2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE)

Fairness Testing through Extreme Value Theory

Verya Monjezi
vmonj@uic.edu
University of Illinois Chicago

Ashutosh Trivedi
ashutosh.trivedi@colorado.edu
University of Colorado Boulder

Vladik Kreinovich
vladik@utep.edu
University of Texas at El Paso

Saeid Tizpaz-Niari
saeid@uic.edu
University of Illinois Chicago

***Abstract*—Data-driven software is increasingly being used as a critical component of automated decision-support systems. Since this class of software learns its logic from historical data, it can encode or amplify discriminatory practices. Previous research on algorithmic fairness has focused on improving "average-case" fairness. On the other hand, fairness at the extreme ends of the spectrum, which often signifies lasting and impactful shifts in societal attitudes, has received significantly less emphasis.**

Leveraging the statistics of extreme value theory (EVT), we propose a novel fairness criterion called *extreme counterfactual discrimination* (ECD). This criterion estimates the worst-case amounts of disadvantage in outcomes for individuals solely based on their memberships in a protected group. Utilizing tools from search-based software engineering and generative AI, we present a randomized algorithm that samples a statistically significant set of points from the tail of ML outcome distributions even if the input dataset lacks a sufficient number of relevant samples.

We conducted several experiments on four ML models (deep neural networks, logistic regression, and random forests) over 10 socially relevant tasks from the literature on algorithmic fairness. First, we evaluate the generative AI methods and find that they generate sufficient samples to infer valid EVT distribution in 95% of cases. Remarkably, we found that the prevalent bias mitigators reduce the average-case discrimination but increase the worst-case discrimination significantly in 35% of cases. We also observed that even the tail-aware mitigation algorithm—MiniMax-Fairness—increased the worst-case discrimination in 30% of cases. We propose a novel ECD-based mitigator that improves fairness in the tail in 90% of cases with no degradation of the average-case discrimination. We hope that the EVT framework serves as a robust tool for evaluating fairness in both average-case and worst-case discrimination.

I. Introduction

Recent technological advancements in training large machine learning (ML) models, such as deep neural networks [1], deep reinforcement learning [2], and large language models [3], [4], have led to a proliferation of data-driven software in almost every aspect of modern socioeconomic infrastructure. These data-driven systems, such as those that decide on recidivism [5], predict benefit eligibility [6], [7], or decide whether to audit a given taxpayer [8], [9], learn their decision logic as ML models by mining simple patterns from historical data. However, these systems often codify and amplify the biases present in the historical data due to various systemic factors. To address this challenge, the software engineering community has developed solutions to characterize, quantify, and mitigate bias in the ML models. We discuss their inadequacies and propose new tools and techniques for the tail of outcome distributions of data-driven software.

Inadequacies of Average-Case Fairness. Although there is an increased participation of minorities (e.g., women) in the labor market (parity in average), they are considerably underrepresented in high-paying occupations and leadership positions [10] (disparity in the extreme). Additionally, the wage gap between privileged and unprivileged individuals continues to be more pronounced in high-paying jobs [11], [10]. Considering these factors, it is indeed surprising that a notable gap exists in the literature regarding the evaluation of algorithmic fairness in the context of extreme outcomes.

One broad class of fairness definitions is individual fairness [12] which requires treating individuals similarly if they are deemed similar based on their non-protected attributes, regardless of their protected attributes. One popular individual fairness notion is *counterfactual discrimination* which necessitates that algorithmic outcomes should be similar for an individual and any related counterfactual individual who differs only in protected attributes. However, these fairness notions primarily focus on the average behavior (expected value or variance) of the model, which can create a false sense of fairness by ignoring the discrimination in socially influential edge cases. This paper presents a framework rooted in EVT to quantify AI fairness within the tail of ML outcomes.

Statistics of the Extreme: Extreme Value Theory. While statistics and machine learning typically focus on "usual" behavior, extreme value theory (EVT) [13] is a branch of statistics that deals with unusual or extreme behaviors. EVT can be applied to model rare events such as the maximum temperature in the summer. Under appropriate assumptions, the statistics of extreme values follow the generalized extreme value (GEV) distribution, which is analogous to the central limit theorem for the statistics of averages or expected values.

Fairness through Extreme Value Theory. The primary focus of this paper centers on a narrow view of *equality of opportunity*, which necessitates similar individuals to be treated similarly at the time of decision-making, as defined by Dwork et al. [12]. We consider the distribution of "counterfactual discrimination," which refers to the distribution of

1558-1225/25/$31.00 ©2025 IEEE
DOI 10.1109/ICSE55347.2025.00070

differences in the ML outcomes when a protected attribute, like race or gender, is altered from observed value A to a counterfactual B. While previous studies have focused on the expected values from this distribution known as average causal discrimination (ACD) [14], representing the "average" change in the outcome when a protected attribute is flipped, this work quantifies the "maximum" change in the ML outcome when a protected attribute is flipped. We call this quantity *extreme counterfactual discrimination* (ECD) and use GEV distributions to model and quantify it. By comparing the GEV distributions of different (sub-)groups, we quantify the fairness of ML models in the extreme tail of outcome distributions as well as the effectiveness of mitigation algorithms in reducing discrimination in the tail.

The ECD metric, proposed in this paper, has a normative implication. It tells us "in the worst-case, how much (dis)advantages are experienced by individuals solely due to their memberships in (un)privileged groups at the time of ML decision-making." Our proposal complements THEMIS [14] that shows the amounts of such discrimination on average. For example, an ACD of +0.05 vs. an ECD of +0.25 for an unprivileged group show that flipping their protected attributes to a privileged group increased their likelihood of receiving favorable ML outcomes by 5% on average, but up to 25% in the worst-case. The statistics of EVT allow us to directly model GEV distributions that bring significant advantages. It directly models the tail distribution, which allows us to investigate the validity of the tail or provide statistical guarantees on the returns/likelihood of extreme discrimination. Other metrics, like the conditional value at risk (CVaR) [15], model the tail of a usual distribution (e.g., normal distributions). Hence, they fail to reason about the validity of the tail and provide any statistical guarantees on extrapolations.

Statistical ECD-Testing Framework. In this paper, we have developed a randomized test-case generation algorithm that explores the tail of ML models and applies the exponentiality test [16], [17] to convince statistical significance. The primary challenge stems from the statistical test's requirements for a certain size of tail samples and the scarcity of samples in the extreme tail (which is precisely why they are considered extreme). If only a subset of these tail samples is included in the analysis, it can result in low confidence in the model due to high variance. On the other hand, selecting a larger number of data points will lead to the erroneous inclusion of non-tail samples and the inference of mixture distributions that violate the asymptotic basis of extreme value theory. Rather than randomly generating the test-cases from the domain of variables [14], we leverage and evaluate various generative methods, such as GANs [18] and VAEs [19], to synthesize samples with realistic combinations of features.

Experiments. We conducted experiments on nine fairness-sensitive datasets with four popular classifiers (an overall 40 training scenarios). Our findings indicate that EVT fits well to the tail of counterfactual bias distributions in 95% of cases that enable us to derive worst-case guarantees. In 25% of scenarios, the worst-case and average-case CD differ significantly across different groups. We also evaluated the characteristics of fairness in the tail over four mitigation algorithms: exponentiated gradient (EG) [20], Fair-SMOTE [21], MAAT [22], and STEALTH [23]. Our results over the mitigated model show that the worst-case and average-case CD differ significantly across different groups in 52% of cases. In addition, the average-based mitigated models significantly increase worst-case discrimination in 35% of the cases, while preserving or improving average fairness in 63% of cases. With tail-based methods, we implement an in-process mitigation strategy that outperforms MiniMax-Fairness [24] and reduces the discrimination in the tail for 90% of cases while improving average fairness in 45% of cases.

Contributions. In this paper, we

- introduce a metric to measure unfairness in the tail of ML outcome distributions,
- present a fairness testing method that generates realistic test-cases and provides statistical guarantees in the tail,
- evaluate the worst-case discrimination for a large set of well-established algorithms and bias mitigators, and
- propose and evaluate a novel tail-aware mitigator.

II. OVERVIEW

We first give a background overview for extreme value theory. We then go through our approach step by step using an example of adult census income dataset, trained using a DNN algorithm.

Extreme Value Theory. Given a set of independent and identically distributed random variables $\{z_1, \dots, z_n\}$, the extreme value theory is concerned with the max statistics of a random process, i.e., $M_n = \max(\{z_1, \dots, z_n\})$. Under some mild assumptions, it has been proved (e.g., see Leadbetter et al. [25]) that M_n belongs to a family of distributions called the *generalized extreme value (GEV)*. There are two basic approaches to infer the parameters of GEV distributions: block maximum and threshold approach [13]. In this paper, we use the threshold approach where extreme events that exceed some high threshold u, i.e., $\{z_i : z_i > u\}$, are extreme values. The GEV distribution has three parameters: a location parameter, a scale parameter, and a shape parameter. When the shape is close to zero or negative, the statistical guarantees on the worst-case discrimination may be feasible.

Threshold Selection. A proper choice of threshold value u is critical to analyze the behavior of GEV. Low values of threshold u might include non-tail samples and lead to mixture distributions that violate the asymptotic basis of the model. On the other hand, high values of threshold u might include only a few tail samples and lead to low confidence in the model due to high variance. In this work, we use coefficient of variation (CV) and provide statistical guarantees in picking thresholds.

Return Level. A *return level* describes by the set of points (m, δ_m) where m is the time period (e.g., the number of queries to the ML software) and the level δ_m is expected to observed during the m period (e.g., maximum discrimination after m interactions).

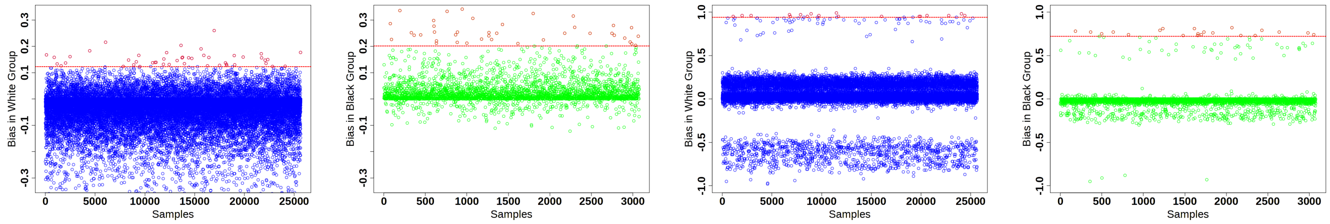

Fig. 1: *Counterfactual discrimination of DNN*: (Left) The observed CD for white with the threshold red line at 0.12; (Mid-Left) The observed CD for black with the threshold line at 0.20. *CD of Mitigated DNN*: (Mid-Right) The observed CD for white with the threshold line sets at 0.81; and (Right) The observed CD for black with the threshold line sets at 0.72.

Dataset. The Adult Census Income dataset [26] is a binary classification dataset used to predict whether an individual has an annual income over $50K$. It consists of $48,842$ instances and 14 attributes. In our study, we consider *race* as the protected attribute and compare the outcomes between white and black individuals.

ML Model and Typical Fairness. We used the same neural network architecture as previous literature on fairness testing [27], [28], [29], which is a six-layered fully-connected neural network with 128 neurons that produces probabilities from the raw logit scores. The model was trained on the Adult Census Income dataset using the `Adam` optimizer with a learning rate of 0.001. The accuracy of the model on the test data is 84%. The true positive rates for white and black individuals are 0.75 and 0.65, respectively. This yields an average odd difference (AOD) of 0.05.

Test-case Generations. The search algorithm samples 25,658 and 3,076 test-cases for white and black groups, respectively. For each sub-group, we compute the likelihood of a favorable outcome for the original sample and its counterfactual, i.e., *counterfactual discrimination* (CD). Figure 1 (left part) shows the CD of these samples for the white and black sub-groups. The mean and standard deviation of CD are -0.04 (+/- 0.05) and 0.03 (+/- 0.04) for the white and black groups.

Inferring Extreme Value Distributions. Figure 1 (left part) shows CD of samples with the threshold (red) lines for white and black groups where red points are extreme values. To infer the parameters of GEV distribution, we set the threshold to 0.12 and 0.2 for white and black groups allowing only 50 samples to exceed the threshold [17]. Given this criterion, the estimated location, scale, and shape are 0.15 (+/- 0.01), 0.03 (+/- 0.01), and -0.08 (+/- 0.01) for white, and 0.28 (+/- 0.01), 0.08 (+/- 0.02), and -0.08 (+/- 0.02) for black, respectively.

Fairness Measures through Extreme Value Distributions. We use the characteristics of GEV distributions to measure the amounts of discrimination between two groups in the tail of the DNN outcomes. Figure 2 shows the GEV density plot for white (left) and black (mid-left) sub-groups. While the average of CDs for these two groups differ by 0.07 (favoring white); the expected extreme CDs differ by 0.13. Crucially, GEV distributions allow us to compute the expected return levels (RL) for a given number of interactions. Table I (original DNN) shows the RLs. For instance, the table indicates that the worst-case CDs of 0.14 and 0.42 are expected in the next 1,000 interactions for white/black sub-groups, respectively.

TABLE I: Return Levels of ECD for original vs. mitigated.

Num. Interactions	Original DNN		Mitigated DNN	
	RL (white)	**RL** (black)	**RL** (white)	**RL** (black)
500	0.12 (+/- 0.01)	0.37 (+/- 0.01)	0.90 (+/- 0.02)	0.78 (+/- 0.1)
1,000	0.14 (+/- 0.03)	0.42 (+/- 0.03)	0.93 (+/- 0.03)	0.80 (+/- 0.18)
2,000	0.16 (+/- 0.04)	0.47 (+/- 0.05)	0.95 (+/- 0.04)	0.81 (+/- 0.19)

Validating Prevalent Fairness Mitigators. We validate the behaviors of popular in-process and pre-process bias mitigators—exponentiated gradient (EG) [20] and Fair-SMOTE [21]—in the tail. When using EG, the average odd difference is 0.02 which shows improvements in the average-case fairness with only 2% accuracy loss. Figure 1 (right parts) shows CD values for the mitigated DNN with a threshold (red) line set at 0.81 and 0.72 for white and black, respectively. The estimated location, scale, and shape of GEV are 0.84 (+/- 0.02), 0.02 (+/- 0.01), and -0.10 (+/- 0.09) for white; and 0.77 (+/- 0.02), 0.04 (+/- 0.02), and -0.05 (+/- 0.05) for Black, respectively, a significant increase in the tail discrimination, see Figure 2 (right parts). The RLs of the mitigated models have also significantly increased, as shown in Table I. Within the white sub-group, in the next 2,000 interactions, we expect an extreme bias of 0.16 in the DNN, whereas an RL of 0.95 is expected for the mitigated DNN. For the black sub-group, the expected RL has increased from 0.47 to 0.81.

Tail-Aware Bias Reduction Algorithms. We first validate the MiniMax-Fainess [24], a tail-aware bias reduction algorithm. Compared to the EG, the MiniMax-Fairness reduces the ECD to -0.27 and degrades the average-based fairness by 0.04. We guide an in-process bias mitigator that reduces the ECD to 0.02 while degrading average-based fairness only by 0.01.

III. Extreme Counterfactual Discrimination

We consider machine learning classifiers with a set of input variables $\mathcal{A}$, which are divided into a protected set of variables Z (e.g., race, sex, and age) and non-protected variables X (e.g., profession, income, and education). A learning problem can be defined as identifying a mapping from the inputs to a probabilistic score of the favorable outcome, inferred from a fixed training dataset $\mathcal{D}_T = \{((\mathbf{x_i}, \mathbf{z_i}), \mathbf{y_i})\}_{i=1}^{N}$, such that the ML model generalizes well to previously unseen situations based on a test dataset $\mathcal{D}_* = \{((\mathbf{x_i^*}, \mathbf{z_i^*}), \mathbf{y_i^*})\}_{i=1}^{M}$.

We abstractly express a machine learning classifier as a function $ML : X \times Z \rightarrow [0, 1]$. The accuracy of model is

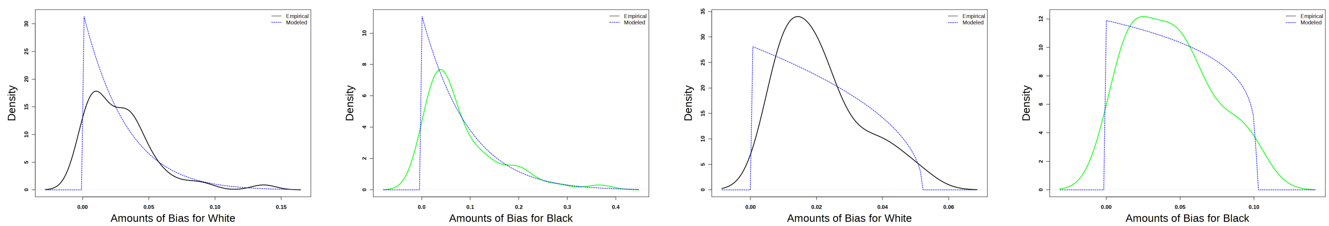

Fig. 2: *The density of GEV for DNN*: (Left) The density for white with the location of 0.15 and scale of 0.03; (Mid-Left) The density for black with the location of 0.28 and the scale of 0.08. *The density of GEV for Mitigated DNN*: (Mid-Right) The density of GEV distributions for white with the location of 0.84 and scale of 0.02; (Right) The density of GEV distributions for black with the location of 0.77 and the scale of 0.04.

measured for the fraction of points in $\mathcal{D}_*$ that satisfy the predicate $ML(x_i^*, z_i^*) \geq 0.5 == y_i^*$. As a convention, we let $Z_i = 1$ indicate membership in a privileged group, and $Z_i = 0$ in the unprivileged group.

Definition III.1 (CD)**.** *Given an individual with non-protected value $X = x$ and protected attribute $Z = z$, the amount of discrimination over the protected attribute i based on the causal fairness notion [30], [14], [27], [28] defines as the difference between ML outcomes over the individual and its counterfactual, i.e., $CD(x, z) = ML(x, z') - ML(x, z)$ where $z'_i = 1 - z_i$ with $z'_j = z_j$ for all protected attributes $1 \leq j \neq i \leq r$ and $-1 \leq CD \leq +1$. The positive values indicate that the ML model disadvantages the individual x in the group z whereas the negative values show unfair advantages for the membership in $Z = z$.*

Considering individuals with $z_i = \{0, 1\}$, the average causal discrimination for sub-groups (**p**rivileged and **u**nprivileged) is: $ACD_p = \mathbb{E}_{z_i=1}\ CD(x, z)$ and $ACD_u = \mathbb{E}_{z_i=0}\ CD(x, z)$.

Previously, THEMIS [14] used z-score testing with an assumption about the normal distribution of counterfactual outcomes to deem discrimination between two groups. From a practical standpoint, it is crucial to ensure fairness on average outcomes (e.g., THEMIS [14]) as well as in the tail.

Definition III.2 (ECD)**.** *Given an ML model and a protected attribute Z_i, our goal is to (1) model the statistics of extreme counterfactual discrimination for each group, i.e., $M_p = \max_{z_i=1} CD(x, z)$ and $M_u = \max_{z_i=0} CD(x, z)$; (2) compute whether the difference between two groups (M_p and M_u) is statistically significant to detect a discrimination in the tail; (3) provide worst-case guarantees on the amounts of discrimination; and (4) mitigate biases in the tail.*

IV. APPROACH

We are interested in determining the maximum values of counterfactual discrimination, denoted as M_p for privileged groups and M_u for unprivileged groups. Since these values for different individuals are independent of each other, we can consider the estimation over a large number of independent and identically distributed random variables.

Extreme value theory is the field of study that examines the limit distributions of such extreme values and the convergence towards these distributions. Our objective, therefore, is to estimate the worst-case counterfactual discrimination by comparing the statistical characteristics of GEV distributions between privileged and unprivileged sub-groups.

However, analyzing extreme values necessitates having an adequate number of samples from the tail behavior of ML models for any given group to have confidence in the results. Our approach comprises three key steps: 1) *Learning the underlying distributions* of the target population to generate valid samples for any sub-group; 2) *Collecting tail samples with statistical guarantees* through a randomized test-case generation algorithm; and 3) *Inferring the tail distributions of counterfactual discrimination* by fitting GEV distributions to each group and comparing the results to determine statistically significant discrimination in worst-case scenarios.

Learning the underlying distributions. The scarcity of samples for some protected groups in datasets can result in statistical uncertainties in extreme value distributions. For instance, in the heart dataset [31], the number of samples for male individuals is notably limited. The conventional approach of sampling data points uniformly at random from the domain of each variable without considering the relationships between variables has the risk of producing samples that do not represent the target group [14], [32], [33], [29]. For example, random generation could result in an income level that is out of line with the general age distribution.

Generative Adversarial Networks (GANs) and Variational Autoencoders (VAE) have been shown to effectively learn and reproduce actual data distributions, making them suitable for generating synthetic data that closely resembles the real-world distributions of sensitive groups [34], [35], [36], [19], [37], [38]. In the GAN paradigm, during the training phase, the generator's primary function is to produce synthetic data samples, while the discriminator is tasked with distinguishing between real and synthetic samples. After multiple rounds of training, the generator learns to generate data so indistinguishable from the original samples. VAEs on the other hand, are trained by encoding input data into a latent representation and recovering it afterward. The decoder then reconstructs the input using the sampled latent points. Training involves optimizing two essential components: the reconstruction loss and the Kullback-Leibler (KL) divergence.

However, in addition to making sure to learn the target distribution of each demographic to alleviate the risk of

Algorithm 1: Tail Sample Generations.

Input: Decision-Support ML Software ML, Generative Adversarial Network GAN, Training Dataset $\mathcal{D}$, Test Samples $\mathcal{D}_*$, Target Group $\mathcal{G}$, Counterfactual Group $\mathcal{G}'$, Low-Bounds on Exp. Test k_{min}, Upper-Bounds on Exp. Test k_{max}, Num. of GAN Samples m, and Timeout $\mathcal{T}$.

1 $Done \leftarrow$ `False`
2 **while** $t \leq \mathcal{T}$ OR $Passed$ **do**
3     $Y \leftarrow ML(\mathcal{D}_*, \mathcal{G})$
4     $Y' \leftarrow ML(\mathcal{D}_*, \mathcal{G}')$
5     $\Delta \leftarrow Y'$ - Y
6     $HQ\_Samples \leftarrow$ `True`
7   **for** $k \leftarrow k_{min}$ *to* k_{max} **do**
8   **if** $size(\Delta) < k_{max}$ **then**
9             $HQ\_Samples \leftarrow$ `False`
10   `Break`
11         $\Theta \leftarrow$ `Select_Top_k`(Δ, k)
12         $\overline{\Theta}$, $\sigma(\Theta) \leftarrow$ `average`(Θ), `std`(Θ)
13         $CV_k \leftarrow \frac{\sigma(\Theta)}{\overline{\Theta}}$
14   **if** $CV_k \geq 1.0 + (\frac{1}{4*k})$ **then**
15             $HQ\_Samples \leftarrow$ `False`
16   `Break`
17   **else**
18             $Passed$, $HQ\_Samples \leftarrow$ `True`, `True`
19   **if** Not $HQ\_Samples$ **then**
20         $\mathcal{D}_* \leftarrow \mathcal{D}_*$ ++ `Data_Generator`$(\mathcal{D}, \mathcal{G}$, m)
21 **return** $\mathcal{D}_*$, Δ

unrealistic data generation, we also need to ensure that they generate samples proportional to the representation of groups in the original dataset. In doing so, we explore Conditional Tabular GAN (CTGAN) [18] and Triplet-based Variational Autoencoder (TVAE) [18]. CTGAN and TVAE are specialized for tabular data, capable of handling mixed variable types and complex relationships, unlike traditional GANs and VAEs which focus more on image data. Previosly, Xiao et al. [39] used CTGAN [18] to generate natural test cases in fairness testing as well. CTGAN allowed them to improve the naturalness of test cases by 20% on average, compared to the baseline [33], [30]. In our application, the generator learns to produce samples for a given protected group as a target that closely reflects the underlying distribution of the group.

Collecting tail samples with statistical guarantees. Algorithm 1 shows our approach to assess the fairness of ML models in the tail. Given a dataset, its protected attribute, the target group, and a ML model; we first initialize the test-case samples $\mathcal{D}_*$ to all samples from the target group (e.g all samples with race white) and compute the likelihood of favorable outcomes of the ML model for this target group (line 3). We do the same for the counterfactual group by flipping the value of the protected attribute (e.g., white to black), (line 4). We set the counterfactual discrimination (Δ) for the target group as the differences in the ML outcomes between the counterfactual and original group (line 5). Then we start our search algorithm to collect enough samples to fit the EVT distributions on the tail. In doing so, we perform the exponential test, adopted from [16], [17] on the current samples $\mathcal{D}_*$ (line 7-18). This test utilized the Coefficient of Variation (CV) to determine the type of extreme value distribution. Specifically, the test goes over k highest values of the counterfactual discrimination and calculates the CV value where k ranges from k_min to k_max (line 13). If for all values of $k \in [k\_min, k\_max]$, the CV is less than $(1.0+\frac{1}{4*k})$, then we are statistically confident that we have enough samples from the tail to infer valid extreme value distribution with exponential or light tails [40], noting that the extra term $(\frac{1}{4*k})$ is to correct the bias in the estimation of CV due to small sample size in the tail [41] (line 17-18). Otherwise, if any values of CV are greater than $(1.0 + \frac{1}{4*k})$, we may not be able to fit an EVT distribution in the tail under the current samples $\mathcal{D}_*$ (line 14-16). Only in this case, we use synthetic data generation methods (CTGAN and TVAE) to generate m data samples, similar to the training data samples of the target group (lines 19-20). We repeat the search until we pass the CV or a timeout occurs.

Inferring the tail distributions of counterfactual discrimination. Given the generated tail samples D_* and the counterfactual discrimination measurements Δ; our final goal is to infer the parameters of GEV distributions to estimate counterfactual discrimination on the tail for each group. Following [17], we initially set the threshold of extreme values to M_{k_max}, i.e., only k_max measurements exceed the threshold. Then, we fit the GEV distribution and analyze the shape of the distributions to decide the validity. If we are statistically confident that the shape is zero or negative ($\xi <= 0$), then the GEV belongs to the type I (exponential) or type III (light), and we compare the expected worst-case values (μ) and the scale (σ).

Given valid GEVs for privileged and unprivileged groups, we measure the amounts of discrimination between them with $\mu_u - \mu_p$, that is the expected worst-case discrimination for a unprivileged group μ_u minus the privileged group μ_p.

Tail-Aware Bias Mitigation. Our approach to mitigating bias in the tail employs in-process bias reduction algorithm via hyperparameter optimization. Our approach extends PARFAIT-ML [42] where we change the objective of optimization from the AOD to ECD while keeping the accuracy constraints the same.

V. EXPERIMENTS

In this section, we first formulate the research questions (RQs). Then, we overview datasets, ML models, bias mitigation algorithms, and our implementations. Finally, we carefully analyze and answer the research questions.

RQ1 (**Generating realistic test cases**) Can the previously proposed algorithm generate realistic data from the underlying distribution of the real dataset?

RQ2 (**Feasibility + Usefulness + Guarantee**) Can extreme value theory (EVT) model and quantify the counterfactual discrimination in the tail of ML outcome distributions with statistical guarantees?

RQ3 (**Average-based Bias Mitigators**) Can we validate the efficacy of the prevalent bias mitigation algorithm [43], [21] via EVT?

RQ4 (**Tail-based Bias Mitigators**) What are the performance of existing tail-aware bias reductions? How does an EVT-based mitigator compare to them?

Dataset. We consider 9 socially critical datasets from the literature of algorithmic fairness. These datasets and their properties are described in Table II. We assume that group 1 is privileged and group 2 is unprivileged.

Training Algorithms and ML Models. We consider 4 popular ML models from the literature. We use a six-layer DNN, following [29], [28], [27]. We trained DNN in TensorFlow [50] and used the same hyperparameters for all tasks with num_epochs, batch_size, and learning_rate are set to 25, 32, and 0.001, respectively. We use the LR, SVM, and Random Forest algorithms from scikit-learn library [51] with the default hyperparameter configuration, similar to [14], [52], [53].

Average-based Bias Mitigation Algorithm. We consider four commonly used (average-based) bias mitigation algorithms, exponentiated gradient (EG) [20] (implemented in both AI Fairness 360 [54] and Fairlearn [43]), Fair-SMOTE [21], MAAT [22], and STEALTH [23]. EG [20] algorithm adapts Lagrange methods to find the multipliers that balance accuracy and fairness. Fair-SMOTE looks for bias in the training data and aims to balance the statistics of sensitive features by generating synthetic samples. MAAT employs a fairness model alongside a performance model to infer the final decision. STEALTH employs a surrogate model to use in predictions and explanations. For evaluating fairness in average-based scenarios, in addition to AOD and EOD metrics, we also included Statistical Parity Difference (SPD), and Disparate Impact (DI) which compare the probabilities of favorable outcomes among protected groups [54].

Tail-aware Bias Reduction. We utilize Minimax-Fairness [24] that takes an iterative game-theoretical approach to reduce the maximum error for protected groups. To investigate the usefulness of the ECD-based mitigator, we adopt a hyperparameter optimization technique, PARFAIT-ML [42] that finds the configurations of ML algorithms to minimize the bias of resultant ML models in the tail. We set ECD as the objective search criteria and run the tool for 1 hour on each benchmark.

Implementation and Technical Details. We run all the experiments on an Ubuntu 20.04.4 LTS server equipped with an AMD Ryzen Threadripper PRO 3955WX CPU and two NVIDIA GeForce RTX 3090 GPUs. We split the dataset into training (60%), validating (20%), and test (20%) data where accuracy, F1, and fairness measures are reported over the test data. We use Fairlearn [43] to quantify the fairness. To measure counterfactual bias, we sample data instances independently and at random for each sub-group. We use the implementation of EG in Fairlearn [43] to study the common bias mitigation algorithm. We repeated each query 100 times and took the average to control the stochastic behavior of the EG with high precision. We obtained the implementation of Minimax-Fairness [24] from their GitHub repository. We also modify the implementation to support training on GPU. We set the $error\_type$, $numsteps$, and $epochs$ to $0/1$ $loss$, 2000, and 50, respectively. We implemented the EVT algorithms in `R` using `evd` and `extRemes` libraries [55]. In Algorithm 1, we set k_{min}, k_{max}, m, $\mathcal{T}$ to 10, 50, 1, and 1200(s), respectively. This choice of k_{min} and k_{max} provides 95% confidence on the feasibility of worst-case guarantees via EVT [17]. We obtained the implementation of Fair-SMOTE [21], MAAT [22], and STEALTH [23] from their GitHub repository and used the recommended configuration to achieve their best results. We repeated each experiment 20 times and conducted 4,400 runs in total. For the statistical tests, we follow prior work [23], [22], [56], [57] and perform a nonparametric test using the Scott-Knott procedure. This involved applying Cliff's Delta and a bootstrap test to assess the results. In our Scott-Knott ranking, we classify results as wins, ties, or losses based on statistically significant improvements, indistinguishable performance, or significant degradations, respectively, compared to the original baseline (vanilla) model. We compare different methods to each other based on number of wins, ties, and losses. **The replication package is available at** https://figshare.com/s/5b4fe7b676e1f7f7b107.

A. *Evaluating Synthetic Data Generation (RQ1)*

We assess the performance of Conditional Tabular GAN (CTGAN) [18] and Triplet-based Variational Autoencoder (TVAE) [18] by comparing their synthetic data against the original dataset, focusing on statistical similarities and distribution characteristics. We aim to determine which model better generates representative test cases for target demographic groups. We also included datasets generated independently at randomly from the domain of variables. For quality assessment, we considered two criteria: similarity to the dataset in several statistical properties and the performance of a downstream ML model trained on generated data versus the actual dataset [58], [59], [60].

Table III shows the results of experiments and the evaluation of metrics for each benchmark. Column `FID` reports the Fréchet Inception Distance [61] (FID) is an inception score-based metric to measure the resemblance between generated and actual datasets. The normalized KL-Divergence, shown in the `KL-D` column, measures the disparity in the informational content between two distributions, with a value of 1.0 indicating minimal divergence. Column `LG-D` indicates the logistic regression detection score [62] that calculates how difficult it is to distinguish real from synthetic data based on the average ROC AUC scores across cross-validation splits. Column `F1 loss` highlights the performance disparities between models trained on actual and synthetic datasets, with values near zero

TABLE II: Datasets used in our experiments.

Dataset	|Instances|	|Features|	Protected Groups		Outcome Label	
			Group 1	*Group 2*	*Label 1*	*Label 0*
Adult *Census* [26] Income	$32,561$	14	Sex-Male	Sex-Female	High Income	Low Income
			Race-White	Race-Black		
German *Credit* [44]	$1,000$	20	Sex-Male	Sex-Female	Good Credit	Bad Credit
Bank Marketing [45]	$45,211$	17	Age-Young	Age-Old	Subscriber	Non-subscriber
Compas [46]	$7,214$	28	Race-Caucasian	Race-Non Caucasian	Did not Reoffend	Reoffend
Default [47]	$30,000$	23	Sex-Male	Sex-Female	Default	Not Default
Heart [31]	297	12	Sex-Male	Sex-Female	Disease	Not Disease
Meps15 [48]	$15,830$	137	Sex-Male	Sex-Female	Utilized Benefits	Not Utilized Benefits
Meps16 [48]	$15,675$	137	Sex-Male	Sex-Female	Utilized Benefits	Not Utilized Benefits
Student Performance [49]	$1,044$	32	Sex-Male	Sex-Female	Pass	Not Pass

TABLE III: Data generation techniques. Legend: **Algorithm**: Generating method, **FID**: Fréchet inception distance, **KL-D**: Kullback–Leibler divergence, **LG-D**: logistic regression detection, **Acc**: accuracy difference, **F1** : Downstream F1 loss.

Dataset	Algorithm	Similarity			ML Perf
		FID	KL-D	LG-D	F1 loss
Adult	CTGAN	.02	**.93**	.74	.03
	TVAE	**.01**	**.93**	**.78**	**.01**
	RND	.11	.19	.01	.25
Compas	CTGAN	**.08**	.97	.61	**.0**
	TVAE	**.08**	**.98**	**.63**	**.0**
	RND	.30	.76	.02	.51
Credit	CTGAN	**.05**	.13	.39	.2
	TVAE	.06	**.15**	**.54**	**.0**
	RND	.07	**.15**	.05	.17
Bank	CTGAN	**.02**	**.88**	**.73**	**.0**
	TVAE	.03	.81	.67	**.0**
	RND	.20	.12	.01	.26
Default	CTGAN	**.03**	**.82**	**.65**	**.01**
	TVAE	.04	.72	.47	.03
	RND	.34	.33	.00	.13
Heart	CTGAN	**.09**	**.93**	**.54**	.16
	TVAE	.13	.77	.35	**.09**
	RND	.13	.31	.15	.39
MEPS15	CTGAN	.26	**.91**	.05	.07
	TVAE	**.06**	.88	**.42**	**.01**
	RND	.78	.89	.00	.45
MEPS16	CTGAN	.26	.9	.06	.10
	TVAE	**.06**	.89	**.39**	**.01**
	RND	.77	**.91**	.00	.34
Students	CTGAN	**.09**	.90	**.22**	.18
	TVAE	**.09**	**.97**	.18	**.02**
	RND	.10	.87	.02	.76

indicating comparable ML performance across both datasets. In this downstream evaluation, we trained two logistic regression models on the actual dataset and the generated data, and then assess their F1 score against the identical test set from the dataset.

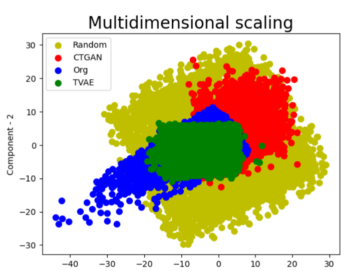

Fig. 3: MDS plot

Our results indicate that both CTGAN and TVAE are effective in learning and replicating the actual data distribution. However, their ability to capture complex feature relationships varies across datasets. For instance, with the Compas dataset, CTGAN's performance stands out: it achieves a KL-divergence of 0.98. Conversely, TVAE shows its strength with the Adult dataset, as supported by all four evaluation metrics. As Table III reveals, data generated randomly tend to deviate significantly from the actual data distribution.

We also employ Multidimensional Scaling (MDS) [63] to visualize these methods. By reducing data to two principal dimensions, MDS provides a visual and analytical means to assess the accuracy with which different generation techniques replicate the characteristics of real dataset. Figure 3 displays this comparison for the Compas dataset, particularly highlighting the alignment of TVAE-generated data with the actual dataset's distribution.

Answer RQ1: CTGAN and TAVE demonstrate their ability to accurately replicate the distribution of actual datasets. In our experiments, they generated data with a KL-Divergence as high as 0.98 and an inception distance as low as 0.008. But, we found that their effectiveness is dataset-dependent.

B. Feasibility, Usefulness, and Guarantee of EVT (RQ2)

One important investigation of this paper is to find out whether Extreme Value Theory (EVT) can effectively model the tail of ML outcome distributions. In Table IV, we present 80 experiments with their corresponding EVT characteristics and the feasibility of EVT to provide fairness guarantees. The number of test cases generated for each group is shown in column $\#N$, determined by the exponential testing in Algorithm 1. The numbers reported in this column include both the original sample size from the dataset and the additional synthetic samples required to pass the test. For instance, a value of 0.1/1.0 indicates that there are 100 original samples with 1000 additional synthetic samples. Columns ACD [14], CVaR [15], and ECD show average, conditional value at risk, and extreme counterfactual discrimination. In the columns (μ, σ, ξ, τ, type) of Table IV, we detail the characteristics of the GEV distribution for each benchmark that informs ECD. Here, μ represents the mean of the extreme value distribution at a specific threshold τ for each combination of algorithm, dataset, and subgroup. For instance, in the DNN application to the Census dataset with sex as protected attribute, we observe an ACD of 0.05, CVaR of 0.08, and ECD of 0.21 where μ_M and μ_F is 0.03 and 0.24, respectively, implying a significant counterfactual discrimination toward female in the tail of DNN's outcome.

The shape ξ indicates the tail behavior of the GEV. A shape ξ around zero or negative suggests that GEV can extrapolate for a long finite (based on Q-Q Plot) or infinite interactions with statistical guarantees, shown with B. In 62 out of 80 scenarios (78%), EVT results in a type III distribution with a negative shape, indicating a finite tail and enabling extrapolation for an unlimited number of queries. For 14 cases (18%), EVT produces a type I distribution with a near-zero

TABLE IV: Characteristics of Extreme Value Distributions. Legend: (*Dataset*) **P**: Protected Attribute, **#N**: Number of test cases, **ACD** = ACD_u - ACD_p: Average Causal Discrimination Difference, **ECD** = $\mu_u - \mu_p$: The Amounts of ECD Tail Discrimination, (*EVT Characteristics*) (μ, σ, ξ, type) for parameters of distributions, τ: threshold. (*Feasibility*) EVT-based extrapolation based on the type of EVT, Q-Q Plot, and its horizon for extrapolations (B). ($\epsilon < 0.01$).

Name				Fairness			EVT Characteristics					Feasibility	
Alg	Dataset	P	#N(k)	ACD [14]	CVaR [15]	ECD	μ	σ	ξ	τ	Type	Q-Q Plot	B
DNN	Census	White	2.56/0	.07	-.23	.13	.15 (+/- .01)	.03 (+/- .01)	-.08 (+/- .01)	.12	III-Finite	Linear (✓)	∞
		Black	3.1/0				.28 (+/- .02)	.08 (+/- .02)	-.08 (+/- .02)	.2	III-Finite	Linear (✓)	∞
	Census	Male	20.1/0	.05	.08	.21	.03 (+/- ϵ)	.01 (+/- ϵ)	.04 (+/- ϵ)	.02	I-Log	Skewed-Right	500
		Female	10.2/0				.24 (+/- .02)	.08 (+/- .02)	-.1 (+/- .02)	.16	III-Finite	Linear (✓)	∞
	Credit	Male	.3/1.0	.05	-.00	.09	.0 (+/- ϵ)	.00 (+/- ϵ)	-141.17 (+/- ϵ)	.00	III-Finite	Linear (✓)	∞
		Female	.7/0				.09 (+/- ϵ)	.01 (+/- ϵ)	-1.05 (+/- ϵ)	.07	III-Finite	Linear (✓)	∞
	Bank	Young	9.4/0	.00	-.05	-.01	.08 (+/- ϵ)	.03 (+/- ϵ)	.17 (+/- ϵ)	.05	II-Infinite	Heavy-Tail (✗)	0
		Old	1.5/0				.07 (+/- .02)	.03 (+/- .02)	-.03 (+/- .02)	.04	I-Log	Skewed-Left	5,000
	Compas	Caucasian	1.4/0	.02	-.01	.11	.02 (+/- ϵ)	.01 (+/- ϵ)	.04 (+/- ϵ)	.02	I-Log	Skewed-Left	500
		Other	3.0/0				.13 (+/- .01)	.02 (+/- .01)	.03 (+/- .01)	.1	I-Log	Skewed-Left	1,000
	Default	Male	11.4/0	.03	-.00	.10	.01 (+/- ϵ)	.01 (+/- ϵ)	.29 (+/- ϵ)	.01	II-Infinite	Heavy-Tail (✗)	0
		Female	17.2/0				.11 (+/- ϵ)	.02 (+/- ϵ)	-.07 (+/- ϵ)	.1	III-Finite	Linear (✓)	∞
	Heart	Male	.1/1.0	-.02	-.00	-.01	.01 (+/- ϵ)	.00 (+/- ϵ)	-.14 (+/- ϵ)	.01	III-Finite	Linear (✓)	∞
		Female	.2/1.0				.00 (+/- ϵ)	.00 (+/- ϵ)	-2.72 (+/- ϵ)	.00	III-Finite	Linear (✓)	∞
	Meps15	Male	8.2/0	.05	-.00	.18	.16 (+/- .01)	.04 (+/- .01)	-.07 (+/- .01)	.11	III-Finite	Linear (✓)	∞
		Female	7.6/0				.34 (+/- .01)	.05 (+/- .01)	-.09 (+/- .01)	.29	III-Finite	Linear (✓)	∞
	Meps16	Male	8.3/0	.07	.00	.20	.18 (+/- .01)	.06 (+/- .01)	.02 (+/- .01)	.12	I-Log	Skewed-Left	5,000
		Female	7.4/0				.38 (+/- .02)	.06 (+/- .02)	-.24 (+/- .02)	.31	III-Finite	Linear (✓)	∞
	Students	Male	.6/0	.02	-.00	.09	.00 (+/- ϵ)	.00 (+/- ϵ)	-347.33 (+/- ϵ)	.00	III-Finite	Linear (✓)	∞
		Female	.5/0				.09 (+/- .01)	.04 (+/- .01)	-.27 (+/- .01)	.04	III-Finite	Linear (✓)	∞
LR	Census	White	25.7/0	.09	-.19	.11	.0 (+/- ϵ)	.00 (+/- ϵ)	.00 (+/- ϵ)	.00	I-Log	Skewed-Left	500
		Black	3.1/0				.11 (+/- ϵ)	.00 (+/- ϵ)	-1.24 (+/- ϵ)	.11	III-Finite	Linear (✓)	∞
	Census	Male	20.1/0	.06	.22	.07	.0 (+/- ϵ)	.0 (+/- ϵ)	-1.0 (+/- ϵ)	.00	III-Finite	Linear (✓)	∞
		Female	10.2/0				.07 (+/- ϵ)	.00 (+/- ϵ)	-1.62 (+/- ϵ)	.07	III-Finite	Linear (✓)	∞
	Credit	Male	.3/.5	.09	-.01	.1	-.02 (+/- ϵ)	.01 (+/- ϵ)	-.74 (+/- ϵ)	-.03	III-Finite	Linear (✓)	∞
		Female	.7/0				.08 (+/- ϵ)	.00 (+/- ϵ)	-1.0 (+/- ϵ)	.08	III-Finite	Linear (✓)	∞
	Bank	Young	9.4/0	.0	-.09	.01	.00 (+/- ϵ)	.00 (+/- ϵ)	-1.0 (+/- ϵ)	.00	III-Finite	Linear (✓)	∞
		Old	1.5/0				.01 (+/- ϵ)	.00 (+/- ϵ)	-1.33 (+/- ϵ)	.01	III-Finite	Linear (✓)	∞
	Compas	Caucasian	1.5/0	-.04	.0	-.06	.00 (+/- ϵ)	.00 (+/- ϵ)	-.73 (+/- ϵ)	.00	III-Finite	Linear (✓)	∞
		Other	3.0/0				.06 (+/- ϵ)	.01 (+/- ϵ)	-.54 (+/- ϵ)	.06	III-Finite	Linear (✓)	∞
	Default	Male	11.4/0	.04	-.07	.03	.00 (+/- ϵ)	.00 (+/- ϵ)	-.66 (+/- ϵ)	.00	III-Finite	Linear (✓)	∞
		Female	17.2/0				.03 (+/- ϵ)	.00 (+/- ϵ)	-1.07 (+/- ϵ)	.03	III-Finite	Linear (✓)	∞
	Heart	Male	.1/1.0	-.02	.02	-.01	.01 (+/- ϵ)	.01 (+/- ϵ)	-1.0 (+/- ϵ)	.00	III-Finite	Linear (✓)	∞
		Female	.2/.5				.00 (+/- ϵ)	.00 (+/- ϵ)	-1.0 (+/- ϵ)	.00	III-Finite	Linear (✓)	∞
	Meps15	Male	8.2/0	.06	-.1	.07	.00 (+/- ϵ)	.00 (+/- ϵ)	-.44 (+/- ϵ)	.00	III-Finite	Linear (✓)	∞
		Female	7.6/0				.07 (+/- ϵ)	.00 (+/- ϵ)	-1.65 (+/- ϵ)	.07	III-Finite	Linear (✓)	∞
	Meps16	Male	8.3/0	.1	-.13	.12	.0 (+/- ϵ)	.00 (+/- ϵ)	-.14 (+/- ϵ)	.00	III-Finite	Linear (✓)	∞
		Female	7.4/0				.12 (+/- ϵ)	.00 (+/- ϵ)	-.99 (+/- ϵ)	.12	III-Finite	Linear (✓)	∞
	Students	Male	.6/0	.04	.0	.1	.00 (+/- ϵ)	.00 (+/- ϵ)	-1.06 (+/- ϵ)	.00	III-Finite	Linear (✓)	∞
		Female	.5/0				.1 (+/- ϵ)	.01 (+/- ϵ)	-2.92 (+/- ϵ)	.08	III-Finite	Linear (✓)	∞
SVM	Census	White	25.7/0	.04	-.11	.03	.28 (+/- .02)	.04 (+/- .02)	-.52 (+/- .02)	.23	III-Finite	Linear (✓)	∞
		Black	3.1/0				.31 (+/- .02)	.06 (+/- .02)	-.33 (+/- .02)	.24	III-Finite	Linear (✓)	∞
	Census	Male	20.1/0	.04	.08	.15	.01 (+/- ϵ)	.00 (+/- ϵ)	-.75 (+/- ϵ)	.01	III-Finite	Linear (✓)	∞
		Female	10.2/0				.16 (+/- .01)	.02 (+/- .01)	-.06 (+/- .01)	.14	III-Finite	Linear (✓)	∞
	Credit	Female	.3/.5	.01	-.03	.02	.01 (+/- ϵ)	.01 (+/- ϵ)	-.07 (+/- ϵ)	.01	III-Finite	Linear (✓)	∞
		Male	.7/0				.03 (+/- ϵ)	.00 (+/- ϵ)	-.41 (+/- ϵ)	.02	III-Finite	Linear (✓)	∞
	Bank	Young	9.4/0	.01	-.09	.07	.1 (+/- .02)	.05 (+/- .02)	.01 (+/- .02)	.05	I-Log	Skewed-Left	1,000
		Old	1.4/0				.17 (+/- .01)	.05 (+/- .01)	.03 (+/- .01)	.12	I-Log	Skewed-Left	1,000
	Compas	Caucasian	1.4/0	.0	.0	.01	.00 (+/- ϵ)	.00 (+/- ϵ)	.12 (+/- ϵ)	.00	II-Infinite	Heavy-Tail (✗)	0
		Other	3.0/0				.01 (+/- ϵ)	.01 (+/- ϵ)	.4 (+/- ϵ)	.01	I-Log	Skewed-Left	2,000
	Default	Male	11.4/0	.0	-.01	.01	.03 (+/- ϵ)	.00 (+/- ϵ)	.00 (+/- ϵ)	.02	I-Log	Skewed-Left	2,000
		Female	17.2				.04 (+/- ϵ)	.01 (+/- ϵ)	-.33 (+/- ϵ)	.03	III-Finite	Linear (✓)	∞
	Heart	Male	.1/1.0	.0	.04	-.01	.01 (+/- ϵ)	.00 (+/- ϵ)	-.18 (+/- ϵ)	.00	III-Finite	Linear (✓)	∞
		Female	.2/.5				.00 (+/- ϵ)	.00 (+/- ϵ)	.00 (+/- ϵ)	.00	I-Log	Skewed-Left	5,000
	Meps15	Male	8.2/0	.02	-.07	.07	.01 (+/- ϵ)	.00 (+/- ϵ)	-.07 (+/- ϵ)	.00	III-Finite	Linear (✓)	∞
		Female	7.6/0				.08 (+/- .01)	.02 (+/- .01)	-.35 (+/- .01)	.06	III-Finite	Linear (✓)	∞
	Meps16	Male	8.3/0	.03	-.13	.12	.00 (+/- ϵ)	.00 (+/- ϵ)	-.26 (+/- ϵ)	.00	III-Finite	Linear (✓)	∞
		Female	7.4/0				.12 (+/- ϵ)	.01 (+/- ϵ)	-.5 (+/- ϵ)	.11	III-Finite	Linear (✓)	∞
	Students	Male	.6/0	.0	.0	.0	.01 (+/- ϵ)	.00 (+/- ϵ)	.36 (+/- ϵ)	.00	II-Infinite	Heavy-Tail (✗)	0
		Female	.5/0				.01 (+/- ϵ)	.01 (+/- ϵ)	.03 (+/- ϵ)	.00	I-Log	Skewed-Left	10,000
RF	Census	White	25.7/0	.04	-.11	.01	.54 (+/- .02)	.07 (+/- .02)	-.16 (+/- .02)	.47	III-Finite	Linear (✓)	∞
		Black	3.1/0				.55 (+/- .03)	.1 (+/- .03)	-.41 (+/- .03)	.42	III-Finite	Linear (✓)	∞
	Census	Male	20.1/0	-.04	.08	-.18	.36 (+/- .01)	.07 (+/- .01)	.00 (+/- .01)	.29	I-Log	Skewed-Left	1,000
		Female	10.3/0				.54 (+/- .02)	.08 (+/- .02)	-.27 (+/- .02)	.44	III-Finite	Linear (✓)	∞
	Credit	Male	.3/1.0	.01	-.03	.03	.07 (+/- .01)	.02 (+/- .01)	-.34 (+/- .01)	.04	III-Finite	Linear (✓)	∞
		Female	.7/0				.1 (+/- .01)	.03 (+/- .01)	.01 (+/- .01)	.07	I-Log	Skewed-Right	500
	Bank	Young	9.5/0	-.01	-.09	-.02	.15 (+/- .01)	.03 (+/- .01)	-.49 (+/- .01)	.11	III-Finite	Linear (✓)	∞
		Old	1.5/0				.13 (+/- .01)	.04 (+/- .01)	-.23 (+/- .01)	.08	III-Finite	Linear (✓)	∞
	Compas	Caucasian	1.5/0	.02	.0	.18	.36 (+/- .02)	.08 (+/- .02)	-.74 (+/- .02)	.23	III-Finite	Linear (✓)	∞
		Other	3.0/0				.54 (+/- .03)	.06 (+/- .03)	-.69 (+/- .03)	.45	III-Finite	Linear (✓)	∞
	Default	Male	11.4/0	-.01	-.01	-.02	.65 (+/- .03)	.09 (+/- .03)	-.21 (+/- .03)	.54	III-Finite	Linear (✓)	∞
		Female	17.2/0				.63 (+/- .02)	.06 (+/- .02)	-.43 (+/- .02)	.56	III-Finite	Linear (✓)	∞
	Heart	Male	.1/2.0	.0	.04	.0	.05 (+/- .01)	.03 (+/- .01)	-.21 (+/- .01)	.01	III-Finite	Linear (✓)	∞
		Female	.2/1.5				.05 (+/- .01)	.04 (+/- .01)	-.12 (+/- .01)	.01	III-Finite	Linear (✓)	∞
	Meps15	Male	8.2/0	-.02	-.07	-.15	.51 (+/- ϵ)	.13 (+/- ϵ)	-1.04 (+/- ϵ)	.34	III-Finite	Linear (✓)	∞
		Female	7.6/0				.36 (+/- .04)	.13 (+/- .04)	-.21 (+/- .04)	.22	III-Finite	Linear (✓)	∞
	Meps16	Male	8.3/0	-.01	-.13	-.03	.38 (+/- .05)	.14 (+/- .05)	-.4 (+/- .05)	.22	III-Finite	Linear (✓)	∞
		Female	7.4/0				.35 (+/- .03)	.08 (+/- .03)	-.38 (+/- .03)	.24	III-Finite	Linear (✓)	∞
	Students	Male	.6/0	.0	.0	.01	.03 (+/- .01)	.02 (+/- .01)	-.38 (+/- .01)	.01	III-Finite	Linear (✓)	∞
		Female	.5/0				.04 (+/- .01)	.02 (+/- .01)	-.29 (+/- .01)	.02	III-Finite	Linear (✓)	∞

shape, implying an infinite but exponentially decaying tail, suitable for extrapolation within bounded queries B. Overall, the worst-case guarantees are achievable in 76 cases (95%).

We examine the relevance of extreme counterfactual discrimination in ML model fairness by employing EVT to measure tail biases, comparing them to established fairness metrics like ACD and CVaR. For instance, in the DNN model trained on the Compas dataset, an ACD of 0.02 and a CVaR of -0.01 indicate fairness in both average and tail cases, yet an ECD of 0.11 suggests a tail-bias toward Caucasians. We classify any ECD difference exceeding 0.05 as discrimination, with its significance indicated by the grayscale in the ECD column. Out of 40 cases, ECD-based discrimination occurs in 19 (48%). In contrast, average-case discrimination (ACD) is observed in 10 out of 40 cases (25%). Notably, in 13 cases (33%), ECD is significantly greater than ACD. In 18 out of 40 experiments, ECD found significant discrimination against the unprivileged group in the tail that missed by the CVaR metric.

Answer RQ2: EVT effectively models extreme counterfactual discrimination (ECD), in 95% of cases, allowing for valid extrapolation of worst-case discrimination. In 33% of cases, ECD shows significantly higher discrimination than the average-case one (ACD [14]). In 18 out of 40 experiments, ECD found significant discrimination against the unprivileged group in the tail that missed by prevalent tail-based metric (CVaR [15]).

C. Validation of Prevalent Bias Mitigation Algorithms (RQ3)

In this analysis, we leverage EVT to assess the effectiveness of prevalent mitigation algorithms like exponentiated gradient (EG) [20] and Fair-SMOTE [21] in the tail. We also include two recent mitigation techniques, MAAT [22] and STEALTH [23] in our experiments to evaluate our approach against more advanced methods. The results are reported in Table V and VI. The column Accuracy Loss shows the accuracy difference between the original and mitigated models with positive values indicating improved accuracy in the mitigated model, columns AOD, EOD, SPD, and DI report the absolute values of average-based fairness measures, and the column ECD shows the amount of discrimination in the tail. Darker gray shades indicate lower rankings, while lighter shades represent higher rankings (no shading indicates the top-ranked method).

We use the Scott-Knott ranking outcomes to compare the four mitigation methods where we consider a statistically significant improvement over the original baseline model (the vanilla model) as a win. While the tables include all metrics, we explain the results for one average-based metric and one tail-based metric. Consider the AOD metric, we find that EG [20] outperforms other methods where it wins in 19 cases (out of 40). STEALTH [23], MAAT [22], and Fair-SMOTE [21] win in 14, 6, and 4 cases in reducing AOD biases. In terms of average AOD over all benchmarks; EG, STEALTH, MAAT, and SMOTE achieve an average of 0.03, 0.07, 0.07, and 0.08, respectively. In terms of number of cases with an AOD bias below or equal to 0.05; we observe that EG, STEALTH, MAAT, and SMOTE have 32, 22, 22, and 9 cases (out of 40), respectively.

When considering ECD metric, STEALTH demonstrates superior performance among the average-based mitigation methods in reducing tail discrimination. Specifically, we find that STEALTH wins in 31 cases (out of 40) whereas EG, MAAT, and Fair-SMOTE win in 15, 12, and 9 cases, respectively. In terms of average ECD over all benchmarks; STEALTH, EG, MAAT, and Fair-SMOTE achieve an average of 0.04, 0.20, 0.08, and 0.12, respectively. In terms of number of cases with an ECD bias below or equal to 0.05; STEALTH, EG, MAAT, and Fair-SMOTE have 32, 11, 15, and 9 cases (out of 40), respectively.

Answer RQ3: While the average-based mitigation methods [20], [21], [22], [23] preserved or improved fairness based on metrics like AOD in 63%, they increase unfairness in tail based on ECD metric in 35% of cases. STEALTH [23] outperformed other mitigation methods significantly based on the ECD metric, failing only in 10% of cases, while preserving/reducing the AOD bias in 65%.

D. Tail-aware Mitigation Algorithms (RQ4)

We first evaluate the effectiveness of MiniMax-Fairness [24], which serves as our baseline, alongside our proposed in-process mitigator (ECD-Fair). Results in Table VII follow a similar format to Table V where we only include the DNN and Logistic regression models since the MiniMax-Fairness only supports these models among our base models. Considering ECD metric, our approach significantly outperforms MiniMax-Fairness. Specifically, ECD-Fair wins in 18 cases (out of 20), while MiniMax-Fairness wins in 10 cases (out of 20). When considering EOD and AOD metrics, ECD-Fair outperforms MiniMax-Fairness with 10 and 9 win cases vs. 7 and 5 win cases (out of 20). In terms of absolute values over all benchmarks, ECD-Fair achieves a average AOD and ECD of 0.04 and 0.03, respectively. The number of cases with AOD and ECD below 0.05 are 15 and 18 (out of 20), respectively.

We also compare ECD-Fair to STEALTH [23] method over the DNN and LR benchmarks as STEALTH outperformed other baseline methods. Based on the EOD and AOD metrics, we find that ECD-Fair wins in 10 and 9 cases (out of 20) vs. STEALTH wins in 6 and 8 cases (out of 20), respectively. When considering the ECD metric, ECD-Fair and STEALTH win in 18 and 16 cases (out of 20), respectively. STEALTH degrades unfairness in tail for 2 benchmarks, while ECD-fair does not increase the unfairness in the tail for any benchmark. Overall, while STEALTH demonstrates a competitive result, ECD-Fair slightly outperforms it for both tail and average metrics.

TABLE V: Average based bias mitigation. Legend: **P**: Protected Attribute, **Acc loss**:Accuracy loss in mitigation , **AOD, EOD, SPD, DI**: Average-based Fairness Measures, **ECD** = $\mu_u - \mu_p$: The Amounts of ECD Tail Discrimination, **NV**: Not Valid.

Name			Exponentiated Gradient (EG) [20]						Fair-SMOTE [21]					
Algorithm	Dataset	P	*Acc loss*	*AOD*	*EOD*	*SPD*	*DI*	*ECD*	Acc loss	*AOD*	*EOD*	*SPD*	*DI*	*ECD*
DNN Avg-based	census	race	−0.01	0.04	0.08	0.06	0.72	0.08	−0.04	0.08	0.11	0.13	1.49	0.1
		sex	−0.03	0.01	0.02	0.09	0.49	0.11	−0.07	0.03	0.05	0.19	0.68	0.2
	credit	sex	−0.09	0.02	0.05	0.02	0.55	0.65	−0.01	0.09	0.15	0.09	0.4	0.11
	bank	age	−0.01	0.03	0.05	0.01	0.4	0.09	0.0	0.03	0.06	0.02	0.26	0.06
	compas	race	−0.01	0.03	0.0	0.08	0.14	0.09	−0.02	0.03	0.01	0.07	0.13	0.15
	default	sex	0.0	0.01	0.01	0.02	0.18	0.3	−0.02	0.03	0.04	0.04	0.25	0.06
	heart	sex	−0.06	0.02	0.03	0.06	0.23	0.17	−0.08	0.15	0.26	0.18	0.45	0.12
	meps15	sex	−0.04	0.02	0.02	0.06	0.61	0.05	−0.03	0.04	0.05	0.06	0.65	0.09
	meps16	sex	0.01	0.01	0.01	0.05	0.57	0.42	−0.03	0.05	0.08	0.07	0.87	0.04
	students	sex	−0.07	0.02	0.02	0.03	0.03	0.09	0.02	0.1	0.05	0.06	0.07	0.05
LR Avg-based	census	race	−0.03	0.04	0.07	0.05	0.53	0.47	0.03	0.3	0.47	0.24	8.14	0.4
		sex	−0.01	0.02	0.03	0.07	0.47	0.19	0.02	0.12	0.13	0.21	0.71	0.29
	credit	sex	0.0	0.07	0.1	0.07	0.5	0.02	0.0	0.11	0.17	0.08	0.41	0.09
	bank	age	−0.11	0.05	0.07	0.03	0.44	0.13	0.0	0.14	0.25	0.06	2.28	0.22
	compas	race	−0.05	0.01	0.01	0.05	0.09	0.22	−0.01	0.03	0.0	0.07	0.13	0.07
	default	sex	−0.05	0.02	0.02	0.02	0.28	0.1	0.02	0.08	0.12	0.05	0.29	0.19
	heart	sex	−0.06	0.12	0.24	0.09	0.45	0.37	0.19	0.19	0.33	0.17	0.48	0.06
	meps15	sex	−0.03	0.02	0.03	0.05	0.58	0.09	−0.03	0.03	0.05	0.05	0.75	0.03
	meps16	sex	0.01	0.01	0.02	0.04	0.63	0.79	−0.03	0.08	0.13	0.08	1.38	0.26
	students	sex	−0.03	0.04	0.03	0.06	0.07	0.05	−0.01	0.08	0.05	0.05	0.07	NV
SVM Avg-based	census	race	−0.05	0.02	0.04	0.04	0.54	0.31	−0.01	0.18	0.29	0.16	3.47	0.08
		sex	−0.04	0.01	0.03	0.07	0.53	0.72	0.07	0.11	0.13	0.19	0.72	0.43
	credit	sex	−0.08	0.04	0.06	0.05	0.23	0.23	−0.02	0.09	0.13	0.07	0.65	0.04
	Bank	age	−0.06	0.01	0.02	−0.0	0.21	0.03	0.01	0.05	0.1	0.02	1.07	0.24
	compas	race	−0.03	0.03	0.0	0.05	0.08	0.02	0.01	0.03	0.0	0.07	0.13	0.06
	default	sex	−0.02	0.02	0.03	0.02	0.29	0.02	−0.02	0.02	0.04	0.02	0.15	0.1
	heart	sex	0.03	0.06	0.19	0.13	0.32	0.1	0.08	0.2	0.33	0.19	0.55	0.06
	meps15	sex	0.0	0.02	0.02	0.04	0.53	0.3	-0.02	0.02	0.04	0.04	0.54	0.1
	meps16	sex	−0.04	0.02	0.03	0.03	0.56	0.52	−0.03	0.02	0.04	0.04	0.7	0.13
	students	sex	−0.09	0.02	0.04	0.02	0.03	0.72	0.03	0.08	0.02	0.06	0.06	0.03
RF Avg-based	census	race	−0.03	0.09	0.11	0.13	1.58	0.07	−0.01	0.1	0.13	0.14	1.85	0.17
		sex	−0.04	0.08	0.07	0.18	0.69	0.04	0.08	0.1	0.11	0.18	0.71	0.07
	credit	sex	−0.05	0.06	0.17	0.08	0.31	NV	0.01	0.08	0.12	0.08	0.53	0.15
	bank	age	0.01	0.01	0.03	0.01	0.18	0.17	0.01	0.02	0.04	0.01	0.27	0.05
	compas	race	−0.02	0.02	0.01	0.07	0.11	0.03	0.0	0.03	0.01	0.07	0.13	0.05
	default	sex	−0.07	0.01	0.02	0.03	0.18	0.02	0.08	0.02	0.03	0.03	0.23	0.06
	heart	sex	−0.09	0.07	0.08	0.18	0.43	0.06	−0.08	0.23	0.36	0.18	0.64	0.12
	meps15	sex	−0.05	0.05	0.07	0.07	0.86	0.04	−0.02	0.03	0.05	0.04	0.44	0.16
	meps16	sex	−0.02	0.02	0.03	0.04	0.57	0.01	−0.03	0.02	0.04	0.04	0.52	0.01
	students	sex	−0.01	0.07	0.01	0.02	0.03	NV	0.0	0.05	0.04	0.05	0.06	0.01

TABLE VI: Average-based bias mitigation (STEALTH [23] and MAAT [22]). Legend is similar to Table V

Name			STEALTH [23]						MAAT [22]					
Algorithm	Dataset	P	*Acc loss*	*AOD*	*EOD*	*SPD*	*DI*	*ECD*	Acc loss	*AOD*	*EOD*	*SPD*	*DI*	*ECD*
DNN Avg-based	census	race	−0.05	0.08	0.12	0.09	2.74	0.03	0.02	0.02	0.04	0.14	1.17	0.13
		sex	0.05	0.18	0.27	0.17	0.87	0.07	0.11	0.1	0.2	0.08	0.43	0.08
	credit	sex	0.01	0.06	0.1	0.06	0.43	0.1	0.19	0.1	0.2	0.07	0.47	0.02
	bank	age	0.0	0.02	0.03	0.0	0.34	0.02	0.06	0.03	0.06	0.02	0.32	0.05
	compas	race	−0.01	0.03	0.0	0.07	0.13	0.03	0.01	0.02	0.0	0.07	0.13	0.11
	default	sex	−0.03	0.02	0.03	0.02	0.23	0.01	0.09	0.04	0.07	0.06	0.41	0.11
	heart	sex	0.0	0.14	0.24	0.12	0.61	0.08	0.06	0.19	0.38	0.1	0.47	0.02
	meps15	sex	−0.03	0.03	0.05	0.04	0.66	0.01	0.08	0.03	0.06	0.08	0.73	0.12
	meps16	sex	−0.04	0.02	0.04	0.04	0.65	0.01	0.08	0.03	0.05	0.08	0.83	0.08
	students	sex	0.05	0.03	0.03	0.02	0.03	0.02	0.1	0.08	0.03	0.06	0.08	0.02
LR Avg-based	census	race	−0.06	0.17	0.25	0.17	4.59	0.12	0.06	0.29	0.52	0.22	5.0	0.16
		sex	0.04	0.19	0.29	0.18	0.87	0.06	0.09	0.05	0.11	0.1	0.52	0.18
	credit	sex	0.01	0.09	0.17	0.06	0.47	0.03	0.14	0.1	0.19	0.07	0.68	0.04
	bank	age	−0.01	0.0	0.01	0.0	0.6	0.02	0.05	0.12	0.22	0.04	0.93	0.13
	compas	race	−0.01	0.03	0.01	0.07	0.13	0.04	−0.01	0.02	0.01	0.07	0.12	0.1
	default	sex	−0.05	0.02	0.03	0.01	0.22	0.02	0.06	0.05	0.09	0.01	0.08	0.12
	heart	sex	0.09	0.21	0.37	0.16	0.68	0.04	0.13	0.18	0.36	0.13	0.48	0.02
	meps15	sex	−0.03	0.02	0.04	0.04	0.56	0.02	0.05	0.03	0.06	0.06	0.66	0.08
	meps16	sex	−0.03	0.03	0.05	0.04	0.74	0.03	0.06	0.02	0.03	0.07	0.77	0.1
	students	sex	0.0	0.08	0.04	0.05	0.06	0.02	0.08	0.07	0.03	0.05	0.07	0.02
SVM Avg-based	census	race	−0.06	0.11	0.18	0.09	4.98	0.06	−0.06	0.1	0.16	0.16	1.84	0.08
		sex	0.04	0.16	0.27	0.14	0.92	0.07	0.09	0.03	0.07	0.12	0.56	0.13
	credit	sex	0.02	0.05	0.09	0.04	0.64	0.01	0.06	0.1	0.17	0.05	0.28	0.03
	bank	age	−0.0	0.01	0.02	0.0	0.75	0.0	0.03	0.05	0.1	0.02	0.56	0.15
	compas	race	−0.01	0.03	0.0	0.07	0.13	0.01	−0.0	0.02	0.0	0.07	0.13	0.11
	default	sex	−0.05	0.01	0.02	0.01	0.21	0.01	0.04	0.02	0.05	0.02	0.12	0.12
	heart	sex	0.09	0.18	0.35	0.15	0.62	0.02	0.14	0.19	0.38	0.12	0.4	0.02
	meps15	sex	−0.03	0.02	0.04	0.03	0.64	0.0	0.03	0.03	0.08	0.05	0.47	0.12
	meps16	sex	0.04	0.02	0.03	0.03	0.57	0.0	0.02	0.05	0.12	0.04	0.43	0.03
	students	sex	−0.01	0.08	0.02	0.05	0.06	0.0	0.1	0.07	0.02	0.05	0.07	0.01
RF Avg-based	census	race	−0.05	0.08	0.15	0.08	3.14	0.06	0.08	0.14	0.26	0.16	1.7	0.06
		sex	0.05	0.15	0.24	0.13	0.87	0.05	0.11	0.09	0.17	0.09	0.46	0.06
	credit	sex	0.01	0.07	0.12	0.05	0.87	0.03	0.21	0.13	0.23	0.07	0.39	0.02
	bank	age	−0.01	0.0	0.01	0.0	0.88	0.01	0.04	0.03	0.05	0.01	0.24	0.07
	compas	race	0.0	0.02	0.0	0.07	0.13	0.03	0.0	0.02	0.0	0.07	0.13	0.09
	default	sex	−0.03	0.01	0.02	0.01	0.23	0.02	0.07	0.02	0.03	0.04	0.32	0.12
	heart	sex	0.09	0.17	0.3	0.14	0.86	0.03	0.16	0.21	0.43	0.11	0.44	0.02
	meps15	sex	−0.03	0.03	0.05	0.04	0.65	0.0	0.06	0.03	0.05	0.07	0.67	0.11
	meps16	sex	−0.03	0.02	0.04	0.04	0.52	0.01	0.06	0.02	0.06	0.06	0.68	0.05
	students	sex	−0.01	0.06	0.04	0.05	0.06	0.01	0.06	0.07	0.01	0.08	0.1	0.01

Answer RQ4: ECD-Fair significantly outperformed MiniMax-Fairness [24], a state-of-the-art tail-aware mitigator. When compared to STEALTH [23], a competitive baseline, we found that ECD-Fair and STEALTH improved fairness in the tail for 90% and 80% of cases, respectively. ECD-Fair and STEALTH reduced the AOD bias in 45% and 40% of cases, respectively.

VI. Discussions

Limitations. One limitation is the lack of ground truth regarding the tail of ML outcome distributions. We can use the maximum individual discrimination in the validation dataset as it gives a lower-bound on the ground truth. Our approach requires the presence of protected attributes during inference. Therefore, it cannot be used to study worst-case fairness for notions such as fairness through unawareness, which requires the removal of protected attributes [12]. Our approach also

TABLE VII: Tail-aware bias mitigation. Legend is similar to Table V

Name			Minimax-Fairness [24]						ECD-Fair					
Algorithm	Dataset	P	*Acc loss*	*AOD*	*EOD*	*SPD*	*DI*	*ECD*	*Acc loss*	*AOD*	*EOD*	*SPD*	*DI*	*ECD*
DNN Tail-based	Census	race	−0.07	0.02	0.04	0.02	0.74	0.14	−0.04	0.08	0.09	0.14	1.65	0.02
		sex	−0.04	0.03	0.05	0.06	0.61	0.03	0.02	0.02	0.03	0.06	0.68	0.04
	Credit	sex	−0.09	0.02	0.04	0.02	0.79	0.02	−0.06	0.04	0.08	0.05	0.21	0.02
	Bank	age	−0.08	0.06	0.05	0.07	0.36	0.05	0.02	0.02	0.04	0.01	0.18	0.02
	Compas	race	−0.02	0.02	0.03	0.04	0.08	0.17	0.02	0.02	0.0	0.06	0.11	0.03
	Default	sex	0.02	0.01	0.02	0.02	0.17	0.08	−0.01	0.04	0.05	0.04	0.32	0.06
	heart	sex	0.06	0.11	0.22	0.06	0.29	0.01	0.04	0.1	0.26	0.16	0.43	0.01
	Meps15	sex	0.0	0.01	0.02	0.04	0.36	0.04	0.03	0.03	0.03	0.06	0.63	0.04
	Meps16	sex	0.04	0.02	0.03	0.04	0.69	0.02	0.0	0.04	0.06	0.05	0.93	0.08
	Students	sex	0.02	0.06	0.02	0.03	0.04	0.21	−0.03	0.06	0.02	0.04	0.04	0.03
LR Tail-based	Census	race	−0.06	0.1	0.13	0.12	3.01	0.17	−0.01	0.03	0.05	0.03	3.38	0.03
		sex	0.04	0.1	0.13	0.15	0.77	0.07	−0.03	0.02	0.04	0.03	0.73	0.05
	Credit	sex	−0.03	0.1	0.17	0.11	0.62	0.13	−0.05	0.05	0.09	0.04	0.15	0.0
	Bank	age	−0.05	0.05	0.09	0.01	0.16	0.02	0.02	0.06	0.11	0.02	0.85	0.04
	Compas	race	−0.01	0.02	0.01	0.06	0.11	0.04	0.04	0.04	0.01	0.08	0.15	0.04
	Default	sex	−0.03	0.03	0.05	0.03	0.46	0.03	−0.01	0.03	0.05	0.03	0.43	0.03
	Heart	sex	−0.07	0.12	0.2	0.21	0.63	0.1	−0.04	0.09	0.27	0.13	0.42	0.02
	Meps15	sex	0.0	0.06	0.09	0.07	1.07	0.06	−0.01	0.03	0.03	0.04	0.77	0.02
	Meps16	sex	−0.04	0.07	0.1	0.07	1.16	0.07	−0.01	0.03	0.04	0.04	0.68	0.03
	Students	sex	0.03	0.04	0.02	0.04	0.04	0.08	0.0	0.04	0.02	0.03	0.04	0.01

depends on the representative individuals sampled from the same training distribution, and may not be valid for out-of-distribution queries. Finally, our approach assumes that flipping the sensitive values leads to valid representations to measure the sensitivity of ML models to the protected attributes.

Threat to Validity. To address the internal validity and ensure our finding does not lead to invalid conclusions, we follow established guidelines and report the statistical significance of measures with the exponential and Scott-Knott statistical testing. To ensure that our results are generalizable, we perform our experiments on three well-established training algorithms from `scikit-learn` and `TensorFlow` libraries with a popular mitigation algorithm from the `Fairlearn` library over 160 fairness-sensitive tasks that have been widely used in the fairness research. It is an open problem whether the algorithms, hyperparameters, and datasets are sufficiently representative to cover challenging fairness scenarios.

VII. Related Work

Fairness Testing of Data-Driven Software. Individual discrimination is a major fairness debugging method [64], [52], [27], [28], [21], [65]. THEMIS [14] is the closest approach. While THEMIS [14] focuses on the average causal discrimination between two subgroups via *counterfactual* queries with prevalent statistical guarantees of normal distributions, we introduce the notion of extreme causal discrimination between two subgroups with exponentially statistical guarantees of extreme value distributions. Rather than randomly sampling data from the domain of variables, we leveraged generative AI models to produce realistic test cases from the tail.

Fairness in the Tail. Multiple works consider the worst-case group fairness [15], [24], [66]. Williamson and Menon [15] leveraged conditional value at risk (CVaR) to minimize the expected loss and the worst-case loss of any group in the upper quantile. We found that CVaR might miss discrimination in the tail and cannot reason about the shape of tail. Diana et al. [24] proposed a constrained optimization objective where the goal is to minimize the expected overall loss for all data instances subject to the hard constraints wherein no group loss can be more than a threshold. We propose an in-process bias mitigator that significantly outperforms this technique as shown in RQ4.

Intersectional Fairness. The keyword "worst-case fairness" has been also used in the relevant fairness literature [67], [68], [69]. However, their notion of fairness still relies on regular "average" fairness metrics like the rate of favorable outcomes per each subgroup. In particular, intersectional fairness concerns about the summary of fairness statistics when there are fairness measures for n subgroups. For example, Ghosh et al. [67] suggests a min-max ratio that takes the maximum for average favorable outcomes of all subgroups and divides it by the min for average favorable outcomes of all subgroups. On the other hand, our fairness measure looks at the tail of ML outcome distributions per each subgroup via EVT and compares the tail distributions between groups to quantify the amounts of discrimination.

Other Application of EVT for Fairness. In addition to its technical applications [70], [17] Extreme value theory has been significantly used to study income and wealth inequalities around the world [71], [72], [73]. Piketty and Saez [71] used the Generalized Pareto Distribution to study the distribution of income in the US between 1913 and 1998. Wang [74] studied the concept of Degree of Matthew Effect in recommendation systems via extreme value theory whereas we consider social bias (discrimination against protected groups) in decision-making systems (based on classifications).

VIII. Conclusion and Future Work

We studied fairness through the lens of extreme value theory. Our proposed approach fitted well to model the worst-cases counterfactual bias with statistical guarantees and revealed the limitations of a state-of-the-art bias reduction algorithm in the worst-case. There are multiple exciting future directions. One direction is to leverage EVT to provide a notion of AI harms to understand if automated decision-support software systematically harms a vulnerable community.

Acknowledgement. The authors thank the anonymous ICSE reviewers for their time and invaluable feedback to improve this paper. Monjezi and Tizpaz-Niari were also affiliated with UT El Paso in the completion of this work. Tizpaz-Niari and Trivedi have been partially supported by NSF under grants CCF-2317206 and CCF-2317207.

References

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[2] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] O. ChatGPT, "Chatgpt: Optimizing language models for dialogue," https://openai.com/blog/chatgpt/, 2022, online.

[5] S. M. Julia Angwin, Jeff Larson and L. Kirchne, "Machine bias," https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2021, online.

[6] D. A. Elyounes, "" computer says no!": The impact of automation on the discretionary power of public officers," *Vand. J. Ent. & Tech. L.*, vol. 23, p. 451, 2020.

[7] S. Ranchordás and L. Scarcella, "Automated government for vulnerable citizens: Intermediating rights," *SSRN Electronic Journal*, 2021.

[8] D. A. Brown, "The IRS is targeting the poorest americans," August 2021, [Online; posted 27-July-2021]. [Online]. Available: https://www.theatlantic.com/ideas/archive/2021/07/how-race-plays-tax-policing/619570/

[9] S. Tizpaz-Niari, V. Monjezi, M. Wagner, S. Darian, K. Reed, and A. Trivedi, "Metamorphic testing and debugging of tax preparation software," in *2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*. IEEE, 2023, pp. 138–149.

[10] B. Petrongolo, "The gender gap in employment and wages," *Nature Human Behaviour*, vol. 3, no. 4, pp. 316–318, 2019.

[11] D. Anderson and D. Shapiro, "Racial differences in access to high-paying jobs and the wage gap between black and white women," *ILR Review*, vol. 49, no. 2, pp. 273–286, 1996.

[12] 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*, 2012, pp. 214–226.

[13] S. Coles, J. Bawa, L. Trenner, and P. Dorazio, *An introduction to statistical modeling of extreme values*. Springer, 2001, vol. 208.

[14] S. Galhotra, Y. Brun, and A. Meliou, "Fairness testing: testing software for discrimination," in *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, ser. ESEC/FSE 2017. New York, NY, USA: Association for Computing Machinery, 2017, p. 498–510. [Online]. Available: https://doi.org/10.1145/3106237.3106277

[15] R. Williamson and A. Menon, "Fairness risk measures," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6786–6797.

[16] J. Diebolt, M. Garrido, and S. Girard, "A goodness-of-fit test for the distribution tail," 2007.

[17] J. Abella, M. Padilla, J. D. Castillo, and F. J. Cazorla, "Measurement-based worst-case execution time estimation using the coefficient of variation," *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, vol. 22, no. 4, pp. 1–29, 2017.

[18] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, *Modeling tabular data using conditional GAN*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[19] Z. Wan, Y. Zhang, and H. He, "Variational autoencoder based synthetic data generation for imbalanced learning," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1–7.

[20] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *International Conference on Machine Learning*. PMLR, 2018, pp. 60–69.

[21] J. Chakraborty, S. Majumder, and T. Menzies, "Bias in machine learning software: Why? how? what to do?" in *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2021. New York, NY, USA: Association for Computing Machinery, 2021, p. 429–440. [Online]. Available: https://doi.org/10.1145/3468264.3468537

[22] Z. Chen, J. M. Zhang, F. Sarro, and M. Harman, "Maat: a novel ensemble approach to addressing fairness and performance bugs for machine learning software," in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2022. New York, NY, USA: Association for Computing Machinery, 2022, p. 1122–1134. [Online]. Available: https://doi.org/10.1145/3540250.3549093

[23] L. Alvarez and T. Menzies, "Don't lie to me: Avoiding malicious explanations with stealth," *IEEE Software*, vol. 40, no. 3, pp. 43–53, 2023.

[24] E. Diana, W. Gill, M. Kearns, K. Kenthapadi, and A. Roth, "Minimax group fairness: Algorithms and experiments," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 66–76. [Online]. Available: https://doi.org/10.1145/3461702.3462523

[25] M. R. Leadbetter, G. Lindgren, and H. Rootzén, *Extremes and related properties of random sequences and processes*. Springer Science & Business Media, 2012.

[26] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/census+income

[27] P. Zhang, J. Wang, J. Sun, G. Dong, X. Wang, X. Wang, J. S. Dong, and T. Dai, "White-box fairness testing through adversarial sampling," in *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, 2020, pp. 949–960.

[28] H. Zheng, Z. Chen, T. Du, X. Zhang, Y. Cheng, S. Ti, J. Wang, Y. Yu, and J. Chen, "Neuronfair: Interpretable white-box fairness testing through biased neuron identification," in *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)*, 2022, pp. 1519–1531.

[29] L. Zhang, Y. Zhang, and M. Zhang, "Efficient white-box fairness testing through gradient search," in *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*, ser. ISSTA 2021, 2021, p. 103–114. [Online]. Available: https://doi.org/10.1145/3460319.3464820

[30] M. Fan, W. Wei, W. Jin, Z. Yang, and T. Liu, "Explanation-guided fairness testing through genetic algorithm," in *Proceedings of the 44th International Conference on Software Engineering*, ser. ICSE '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 871–882. [Online]. Available: https://doi.org/10.1145/3510003.3510137

[31] "UCI:heart disease data set," 2001. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[32] S. Udeshi, P. Arora, and S. Chattopadhyay, "Automated directed fairness testing," in *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*, ser. ASE '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 98–108. [Online]. Available: https://doi.org/10.1145/3238147.3238165

[33] H. Zheng, Z. Chen, T. Du, X. Zhang, Y. Cheng, S. Ji, J. Wang, Y. Yu, and J. Chen, "Neuronfair: interpretable white-box fairness testing through biased neuron identification," in *Proceedings of the 44th International Conference on Software Engineering*, ser. ICSE '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1519–1531. [Online]. Available: https://doi.org/10.1145/3510003.3510123

[34] Z. Zhao, A. Kunar, R. Birke, and L. Y. Chen, "Ctab-gan: Effective table data synthesizing," in *Proceedings of The 13th Asian Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, V. N. Balasubramanian and I. Tsang, Eds., vol. 157. PMLR, 17–19 Nov 2021, pp. 97–112. [Online]. Available: https://proceedings.mlr.press/v157/zhao21a.html

[35] E. Nazari, P. Branco, and G.-V. Jourdan, "Autogan: An automated human-out-of-the-loop approach for training generative adversarial networks," *Mathematics*, vol. 11, no. 4, 2023. [Online]. Available: https://www.mdpi.com/2227-7390/11/4/977

[36] A. Rajabi and O. O. Garibay, "Distance correlation gan: Fair tabular data generation with generative adversarial networks," in *Artificial Intelligence in HCI: 4th International Conference, AI-HCI 2023, Held as Part of the 25th HCI International Conference, HCII 2023, Copenhagen, Denmark, July 23–28, 2023, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2023, p. 431–445. [Online]. Available: https://doi.org/10.1007/978-3-031-35891-3_26

[37] X. Zhang, Y. Fu, A. Zang, L. Sigal, and G. Agam, "Learning classifiers from synthetic data using a multichannel autoencoder," *ArXiv*, vol. abs/1503.03163, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:8164829

[38] Z. Islam, M. Abdel-Aty, Q. Cai, and J. Yuan, "Crash data augmentation using variational autoencoder," *Accident Analysis and Prevention*, vol. 151, p. 105950, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000145752031770X

[39] Y. Xiao, A. Liu, T. Li, and X. Liu, "Latent imitator: Generating natural individual discriminatory instances for black-box fairness testing," in *Proceedings of the 32nd ACM SIGSOFT international symposium on software testing and analysis*, 2023, pp. 829–841.

[40] J. D. Castillo, J. Daoudi, and R. Lockhart, "Methods to distinguish between polynomial and exponential tails," *Scandinavian Journal of Statistics*, vol. 41, no. 2, pp. 382–393, 2014.

[41] R. Sokal and F. Rohlf, "Biometry: The principles and practice of statistics in biological research 3rd edition wh freeman and co," *New York*, 1995.

[42] S. Tizpaz-Niari, A. Kumar, G. Tan, and A. Trivedi, "Fairness-aware configuration of machine learning libraries," in *Proceedings of the 44th International Conference on Software Engineering*, 2022, pp. 909–920.

[43] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, "Fairlearn: A toolkit for assessing and improving fairness in AI," Microsoft, Tech. Rep. MSR-TR-2020-32, May 2020. [Online]. Available: https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[44] "UCI machine learning repository (german credit)," 2017. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[45] "UCI machine learning repository (bank marketing)," 2017. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/bank+marketing

[46] ProPublica, "Compas software ananlysis," https://github.com/propublica/compas-analysis, 2021, online.

[47] "UCI:default of credit card clients data set," 2009. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

[48] "Medical expenditure panel survey," 2014. [Online]. Available: https://meps.ahrq.gov/mepsweb/

[49] "Student performance data set," 2014. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Student+Performance

[50] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

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

[53] J. Chakraborty, S. Majumder, Z. Yu, and T. Menzies, "Fairway: a way to build fair ml software," in *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2020, pp. 654–665.

[54] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović *et al.*, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4–1, 2019.

[55] E. Gilleland, M. Ribatet, and A. G. Stephenson, "A software review for extreme value analysis," *Extremes*, vol. 16, no. 1, pp. 103–119, 2013.

[56] M. Hess and J. Kromrey, "Robust confidence intervals for effect sizes: A comparative study of cohen's d and cliff's delta under non-normality and heterogeneous variances," *Paper Presented at the Annual Meeting of the American Educational Research Association*, 01 2004.

[57] N. Mittas and L. Angelis, "Ranking and clustering software cost estimation models through a multiple comparisons algorithm," *IEEE Transactions on Software Engineering*, vol. 39, no. 4, pp. 537–551, 2013.

[58] F. Yang, Z. Yu, Y. Liang, X. Gan, K. Lin, Q. Zou, and Y. Zeng, "Grouped correlational generative adversarial networks for discrete electronic health records," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2019, pp. 906–913.

[59] L. Theis, A. van den Oord, and M. Bethge, "A note on the evaluation of generative models," *CoRR*, vol. abs/1511.01844, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:2187805

[60] V. S. Chundawat, A. K. Tarun, M. Mandal, M. Lahoti, and P. Narang, "Tabsyndex: a universal metric for robust evaluation of synthetic tabular data," *arXiv preprint arXiv:2207.05295*, 2022.

[61] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6629–6640.

[62] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Oct 2016, pp. 399–410.

[63] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[64] A. Agarwal, P. Lohia, S. Nagar, K. Dey, and D. Saha, "Automated test generation to detect individual discrimination in ai models," *arXiv preprint arXiv:1809.03260*, 2018.

[65] V. A. Dasu, A. Kumar, S. Tizpaz-Niari, and G. Tan, "Neufair: Neural network fairness repair with dropout," ser. ISSTA 2024. New York, NY, USA: Association for Computing Machinery, 2024, p. 1541–1553. [Online]. Available: https://doi.org/10.1145/3650212.3680380

[66] S. Shekhar, G. Fields, M. Ghavamzadeh, and T. Javidi, "Adaptive sampling for minimax fair classification," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 535–24 544, 2021.

[67] A. Ghosh, L. Genuit, and M. Reagan, "Characterizing intersectional group fairness with worst-case comparisons," in *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*. PMLR, 2021, pp. 22–34.

[68] M. Zhang and J. Sun, "Adaptive fairness improvement based on causality analysis," in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2022, pp. 6–17.

[69] Z. Chen, J. M. Zhang, F. Sarro, and M. Harman, "Fairness improvement with multiple protected attributes: How far are we?" in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 2024, pp. 1–13.

[70] S. Tizpaz-Niari and S. Sankaranarayanan, "Worst-case convergence time of ml algorithms via extreme value theory," in *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI*, ser. CAIN '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 211–221. [Online]. Available: https://doi.org/10.1145/3644815.3644989

[71] T. Piketty and E. Saez, "Income inequality in the united states, 1913–1998," *The Quarterly journal of economics*, vol. 118, no. 1, pp. 1–41, 2003.

[72] E. Saez, "Income and wealth concentration in a historical and international perspective, uc berkeley and nber, forthcoming in john quigley," in *Poverty, the Distribution of Income, and Public Policy, A conference in honor of Eugene Smolensky*, 2004.

[73] A. B. Atkinson, "Income Inequality in OECD Countries: Data and Explanations," *CESifo Economic Studies*, vol. 49, no. 4, pp. 479–513, 12 2003. [Online]. Available: https://doi.org/10.1093/cesifo/49.4.479

[74] H. Wang, "Fairness metrics for recommender systems," in *2022 9th international Conference on Wireless Communication and Sensor Networks (ICWCSN)*, 2022, pp. 89–92.