI applied to finance problems, explainable AI, sequential decision making and quantum computing.

Min Wu (Fellow, IEEE) received the B.E. degree in automation (Highest Hons.) and the B.A. degree in economics from Tsinghua University, Beijing, China, in 1996, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 2001.

She is currently a Christine Y. Kim Eminent Professor in Information Technology and an Associate Dean of Engineering, and a Distinguished University Professor with the University of Maryland, College Park, MD, USA. Her research interests include information security and forensics, multimedia signal processing, and applications of data science and machine learning in health and IoT.

Prof. Wu received a U.S. NSF CAREER Award, a U.S. ONR Young Investigator Award, a TR100 Young Innovator Award from the MIT Technology Review, an IEEE Harriett B. Rigas Education Award, a J.P. Morgan AI Faculty Research Award, and an IEEE Signal Processing Society Meritorious Service Award. She has served as the Chair of IEEE Technical Committee on Information Forensics and Security, the Vice President-Finance of IEEE Signal Processing Society, and the Editor-in-Chief of IEEE SIGNAL PROCESSING MAGAZINE. She has been elected to serve as the 2024-2025 President of the IEEE Signal Processing Society. She is a Fellow of AAAS, the U.S. National Academy of Inventors, and the Washington Academy of Sciences.