A GENERIC FRAMEWORK FOR CONFORMAL FAIRNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal Prediction (CP) is a popular method for uncertainty quantification with machine learning models. While the method provides probabilistic guarantees regarding the coverage of the true label, these guarantees are agnostic to the presence of sensitive attributes within the dataset. In this work, we formalize *Conformal Fairness*, a notion of fairness using conformal predictors, and provide a theoretically well-founded algorithm and associated framework to control for the gaps in coverage between different sensitive groups. Our framework leverages the exchangeability assumption (implicit to CP) rather than the typical IID assumption, allowing us to apply the notion of Conformal Fairness to data types and tasks that are not IID, such as graph data. Experiments were conducted on graph and tabular datasets to demonstrate that the algorithm can control fairness-related gaps in addition to coverage aligned with theoretical expectations.

1 INTRODUCTION

Machine learning (ML) models are increasingly used to make critical decisions in many fields of human endeavor making it essential to quantify the uncertainty associated with their predictions. Conformal Prediction (CP) is a distribution-free framework (Vovk et al., 2005) which produces confidence sets with rigorous theoretical guarantees and has become popular in real-world applications (Cherian & Bronner, 2020). Post-hoc CP allows for facile integration into ML pipelines, while its weaker requirement of a *statistical exchangeability* assumption makes it applicable to a wide variety of data types, including graph data (H. Zargarbashi et al., 2023; Huang et al., 2024).

Relatedly, ensuring the fairness of machine learning models is vital for their high-stakes deployments in critical decision-making. Biases affect ML models at different stages - from data collection to algorithmic learning stages (Mehrabi et al., 2021). During the data collection stage, measurement and representation biases can skew how each feature is interpreted, leading to inaccurate determinations by learning models. Algorithmic bias, caused by model design choices and prioritization of specific metrics while learning the model, can also lead to unfair outcomes. Many models inherit biases from historical outcomes (Kallus & Zhou, 2018; Dwork et al., 2012) and inadvertently skew decisions towards members of certain advantaged groups (Mehrabi et al., 2021). These biases have led to several global actors proposing and requiring practitioners to adhere to certain *fairness* standards (Hirsch et al., 2023). To facilitate ML pipeline and model adherence to socio-cultural or regulatory fairness standards, researchers have proposed methods to either construct fair-predictors (Alghamdi et al., 2022; Creager et al., 2019; Zhao et al., 2023) or audit fairness claims made by deployed machine learning models (Ghosh et al., 2021; Maneriker et al., 2023; Yan & Zhang, 2022).

However, these efforts on fairness (predictors, auditing, and uncertainty quantification) primarily focus on binary classification, often implicitly relying on the independent and identically distributed (IID) assumption, and do not, for the most part, bridge both fairness and uncertainty quantification. The need to both quantify uncertainty and ensure that fairness considerations are met is critical. A few researchers have started to examine how to assess (and possibly improve) the prediction quality of unreliable models (Wang & Wang, 2024) while meeting socio-cultural or regulatory standards of fairness. However, these efforts are limited in that they either require knowledge of group membership at inference time (a somewhat impractical assumption) (Lu et al., 2022) or are model specific (Wang & Wang, 2024).

Key Contributions: To redress these concerns, we propose a novel and comprehensive Conformal Fairness (CF) Framework.

First, we develop the theoretical insights that facilitate how our framework leverages CP's distribution-free approach to build and construct fair uncertainty sets according to user-specified notions of fairness. Our framework is not only comprehensive but also highly flexible, as it can be adapted to bespoke user-specified fairness criteria. This adaptability ensures that the framework can be customized to meet the specific needs of different users, enhancing its practicality and usability.

Second, the weaker (exchangeability) assumptions required by CP allow us to extend the utility of our framework to fairness problems in graph models. Graph models, in particular, suffer from the *homophilly effect*, which exacerbates inherent segregation due to node linkages and causes further biases in predictions (Dong et al., 2023).

Third, we discuss how our approach serves as a fairness auditing tool for conformal predictors. This function is important as it allows one to verify the fairness of the model, ensuring that fairness is not just a theoretical concept but a practical reality in predictive modeling.

Finally, we demonstrate the effectiveness of our CF Framework by evaluating fairness using multiple popular fairness metrics for multiple different conformal predictors on both real-world graph and tabular fairness datasets.

2 BACKGROUND

2.1 CONFORMAL PREDICTION

Conformal Prediction (Vovk et al., 2005) is a framework for quantifying the uncertainty of a model by constructing prediction sets that satisfy a *miscoverage* guarantee. For expository simplicity, we will focus on split (or inductive) conformal prediction (CP) in the classification setting. Given a calibration dataset, $\mathcal{D}_{\text{calib}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ and a test point $(\boldsymbol{x}_{n+1}, y_{n+1})$, where $\boldsymbol{x}_i \in \mathcal{X} = \mathbb{R}^d$ and $y_i \in \mathcal{Y} = \{0, \ldots, K-1\}$, CP is used to construct a prediction set $\mathcal{C}(\boldsymbol{x}_{n+1})$ such that:

$$\alpha - \frac{1}{n+1} < \Pr\big[y_{n+1} \notin \mathcal{C}_{\hat{q}(\alpha)}(\boldsymbol{x}_{n+1})\big] \leq \alpha, \tag{1}$$

where $\alpha \in [0, 1]$ is the miscoverage bound. Concretely, given a non-conformity score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, let $\hat{q}(\alpha) = \text{Quantile}\Big(\frac{\lceil (n+1)(1-\alpha) \rceil}{n}; \{s(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}\Big)$. Then, $\mathcal{C}_{\hat{q}(\alpha)}(\boldsymbol{x}_{n+1}) = \{y \in \mathcal{Y} : s(\boldsymbol{x}, y) \leq \hat{q}(\alpha)\}$ satisfies Equation 1.

Evaluating CP: *Coverage* quantifies the true test time probability $\Pr\big[y_{n+1} \in \mathcal{C}_{\hat{q}(\alpha)}(x_{n+1})\big]$ while *efficiency* is the average test prediction set size, $|\mathcal{C}(x_{n+1})|$. Intuitively, there is an inverse relationship between coverage and efficiency, as a higher desired coverage is harder to achieve so the method may produce larger prediction sets to satisfy the guarantee. In CP, the only assumption made about the data is that $\mathcal{D}_{\text{calib}} \cup \{(\boldsymbol{x}_{n+1}, y_{n+1})\}$ is *exchangeable* – a weaker notion than iid, enabling its use on non-iid data, including graph data.

Graph CP: In this work, we focus on the node classification task. Given an attributed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{X})$, where $\mathcal{V}$ is the set of nodes, $\mathcal{E}$ is the set of edges, and $\boldsymbol{X}$ is the set of node attributes. Let $\boldsymbol{A}$ be the adjacency matrix for the graph. Further, let $\mathcal{Y} = \{0, \ldots, K-1\}$ denote the set of classes associated with the nodes. For $v \in \mathcal{V}$, $\boldsymbol{x}_v \in \mathbb{R}^d$ denotes its features and $y_v \in \mathcal{Y}$ denotes its true class. The task of node classification is to learn a model that predicts the label for each node given node features and the adjacency matrix, i.e. $(\boldsymbol{X}, \boldsymbol{A}, v) \mapsto y_v$. In the transductive setting, the entire graph, including test points, is accessible during the base model training. In this scenario, for any trained permutation-equivariant function (e.g. GNN) trained on a set of training/validation nodes, the scores produced on the calibration set and test set are exchangeable, thus enabling CP to be applied (H. Zargarbashi et al., 2023; Huang et al., 2024).

2.2 FAIRNESS METRICS

Statistical/group fairness metrics aim to observe bias in the predictions of a model between different groups (defined by a sensitive attribute e.g., gender, race, ethnicity) in a dataset. This work considers several popular fairness metrics, including equal opportunity, equalized odds, demographic parity, predictive equality, and predictive parity. For generality, we define the metrics for the multiclass setting with an n-ary sensitive attribute. Let $\mathcal{Y}^+$ denote the set of advantaged labels (e.g., "is_approved" in a loan approval task), Y be the true label, and $\hat{Y}$ be the predicted label from a model. Let $\mathcal{G}$ be the set of all groups for the sensitive attribute(s). Table A1 discusses the formal definitions of different fairness metrics considered in this work.

Achieving *exact fairness* (i.e., equality in Table A1) can be challenging and, in some cases, impossible (Barocas et al., 2023). Often, regulatory requirements focus on the differences across groups for a given positive label. For example, while exact Demographic Parity is challenging to achieve, many regulatory bodies instead focus on **Disparate Impact**. Disparate Impact considers the *ratio* between the groups – rather than the difference.

3 CONFORMAL FAIRNESS (CF) FRAMEWORK

In this section, we propose a theoretically well-founded framework using conformal predictors to control for the gaps in coverage between different sensitive groups. The framework is motivated by the standard CP algorithm to determine the conditional coverage *given* a score threshold, λ. The conditional coverage for each sensitive group and each advantaged label is leveraged to evaluate if fairness is achieved for some closeness threshold c for different fairness metrics (can also be user-specified). By searching over a set Λ, we can provide an optimal threshold λ_{opt} for fairness to be achieved.

3.1 EXEMPLAR CONFORMAL FAIRNESS (CF) METRICS

We define metrics for conformal fairness by levering popular fairness metric definitions for the multiclass setting in Table A1. Essentially achieved by replacing equivalence to the prediction, $\cdot = \hat{Y}$, with membership in the prediction set, $\cdot \in \mathcal{C}_\lambda(X)$, as shown in Table 1.

Table 1: Conformal Fairness Metrics.

Metric	**Definition**
Demographic (or Statistical) Parity	$\Pr\Big[y \in \mathcal{C}_\lambda(X) \,\Big\vert\, X \in g_a\Big] = \Pr\Big[y \in \mathcal{C}_\lambda(X) \,\Big\vert\, X \in g_b\Big],\ \forall g_a, g_b \in \mathcal{G},\ \forall y \in \mathcal{Y}^+$
Equal Opportunity	$\Pr\Big[y \in \mathcal{C}_\lambda(X) \,\Big\vert\, Y = y, X \in g_a\Big] = \Pr\Big[y \in \mathcal{C}_\lambda(X) \,\Big\vert\, Y = y, X \in g_b\Big],\ \forall g_a, g_b \in \mathcal{G},\ \forall y \in \mathcal{Y}^+$
Predictive Equality	$\Pr\Big[y \in \mathcal{C}_\lambda(X) \,\Big\vert\, Y \neq y, X \in g_a\Big] = \Pr\Big[y \in \mathcal{C}_\lambda(X) \,\Big\vert\, Y \neq y, X \in g_b\Big],\ \forall g_a, g_b \in \mathcal{G},\ \forall y \in \mathcal{Y}^+$
Equalized Odds	Equal Opp. and Pred. Equality
Predictive Parity	$\Pr\Big[Y = y \,\Big\vert\, y \in \mathcal{C}_\lambda(X), X \in g_a\Big] = \Pr\Big[Y = y \,\Big\vert\, y \in \mathcal{C}_\lambda(X), X \in g_b\Big],\ \forall g_a, g_b \in \mathcal{G},\ \forall y \in \mathcal{Y}^+$

3.2 CONFORMAL FAIRNESS (CF) THEORY

Before presenting our framework, we first lay out the necessary theoretical groundwork. Detailed proofs are in Appendix B. For ease of exposition, we may equivalently control for either coverage or miscoverage.

Filtering $\mathcal{D}_{\text{calib}}$: Each fairness metric is evaluated on a subset of the population, defined by a condition on the data (i.e., membership in a group, true label value). We filter $\mathcal{D}_{\text{calib}}$ based on the corresponding fairness metric for the necessary guarantees. By doing so, we can provide probabilistic bounds to satisfy the required conditional miscoverages as stated in Lemma 3.1.

Lemma 3.1. *Calibrating on $\mathcal{D}_{\text{calib}} \cap \mathcal{R}$, where $\mathcal{R} \subset \mathcal{D}$ serves as filter, guarantees that:*

$$\alpha - \frac{1}{|\mathcal{D}_{\text{calib}} \cap \mathcal{R}| + 1} < \Pr[y_{n+1} \notin \mathcal{C}_\lambda(x_{n+1}) \mid (\boldsymbol{x}_{n+1}, y_{n+1}) \in \mathcal{D} \cap \mathcal{R}] \leq \alpha \tag{2}$$

Adhering to standard notation, it is implicit the test point is in $\mathcal{D}$ so we will omit it and write: $\Pr[y_{n+1} \notin \mathcal{C}_\lambda(\boldsymbol{x}_{n+1}) \mid (\boldsymbol{x}_{n+1}, y_{n+1}) \in \mathcal{R}]$.

Prior work (Ding et al., 2024; Vovk et al., 2005; Lei et al., 2016) focused on the upper bound; however, for our framework, the lower bound is necessary.

Inverse Quantile: In standard CP, given a miscoverage level, α, we can get a threshold by computing the $(1 - \alpha)$-quantile on the non-conformity scores of $\mathcal{D}_{\text{calib}}$. This threshold is then used

Algorithm 1 Conformal Fairness Framework

```
procedure CONFORMAL_FAIRNESS($\mathcal{D}_{\text{calib}}$, $\mathcal{Y}$, $\mathcal{Y}^+$, $\mathcal{G}$, $c$, $\Lambda$, filter_fn)
    satisfying_lambdas = [$\lambda$ for $\lambda \in \Lambda$
                    if SATISFY_LAMBDA($\mathcal{D}_{\text{calib}}, \mathcal{Y}, \mathcal{Y}^+, \mathcal{G}, c, \lambda$, filter_fn)]
    $\lambda_{\text{opt}} = \min\{\text{satisfying\_lambdas}\}$
     return $\lambda_{\text{opt}}$
end procedure

procedure SATISFY_LAMBDA($\mathcal{D}_{\text{calib}}$, $\mathcal{Y}$, $\mathcal{Y}^+$,$\mathcal{G}$, $c$,$\lambda$, filter_fn)
    label_miscoverages = $[0]_{(\mathcal{G}_i, y) \in \mathcal{G} \times \mathcal{Y}}$
    interval_widths = $[0]_{(\mathcal{G}_i, y) \in \mathcal{G} \times \mathcal{Y}}$
    for $(g, y) \in \mathcal{G} \times \mathcal{Y}^+$ do
        $\mathcal{D}_{\text{calib}(g,y)}$ = filter_fn($\mathcal{D}_{\text{calib}}, (g, y)$)
        $\mathcal{S}_{(g,y)} = \left[s(\boldsymbol{x}_i, y_i) \text{ for } (\boldsymbol{x}_i, y_i) \in \mathcal{D}_{\text{calib}(g,y)}\right]$
        interval_widths[$(g, y)$] = $\frac{1}{|\mathcal{D}_{\text{calib}(g,y)}|+1}$          ▷ Uses Lemma 3.1
        label_miscoverages[$(g, y)$] = $Q^{-1}(\lambda, \mathcal{S}_{(g,y)})$          ▷ Uses Lemma 3.2
    end for
    for $y \in \mathcal{Y}^+$ do          ▷ Uses Lemma 3.3
        $\alpha_{\text{min}}$ = min(label_miscoverages[$(\cdot, y)$] − interval_widths[$(\cdot, y)$])
        $\alpha_{\text{max}}$ = max(label_miscoverages[$(\cdot, y)$])
        if $\alpha_{\text{max}} - \alpha_{\text{min}} > c$ then return False
        end if
    end for
     return True
end procedure
```

to construct prediction sets for test points. Here, given a threshold, λ, we want to determine the corresponding miscoverage level α. We achieve this by computing an *inverse λ-quantile*. Formally, if $(\boldsymbol{x}_{n+1}, y_{n+1})$ is a test point, then the inverse λ-quantile is:

$$Q^{-1}(\lambda, S) \coloneqq \Pr[s(\boldsymbol{x}_{n+1}, y_{n+1}) \leq \lambda].$$

Lemma 3.2 asserts that the miscoverage level is within a bounded interval of length $\frac{1}{|\mathcal{D}_{\text{calib}}|+1}$.

Lemma 3.2. *For $\lambda \in [0, 1]$ and $n = |\mathcal{D}_{\text{calib}}|$,*

$$\frac{\sum_{i=1}^{n} \mathbf{1}[s(\boldsymbol{x}_i, y_i) > \lambda]}{n+1} < \Pr[y_{n+1} \notin \mathcal{C}_\lambda(x_{n+1})] < \frac{\sum_{i=1}^{n} \mathbf{1}[s(\boldsymbol{x}_i, y_i) > \lambda] + 1}{n+1}, \tag{3}$$

CF for a Fixed Label: Unlike standard CP, where miscoverage is w.r.t the correct label, y_i, the CF metrics are concerned with the miscoverage of a fixed advantaged label, $\tilde{y} \in \mathcal{Y}^+$, as seen in Table 1. Lemma 3.3 asserts that we can perform CP using a fixed label and get the same coverage guarantees.

Lemma 3.3. *Equation 1 holds if we replace $\{(\boldsymbol{x}_i, y_i)\}$ with $\{(\boldsymbol{x}_i, \tilde{y})\}$ for a fixed $\tilde{y} \in \mathcal{Y}$.*

Connecting Theory to the Framework: For a particular fairness metric, we can filter the calibration set based on the conditional from Table 1 and achieve bounds on the conditional miscoverage with Lemma 3.1. By Lemma 3.3, the bounds continue to hold when considering the conditional miscoverage for a fixed positive label. We can use Lemma 3.2 to perform an inverse quantile to compute the miscoverage under various λ thresholds. With the miscoverages for a fixed positive label and each sensitive group, we can compute the worst pairwise coverage gap across the groups using the bounds given by Lemma 3.3 to evaluate and control fairness at the desired closeness threshold.

3.3 CORE CONFORMAL FAIRNESS (CF) ALGORITHM

Input: The input to the core CF algorithm 1, include the calibration set, $\mathcal{D}_{\text{calib}}$, the set of (positive) labels, the set of sensitive groups, $\mathcal{G}$, a closeness threshold, c, a lambda threshold search space, Λ, and a filtering function. The filtering function is defined based on the conditional event for the corresponding fairness metric in Table 1. For example, for Demographic Parity and Equal Opportunity, the function would filter the calibration set where $\boldsymbol{x} \in g$ and $(\boldsymbol{x}_i \in g) \cap (y_i = y)$, respectively.

Choosing Λ: The algorithm accepts a user-provided search space, Λ, which avoids degenerate thresholds and can guarantee desirable conditions. For our experiments, we set $\Lambda = [\hat{q}, \max\{\mathcal{S}_{calib}\}]$, where $\mathcal{S}_{calib} = \{s(\boldsymbol{x}_i, y_i) : (\boldsymbol{x}_i, y_i) \in \mathcal{D}_{\text{calib}}\}$ is the set of non-conformity scores. This ensures that $\lambda_{opt} \geq \hat{q}(\alpha)$, guaranteeing the $1 - \alpha$ coverage level for the correct label. Since $\lambda_{opt} \geq \hat{q}(\alpha)$, the miscoverage decreases for larger thresholds and still satisfies the α miscoverage requirement. That is, $\alpha \geq \Pr\big[y_{n+1} \notin \mathcal{C}_{\hat{q}(\alpha)}(x_{n+1})\big] \geq \Pr\big[y_{n+1} \notin \mathcal{C}_{\lambda_{opt}}(x_{n+1})\big]$.

Procedure: For each $\lambda \in \Lambda$ and $(g, y) \in \mathcal{G} \times \mathcal{Y}^+$, $\mathcal{D}_{\text{calib}}$ is filtered with the *filter_fn* and the non-conformity scores are, $\mathcal{S}_{(g,y)}$, are computed. The inverse quantile is computed with the λ threshold on $\mathcal{S}_{(g,y)} \cup \{s(\boldsymbol{x}_{n+1}, y_{n+1})\}$. We then compare the conditional coverages for a fixed $y \in \mathcal{Y}^+$ across the different sensitive groups and check if the worst-case violation (i.e., the maximum difference in conditional coverage for a fixed label) is within our desired closeness threshold. If it isn't, then that particular λ is rejected. Of all the accepted λs, we choose the minimum to minimize the prediction set size (i.e. get the best efficiency). For fairness metrics with multiple conditions (e.g. Equalized Odds), the framework is first executed for each condition. Then, the optimal lambda is chosen from the intersection of the *satifying_lambdas*.

Using multiple λ thresholds: We also consider a classwise approach where we choose a $[\lambda^0_{opt}, \ldots, \lambda^{k-1}_{opt}] = \boldsymbol{\lambda_{opt}} \in \mathbb{R}^K$ for each of the K classes. λ^i_{opt} is only required to satisfy the closeness threshold for the i^{th} class, thus allowing for smaller λ^i_{opt} to be chosen for most classes. In theory, you can choose different lambdas for each $(g, y) \in \mathcal{G} \times \mathcal{Y}$ pair; however, in an online setting where data comes in as a stream, knowledge of the sensitive attribute may be unavailable. This setting is interesting as one may use it even if the sensitive information is explicitly removed. For example, loan applications may be race or gender-blind to enforce fairer judgment.

3.4 FRAMEWORK EXTENSIBILITY

Algorithm 1 can be directly applied to control for the coverage difference with Demographic Parity, Equal Opportunity, Predictive Equality, and Equalized Odds. The following modifications are necessary to accommodate Disparate Impact, Predictive Parity, and some user-defined variants.

Disparate Impact: The standard criterion for Disparate Impact is the 80% *Rule* (EEOC, 1979; Feldman et al., 2015). To control Disparate Impact to adhere to the Four-Fifths Rule, we change Line 18 to check if $\alpha_{\min}/\alpha_{\max} < c$, where $c = 0.8$. The rest of the algorithm stays the same.

Predictive Parity: Predictive Parity seeks to balance the Positive Predictive Value (PPV) across the different sensitive groups (Verma & Rubin, 2018). It differs from the other fairness metrics in Table 1 as it is conditioned on the prediction set. Given the objective of balancing conditional coverage, using the definition of predictive parity, and Bayes' theorem, we get

$$\Pr[Y = y \mid y \in \mathcal{C}_\lambda(X), X \in g_a] = \underbrace{\frac{\Pr[y \in \mathcal{C}_\lambda(X) \mid Y = y, X \in g_a]}{\Pr[y \in \mathcal{C}_\lambda(X) \mid X \in g_a]}}_{\text{Equal Opportunity over Demographic Parity}} \cdot \underbrace{\Pr[Y = y \mid X \in g_a]}_{\text{Conditional Label Probability}},$$

for $y \in \mathcal{Y}^+$ and $g_a \in \mathcal{G}$. A solution is guaranteed for any choice of $\mathcal{Y}^+ \subseteq \mathcal{Y}$ if c is greater than the maximum pairwise total variation distance of the group-conditioned label distribution. Formally,

Theorem 3.4. *Let W be a random variable for a label distribution over $\mathcal{Y}$. Let $W_i \sim W|(X \in g_i)$ – the label distribution conditioned on group membership. Then there exists λ such that for $c \geq \max\{D_{TV}(W_i, W_j) \mid i, j \in \{1, \ldots, |\mathcal{G}|\}\}$, where D_{TV} is the total variation distance, the coverage difference for Predictive Parity is satisfied.*

One can define a *modified total variation distance* as

$$D^+_{TV}(W_i, W_j) \coloneqq \sup_{k \in \mathcal{Y}^+} |\Pr[W_i = k] - \Pr[W_j = k]|,$$

and use this in place of D_{TV} in Theorem 3.4 for a weaker assumption about c, which still gives a satisfying λ.

Since Equal Opportunity, Demographic Parity, and Conditional Label Probability are all bounded within intervals, we can compute an interval for which Predictive Parity is satisfied and then use our framework to find λs where the coverages satisfy the coverage difference requirement. More theoretical details about the interval guarantees and a proof of Theorem 3.4 are in Appendix C.

To control for arbitrarily small values of c, we can either assume that the label distribution is independent of group membership or use the following *Predictive Parity Proxy* (an example of a user-defined metric). For all $g_a, g_b \in \mathcal{G}, y \in \mathcal{Y}^+$

$$\big|(\Pr[Y = y \mid y \in \mathcal{C}_\lambda(X), X \in g_a] - \Pr[Y = y \mid X \in g_a]) \\ - (\Pr[Y = y \mid y \in \mathcal{C}_\lambda(X), X \in g_b] - \Pr[Y = y \mid X \in g_b])\big| < c \quad (4)$$

Proofs and technical details on these modifications can be found in Appendix C.

3.5 LEVERAGING THE CF FRAMEWORK FOR FAIRNESS AUDITING

Using the Conformal Fairness Framework, one can audit if the disparity of a conformal predictor between multiple groups violates a user-specified fairness criterion. Specifically, we have thus far focused on fairness criteria concerning bounding the disparity between groups using the fairness metrics described in Table 1 by some closeness threshold, c. It is straightforward to support user-defined fairness metrics concerning label coverage. While Algorithm 1, as presented, gives a method of finding an optimal λ threshold which satisfies the fairness guarantees using Lemmas 3.1, 3.2, and 3.3, the same `satisfy_lambda` procedure can be leveraged to check if a *given* λ used by a conformal predictor satisfies the same fairness guarantees. Notably, the CF framework can also be leveraged even if the conformal predictor is treated as a black-box model. In this case, we can construct a $\mathcal{D}_{\text{audit}}$ set exchangeable with the calibration data used for the conformal predictor. Using $\mathcal{D}_{\text{audit}}$, we can determine if the conformal predictor satisfies the corresponding fairness guarantee given the fairness metric and the λ threshold used.

3.6 NON-CONFORMITY SCORES

There are several choices for the non-conformity score for performing fair conformal prediction with classification tasks. We currently implement TPS (Sadinle et al., 2019), APS (Romano et al., 2019), DAPS (H. Zargarbashi et al., 2023), and CFGNN (Huang et al., 2024) in the CF framework, though any non-conformity score can be used. More details on the specifics of each non-conformity score can be found in Appendix D.2.

4 EXPERIMENTS

4.1 SETUP

Datasets: To evaluate the CF Framework, we used five multi-class datasets Pokec-n (Takac & Zabovsky, 2012), Pokec-z (Takac & Zabovsky, 2012), Credit (Agarwal et al., 2021), ACSIncome (Ding et al., 2021), and ACSEducation (Ding et al., 2021) (see Table 2 for details). For each dataset, we use a $30\%/20\%/25\%/25\%$ stratified split of the labeled points for $\mathcal{D}_{\text{train}}/\mathcal{D}_{\text{valid}}/\mathcal{D}_{\text{calib}}/\mathcal{D}_{\text{test}}$.

Table 2: Dataset Statistics. T refers to Tabular and G refers to Graph.

Name	Type	Size	# Labeled	# Groups	# Classes
ACSIncome	T	$1,664,500$	ALL	race(9)	4
ACSEducation	T	$1,664,500$	ALL	race(9)	6
Name	**Type**	$(\lvert\mathcal{V}\rvert, \lvert\mathcal{E}\rvert)$	**# Labeled**	**# Groups**	**# Classes**
Credit	T/G	$(30,000,\ 1,436,858)$	ALL	age(2)	4
Pokec-n	G	$(66,569,\ 729,129)$	$8,797$	region(2), gender(2)	4
Pokec-z	G	$(66,569,\ 729,129)$	$8,797$	region(2), gender(2)	4

Models: For the graph datasets, we evaluated with GCN (Kipf & Welling, 2017), GraphSAGE (Hamilton et al., 2017), or GAT (Veličković et al., 2018) as the base model (results reported are for the highest performing base model). For Credit, we evaluated additionally considered XGBoost (Chen & Guestrin, 2016) (i.e., ignoring the graph structure) as we empirically observed this approach to outperform the graph neural network baselines in terms of efficiency for this dataset. The choice of ignoring edge information while training Credit on XGBoost does not prohibit us from using CFGNN or DAPS - which utilize the edge information. The conformal predictor simply requires the softmax logits from the base model (i.e. XGBoost) but is otherwise model agnostic. For ACSIncome and ACSEducation, we used an XGBoost model. Each model's hyperparameters were tuned as discussed in Appendix D.3.

Baseline: For each dataset and CP non-conformity score, we built a conformal predictor. Then, we assess fairness according to the specific fairness metric using the conformal quantile, $\hat{q}$, using the 90% quantile ($\alpha = 0.1$) from the calibration phase.

Evaluation Metrics: We report the *worst fairness disparity* and *efficiency*. For Disparate Impact, the worst fairness disparity is the *minimum* $\alpha_{\text{min}}/\alpha_{\text{max}}$ across the positive labels. For the remaining metrics, we record the *maximum* $\alpha_{\text{max}} - \alpha_{\text{min}}$ across the positive labels.

4.2 RESULTS

For each figure, we use a black line to indicate the base conformal predictor's *average worst-case* fairness disparity across different thresholds, the bar plot for the *worst* fairness disparity using the CF Framework, and a black dot to denote the desired fairness disparity. We report the average base performance for simplicity and readability of the figures. In every experiment, except for Figure 2, the CF framework was better than the average base conformal predictor. We provide a more granular version of Figure 2 in Figure E4, where it is clear that the framework performs better for every closeness threshold.

Controlling for Fairness Disparity: For different closeness thresholds, our CF Framework effectively controls the fairness disparity for several metrics compared to the base conformal predictor. In Figure 1 and 2, we can observe that in terms of fairness disparity, our CF Framework **precisely** (note step-wise change with c on violations) improves upon the baseline conformal predictor. As with algorithmic fairness, a trade-off is involved in that there is a slightly worse efficiency. From Figure 2, we continue to observe this for both standard and graph-based conformal predictors. Furthermore, if the base conformal predictor is already "fair" according to our fairness disparity criterion, then the CF Framework will report the results accordingly. This phenomenon is observed with the CFGNN results in Figure 2, where the CF Framework matches the baseline regarding both evaluation metrics. This behavior of the CF Framework makes it suitable to leverage for black box fairness auditing (as noted previously). We present additional results, for example, the disparity results for the CF Framework without classwise lambdas in Appendix E. Notably, the prediction set sizes are more prominent due to selecting a larger λ than the classwise approach (see Figure E3 vs Figure 1).

Controlling for Disparate Impact: For Disparate Impact, we present results for the standard *80% Rule*. In Table 3, we see that using the CF Framework can significantly improve upon the base conformal predictor for the *80% Rule*. For the base conformal predictor, the disparate impact value is far below the desired 0.8, and in some cases less than 0.4 as with Credit with TPS and ACSIncome dataset. Our framework, however, is close to the 0.8 value and in some cases surpasses it, like in Credit with CFGNN, with minor effects on the efficiency for both datasets.

Table 3: *80% Rule* for Credit and ACSIncome. Our framework surpasses the base conformal predictor achieving a disparate impact value of 0.80 or higher

		APS		**TPS**		**CFGNN**		**DAPS**	
		Base	CF	Base	CF	Base	CF	Base	CF
Credit	Disp. Impact	0.646	**0.821**	0.252	**0.793**	**0.922**	**0.922**	0.539	**0.809**
	Efficiency	2.326	2.513	2.268	2.558	2.202	2.202	2.254	2.526
ACSIncome	Disp. Impact	0.397	**0.797**	0.356	**0.798**	N/A	N/A	N/A	N/A
	Efficiency	2.212	2.674	2.109	2.679	N/A	N/A	N/A	N/A

Agnostic to Non-Conformity Score: As discussed earlier, the CF Framework can support a variety of non-conformity scores, emphasizing the agnostic nature of our framework. We achieved effective results for conformal predictors with different underlying non-conformity score functions for all the experiments. Further results can be found in Appendix E.

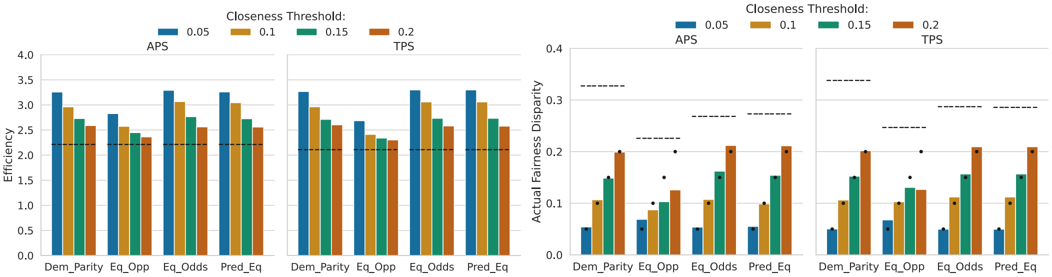

Figure 1: **ACSIncome**. The left two plots are efficiency results, while the right two are the fairness disparities for (a) APS and (b) TPS. In all cases, our framework gives results at or better than the desired threshold and better than the baseline.

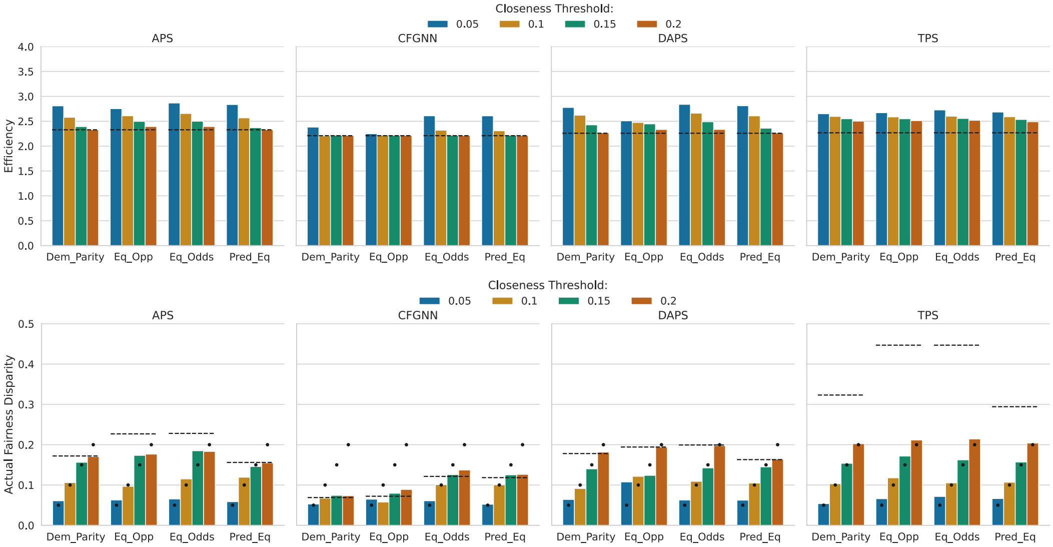

Figure 2: **Credit**. The top four plots are efficiency results, while the bottom four are the fairness disparities for (a) APS, (b) CFGNN, (c) DAPS, and (d) TPS. In all cases, our framework achieves the desired coverage gap better than the baseline, with a minor impact on efficiency.

Intersectional Fairness: When characterizing data points into groups, we are not limited to a single sensitive attribute. In many applications, there can be multiple sensitive attributes (e.g., race and gender) that need to be considered. Our CF Framework is not limited to analyzing a single sensitive attribute. To demonstrate this, we conduct an experiment with the Pokec-n dataset. Pokec-n has two sensitive attributes, namely *region* and *gender*. We treat each combination of region and gender as a separate sensitive group and apply the CF framework to control for fairness disparities. Figure 3 shows that the CF framework improves upon the base conformal predictor regarding fairness disparity. This improvement is starker with the graph-based conformal predictors, CFGNN, and DAPS as seen in Figure 3 plots (b) and (c).

One challenge intersectional fairness introduces is the multiplicative increase in the number of groups that must be calibrated and evaluated (combinations of sensitive attributes and classes). This places a stronger requirement on the number of data points necessary to meet the coverage guarantees we discussed in Section 3.2 (guarantees are more challenging to meet as the size of $\mathcal{D}_{(g,y)}$ gets smaller). This problem is exacerbated (in empirical results) for datasets with only a few labeled points such as Pokec-n. For Pokec-n, using a standard data split, the calibration set has around 2200 data points. The calibration set is then further split to get the conditional positive label coverage for each positive label and group pair. This results in the calibration being done with sets of fewer than a few hundred points, which is much lower than the suggested 1000 points in the literature (Angelopoulos & Bates, 2021). In Figure 3, the effect of this challenge is seen with the fairness disparity given by the CF Framework being slightly above the desired closeness threshold for $c = 0.1$. However, despite this disadvantage for many metrics, the guarantees are still being met, even for intersectional fairness.

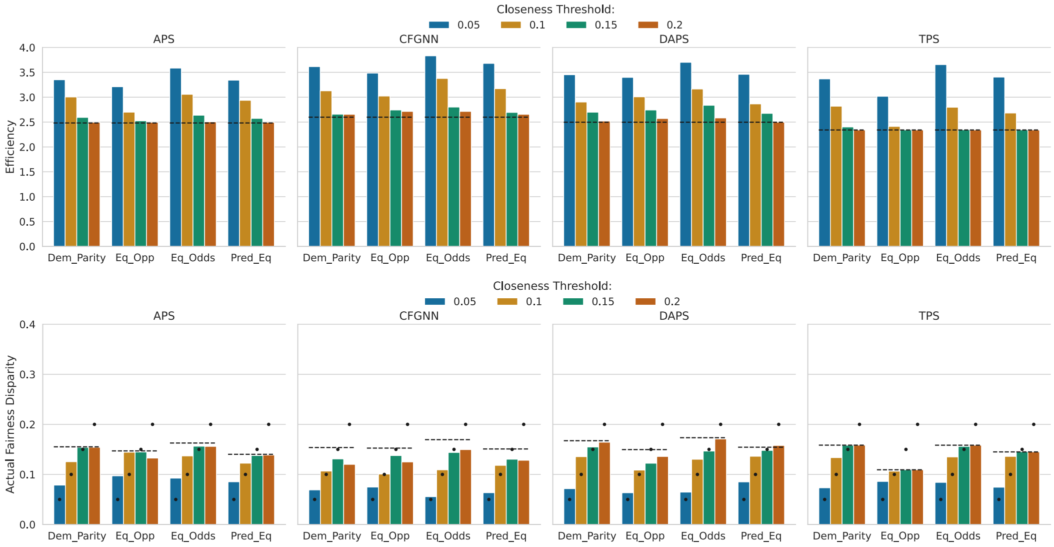

Figure 3: **Pokec-n** using **both** sensitive attributes. The top four plots are the efficiency results, while the bottom four are the fairness disparities for (a) APS, (b) CFGNN, (c) DAPS, and (d) TPS. We observe that CFGNN (b) and DAPS (c) achieve the desired fairness coverage thresholds better than standard CP methods.

Predictive Parity Proxy: As discussed, the CF framework is extensible to user-defined fairness notions. We consider the Predictive Parity Proxy in Equation 4 as an example of a user's ability to provide a reasonable fairness measure (Disparate Impact, above is another example). An experiment on ACSEducation in Table 4 demonstrates we can control for arbitrarily small values of c, unlike the standard notion of Predictive Parity. Additionally, it empirically illustrates that we can control for disparities of probabilities conditioned on the prediction set. This metric can also be applied in the graph setting, as seen in Appendix E.

Table 4: **ACSEducation**. The worst-case fairness disparity, based on the Predictive Parity Proxy, with our method is below the desired c threshold, while the *avearge* baseline disparity is much higher (> 0.30) than all of the c thresholds we consider.

	Closeness Threshold (c)	**0.05**	**0.10**	**0.15**	**0.20**	**Base (Average)**
TPS	Max Fairness Disparity	0.038	0.091	0.167	0.199	0.319
	Efficiency	4.551	3.761	3.348	3.210	2.828
APS	Max Fairness Disparity	0.044	0.093	0.152	0.166	0.411
	Efficiency	5.185	3.975	3.499	3.374	2.982

4.3 DISCUSSION

Very few prior efforts study fairness and conformal prediction (Wang et al., 2024; Lu et al., 2022; Liu et al., 2022). One line of work has focused on applying fairness notions toward CP problems for regression tasks, explicitly focusing on Demographic Parity (Liu et al., 2022) and Equal Opportunity (Wang et al., 2024), respectively. Our work differs in its breadth and flexibility (supporting a range of fairness metrics and conformity scores) and its focus on classification. While our work does not address the regression tasks directly, this is a potential direction for future work. Another line of work focuses on applying the notion of Overall Accuracy Equality for CP (Lu et al., 2022). This effort considers a specific medical application of detecting malignant skin conditions and applies group-balanced CP (Vovk, 2012). Our CF framework generalizes group-balanced CP to consider the notion of coverage for a particular label, thus allowing us to evaluate disparity based on classical fairness metrics in a manner that does not require knowledge of group membership at prediction time (or in an online setting), unlike in Lu et al. (2022) which relies on having the group membership information a priori.

5 CONCLUSION

In this work, we formalize the notion of Conformal Fairness using conformal predictors and propose a novel and comprehensive Conformal Fairness (CF) Framework. We provide a theoretically grounded algorithm that can be used to control for the gaps in conditional coverage, defined based on different fairness metrics, across sensitive groups. We conduct experiments with conformal predictors for both tabular and graph datasets, leveraging the exchangeability assumption of (graph) conformal prediction. We present results for Conformal Fairness based on various classical and user-defined fairness metrics on conformal predictors with various non-conformity score functions. We further present results on the framework's effectiveness in evaluating intersectional fairness with conformal predictors. We further describe how the CF framework can be practically leveraged for applications, including fairness auditing of conformal predictors. Future work could include expanding the CF framework to control for coverage gaps for regression tasks and enhancing the theory to loosen assumptions of conformal prediction and look at non-exchangeable variations.

REFERENCES

Chirag Agarwal, Himabindu Lakkaraju, and Marinka Zitnik. Towards a unified framework for fair and stable graph representation learning. In *Uncertainty in Artificial Intelligence*, pp. 2114–2124. PMLR, 2021.

Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, Peter Michalak, Shahab Asoodeh, and Flavio Calmon. Beyond adult and compas: Fair multi-class prediction via information projection. *Advances in Neural Information Processing Systems*, 35:38747–38760, 2022.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *CoRR*, abs/2107.07511, 2021. URL `https://arxiv.org/abs/2107.07511`.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL `https://doi.org/10.1145/2939672.2939785`.

John Cherian and Lenny Bronner. How the washington post estimates outstanding votes for the 2020 presidential election. *Retrieved September*, 13:2023, 2020.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pp. 1436–1445. PMLR, 2019.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

Tiffany Ding, Anastasios N. Angelopoulos, Stephen Bates, Michael I. Jordan, and Ryan J. Tibshirani. Class-conditional conformai prediction with many classes. In *Proceedings of the 37th*

International Conference on Neural Information Processing Systems, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

Yushun Dong, Jing Ma, Song Wang, Chen Chen, and Jundong Li. Fairness in graph mining: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10583–10602, 2023.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

The U.S. EEOC. Uniform guidelines on employee selection procedures. March 1979.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pp. 259–268, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336642. doi: 10.1145/2783258.2783311. URL `https://doi.org/10.1145/2783258.2783311`.

Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S. Meel. Justicia: A stochastic sat approach to formally verify fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35 (9):7554–7563, May 2021. doi: 10.1609/aaai.v35i9.16925. URL `https://ojs.aaai.org/index.php/AAAI/article/view/16925`.

Soroush H. Zargarbashi, Simone Antonelli, and Aleksandar Bojchevski. Conformal prediction sets for graph neural networks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12292–12318. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/h-zargarbashi23a.html`.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. *CoRR*, abs/1706.02216, 2017. URL `http://arxiv.org/abs/1706.02216`.

Dennis Hirsch, Timothy Bartley, Aravind Chandrasekaran, Davon Norris, Srinivasan Parthasarathy, and Piers Norris Turner. *Business Data Ethics: Emerging Models for Governing AI and Advanced Analytics*. Springer Nature, 2023.

Kexin Huang, Ying Jin, Emmanuel Candès, and Jure Leskovec. Uncertainty quantification over graph with conformalized graph neural networks. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning*, pp. 2439–2448. PMLR, 2018.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=SJU4ayYgl`.

Jing Lei, Max Grazier G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry A. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113:1094 – 1111, 2016. URL `https://api.semanticscholar.org/CorpusID:13741419`.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.

Meichen Liu, Lei Ding, Dengdeng Yu, Wulong Liu, Linglong Kong, and Bei Jiang. Conformalized fairness via quantile regression. *Advances in Neural Information Processing Systems*, 35:11561–11572, 2022.

Tianci Liu, Haoyu Wang, Yaqing Wang, Xiaoqian Wang, Lu Su, and Jing Gao. Simfair: A unified framework for fairness-aware multi-label classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14338–14346, Jun. 2023. doi: 10.1609/aaai.v37i12.26677. URL https://ojs.aaai.org/index.php/AAAI/article/view/26677.

Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12008–12016, Jun. 2022. doi: 10.1609/aaai.v36i11.21459. URL https://ojs.aaai.org/index.php/AAAI/article/view/21459.

Pranav Maneriker, Codi Burley, and Srinivasan Parthasarathy. Online fairness auditing through iterative refinement. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 1665–1676, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599454. URL https://doi.org/10.1145/3580305.3599454.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Julien Rouzot, Julien Ferry, and Marie-José Huguet. Learning optimal fair scoring systems for multi-class classification. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 197–204, 2022. doi: 10.1109/ICTAI56018.2022.00036.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Lubos Takac and Michal Zabovsky. Data analysis in public social networks. In *International scientific conference and international workshop present day trends of innovations*, volume 1, 2012.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018. URL https://arxiv.org/abs/1710.10903.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pp. 1–7, 2018.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In Steven C. H. Hoi and Wray Buntine (eds.), *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pp. 475–490, Singapore Management University, Singapore, 04–06 Nov 2012. PMLR. URL https://proceedings.mlr.press/v25/vovk12.html.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Fangxin Wang, Lu Cheng, Ruocheng Guo, Kay Liu, and Philip S. Yu. Equal opportunity of coverage in fair regression. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

Ziyan Wang and Hao Wang. Variational imbalanced regression: Fair uncertainty quantification via probabilistic smoothing. *Advances in Neural Information Processing Systems*, 36, 2024.

Tom Yan and Chicheng Zhang. Active fairness auditing. In *International Conference on Machine Learning*, pp. 24929–24962. PMLR, 2022.

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2): 2473–2480, 2009.

Chen Zhao, Le Wu, Pengyang Shao, Kun Zhang, Richang Hong, and Meng Wang. Fair representation learning for recommendation: A mutual information perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):4911–4919, Jun. 2023. doi: 10.1609/aaai.v37i4.25617. URL https://ojs.aaai.org/index.php/AAAI/article/view/25617.