TRAINING FAIR TABULAR FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Tabular Foundation Models (TFMs) have emerged as leading methods for tabular predictive tasks, leveraging in-context learning to predict on new data without task-specific training. Despite the increased use of TFMs in high-stakes decision-making, their fairness properties remain largely unexplored. In this work, we incorporate fairness constraints directly into TFM training, enabling fair predictions in a single forward pass. Our approach addresses two key challenges: limited access to sensitive attributes in training data, and the incompatibility of existing fairness techniques with the in-context learning paradigm. We propose `FairTFM`, a scalable training strategy based on synthetic fairness tasks and a fairness-aware architecture using a gradient reversal layer, which encourages the model to learn representations invariant to sensitive attributes. Experiments on 132 fairness tasks show consistent improvements in fairness while maintaining competitive accuracy.

1 INTRODUCTION

Tabular data, organized in rows and columns, is the dominant modality for decision-making tasks in domains such as healthcare and finance. While tree-based models, such as XGBoost (Chen & Guestrin, 2016), have long dominated tabular learning, tabular foundation models (TFMs) have emerged as strong alternatives (Hollmann et al., 2023). These models are typically pretrained on large collections of tabular data and can adapt to new tasks via in-context learning (ICL) (Brown et al., 2020), using only a few labeled examples. Unlike traditional approaches, TFMs do not require task-specific training or hyperparameter tuning. Notably, models such as TabDPT (Ma et al., 2025), TabPFNv2 (Hollmann et al., 2025), and TabICLv2 (Qu et al., 2026) can match or outperform heavily tuned tree-based models across a wide range of datasets (Erickson et al., 2025).

Even as TFMs are deployed in high-stakes decision-making, their fairness properties remain largely unexplored. Recent work shows that, despite strong predictive performance, TFMs can exhibit biased outcomes similar to traditional models (Kenfack et al., 2026). While fairness-aware training has been extensively studied in classical machine learning (Mehrabi et al., 2021), these approaches do not readily extend to the ICL paradigm, where predictions must be produced in a single forward pass without task-specific optimization. This gap calls for fairness-aware training methods specifically designed for TFMs. Hence, in this work, we propose a training framework that incorporates fairness as an objective in TFMs alongside predictive performance. We train TFMs under explicit fairness constraints, enabling fair predictions directly through ICL without requiring post-hoc correction or task-specific retraining. To our knowledge, the resulting model is the first TFM that accounts for statistical group fairness notions, such as demographic parity and equalized odds, during pretraining.

As a running example, consider a bank that uses a TFM to predict whether a client's income will exceed $50,000. Anti-discrimination regulations forbid disadvantaging clients based on gender, which could occur downstream if the income predictions are biased. Since an off-the-shelf TFM makes predictions without modification to its weights, there is no opportunity to apply fairness constraints specific to this task. With our approach, the bank simply marks gender as the sensitive attribute in the context set, and the pretrained model produces fairer predictions in a single forward pass.

Our method for fair pretraining relies on two key components. (i) **Synthetic fairness task generation:** given a dataset, we randomly designate an input feature as a sensitive attribute, treat it as categorical, and optimize for fairness with respect to it. Repeating this process for every dataset sampled during pretraining enables scalable fairness-aware training across diverse tasks. (ii) **Fairness-aware architecture and training:** we extend transformer-based TFMs with a dedicated encoder for sensitive attributes and introduce a dual-head prediction mechanism. In addition to the label prediction head,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

we include a sensitive attribute predictor connected through a gradient reversal layer (Ganin et al., 2016), encouraging the model to learn representations invariant to sensitive attributes.

We evaluate our approach on 120 fairness tasks derived from the ACS PUMS datasets (Ding et al., 2021) with varying sensitive attributes, and on 12 additional tasks built from six widely used fairness benchmarks to test generalization. We measure fairness using demographic parity, equal opportunity, and equalized odds, comparing against strong baselines. Our results show that the proposed `FairTFM` framework consistently improves fairness metrics while maintaining competitive predictive performance, improving fairness by 32–75% relative to the strongest TFM baseline at accuracy costs of 2–11%.

Our contributions are as follows:

- We propose a scalable strategy for creating fairness tasks with a wide variety of sensitive attribute relationships, based on a stick-breaking group construction that requires no labelled sensitive attributes.
- We design a fairness-aware transformer architecture using adversarial learning that enables fair predictions in a single forward pass, handling a sensitive attribute that changes identity across tasks and is masked at inference.
- We provide an extensive empirical evaluation on a diverse set of tasks, demonstrating improved fairness metrics while preserving accuracy, including comparisons with fairness-aware baselines and probing analyses of the learned representations.

2 Related work

Fairness Prior work on fair machine learning typically intervenes at one of three stages: pre-processing the data, incorporating fairness constraints during training, or post-processing model outputs (Kamiran & Calders, 2009; Zemel et al., 2013; Hardt et al., 2016; Zhang et al., 2018). These approaches have been studied extensively for conventional supervised models, but they often assume access to the training pipeline or to calibrated model outputs. This assumption is less compatible with ICL, where a pretrained model is used as a frozen predictor at inference time. Our setting is therefore closer to learning fairness-aware representations during pretraining so that fair behaviour can be obtained in a single forward pass. While adversarial fairness objectives based on gradient reversal have previously been studied in unsupervised domain adaptation (Ganin & Lempitsky, 2015), integrating such objectives into TFM pretraining introduces fundamentally different challenges, including heterogeneous tasks, dynamically changing sensitive attributes across sampled tasks, and the requirement of single-pass in-context inference without task-specific optimization. Our contribution is therefore not merely the use of adversarial learning for fairness, but the introduction of a scalable fairness-aware pretraining framework for TFMs that enables fairer predictions across unseen downstream tasks while preserving the inference properties of in-context learning.

Tabular foundation models Recent tabular foundation models such as TabPFN, TabDPT, and TabICL demonstrate that pretrained transformers can be highly competitive on tabular prediction tasks (Hollmann et al., 2023; 2025; Ma et al., 2025; Grinsztajn et al., 2026; Qu et al., 2025; 2026). However, this literature emphasizes predictive accuracy, and does not directly address the potential for biased outcomes. FairPFN (Robertson et al., 2025) does incorporate fairness into TFM pretraining under a causal notion of fairness, which contrasts with our work which targets statistical group fairness notions such as demographic parity, equal opportunity, and equalized odds. Statistical fairness is data-driven and focuses on distributional fairness (outcome-based), while causal fairness is more interventional and does not necessarily ensure outcome parity across groups. Our work focuses directly on optimizing statistical fairness notions commonly used in tabular classification benchmarks (Le Quy et al., 2022). We include FairPFN as a baseline in our experiments and find that its causal objective does yield partial group-fairness benefits, but at a predictive cost that leaves it Pareto-dominated by `FairTFM`, and without any mechanism for controlling the fairness–utility trade-off. We provide a more detailed discussion of these connections in Appendix A.

3 Training fair tabular foundation models

In this section, we present our fairness-aware tabular foundation model for statistical fairness. We describe how fairness tasks are synthesized from generic data priors, introduce the model architecture, and specify the pretraining objective.



(a) Fairness task sampling from a prior.

(b) Overview of the FairTFM architecture.

Figure 1: (a) We construct fairness tasks from a prior data generator by first sampling a dataset (X', y), then randomly removing one feature to serve as the sensitive attribute s and converting it to a categorical variable, yielding the triplet (X, y, s). The resulting data are split into context $(X_{\text{ctx}}, y_{\text{ctx}}, s_{\text{ctx}})$ and query $(X_{\text{qy}}, y_{\text{qy}}, s_{\text{qy}})$ sets for ICL. (b) FairTFM jointly processes context and query inputs with a transformer. In addition to the main prediction head for the target label $\hat{y}_{\text{qy}}$, a second head predicts the sensitive attribute $\hat{s}_{\text{qy}}$. A gradient reversal layer on the sensitive branch (red arrow) encourages the model to learn representations that are invariant to s, thereby reducing reliance on sensitive-correlated features while preserving predictive performance.

3.1 Fairness task sampling from a dataset prior

Most TFMs are pretrained on tasks sampled from a dataset prior, whether synthetic or representing real datasets, in order to expose the model to a broad distribution of supervised learning problems. We follow the same principle, but augment it to generate *fairness tasks* in a self-supervised manner (Chen et al., 2020; Balestriero et al., 2023) that captures diverse forms of group-dependent bias. A similar process was employed in TabDPT (Ma et al., 2025) for crafting diverse predictive tasks from real-world datasets, and in CausalPFN (Balazadeh et al., 2025; Stith et al., 2026) for generating synthetic causal inference tasks.

Concretely, starting from a sampled dataset, we randomly designate one input feature as the sensitive attribute and remove it from the feature set, as illustrated in Figure 1a. When the selected feature is continuous, we convert it into a categorical attribute using a "stick-breaking" discretization scheme. We first sample a sequence of proportions from Beta distributions and use them to construct mixture weights via a Dirichlet process (Khan et al., 2012), where each component represents a fraction of the remaining mass. These weights are then mapped to empirical quantiles of the normalized feature, producing data-dependent cut points that partition the feature into discrete groups. This procedure yields flexible, non-uniform bins that adapt to the underlying feature distribution while retaining a probabilistic interpretation.

In the running example, the bank's dataset corresponds to one predictive task requiring fairness considerations, where gender plays the role of the designated sensitive attribute. During pretraining, any feature in a dataset can play the role of sensitive attribute in some sampled tasks, and an ordinary predictive feature in others.

3.2 Network architecture

We build on TabPFN's transformer encoder for tabular data (Vaswani et al., 2017), using the nanoTabPFN architecture (Pfefferle et al., 2025) as a lightweight backbone that also has less input pre-processing so that we mitigate confounding effects in our controlled experiments. In this architecture, each input pair (X_i, y_i) is represented as a sequence of d-dimensional tokens and is processed by alternating self-attention over rows and columns, enabling ICL along both axes.

During pretraining, available input data is divided into context $(X_{\text{ctx}}, y_{\text{ctx}})$ and query $(X_{\text{qy}}, y_{\text{qy}})$ sets. As in TabPFN, the query labels y_{qy} are masked and the corresponding output token is mapped to class logits through a multi-layer perceptron (MLP). For inference, available labelled data (what would be training data for a non-foundational predictive model) is provided as context, and the TFM predicts y_{qy} on unlabelled query data.

Our fairness-aware extension, which we call Fair Tabular Foundation Model (FairTFM), is shown in Figure 1b. Rather than operating only on pairs (x_i, y_i), FairTFM processes triplets (x_i, y_i, s_i), where s_i denotes the d-dimensional token of the sensitive attribute. FairTFM incorporates a shared transformer backbone that learns representations, and two predictive MLP heads for the target label $\hat{y}_{\text{qy}}$ and sensitive attribute $\hat{s}_{\text{qy}}$. Because the chosen attention mechanism uses no column positional

encodings, predictions are invariant to the ordering of input columns (Pfefferle et al., 2025). Hence, when a feature is randomly designated as the sensitive attribute during pretraining, its index in the table does not have an impact on predictions.

To discourage the learned representation from encoding sensitive information, the secondary prediction head that predicts $\hat{s}_{\text{qy}}$ is preceded by a gradient reversal layer (GRL) (Ganin & Lempitsky, 2015). The head itself is trained to minimize prediction error of s_{qy}, and would naturally exploit any information about s encoded in the transformer's representation. However, the GRL multiplies the gradient signal passing to the transformer by -1, such that the representation is adversarially trained to contain no information about s, making the prediction head's task harder. Intuitively, the GRL forces the transformer to reduce reliance on rows and features that are informative for s, including proxies for s, leading to fairer predictions when the same representation is used by the y head.

A key advantage of the GRL formulation in our setting is that it does not require training a separately optimized adversary for each sampled fairness task, even though the designated sensitive attribute changes across tasks during pretraining. In contrast, an adversarial setup with separately trained classifier adversaries would require fitting or re-optimizing a new adversary for each sampled fairness task during pretraining. Such per-task adversarial training would substantially increase pretraining complexity and make large-scale task sampling less practical.

Since the fairness tasks used for pretraining sample the sensitive attribute at random from the available features (Section 3.1), one may worry that the GRL degrades the embedding broadly rather than removing only sensitive information. The key design point is that the sensitive attribute is removed from X' and provided through a dedicated encoder as an explicit token for every context row (Algorithm 1, lines 3 and 8–9). The s-head only predicts s, so the reversed gradient only penalizes information about the attribute signalled by the in-context s tokens; the learned invariance to s is conditional and in-context. Moreover, since each feature of X' is treated as sensitive in some sampled tasks and not in others, the overall objective is not optimized by degrading any fixed feature. In the running example, the query representations suppress gender *and its proxies* because the bank designates gender as the sensitive attribute in the context examples.

At inference time, we do not assume access to the sensitive attribute s_{qy}. Since the transformer is trained to create representations that do not contain information about s_{qy}, providing it would not be particularly fruitful. Hence, we replace s_{qy} values with a fixed learnable parameter during both pretraining and inference. This allows the transformer encoder to operate on d-dimensional tokens representing (x_i, y_i, s_i) and produce query sample embeddings, while ensuring that s_{qy} is never directly exposed to the model. Beyond obviating the need for direct access to sensitive information at inference, this design discourages the model from exploiting s_{qy} as a predictive shortcut, which would undermine the objective of the GRL.

3.3 Training objective

`FairTFM` is trained from scratch on purely synthetic fairness tasks derived from TabICL's prior (Qu et al., 2025) (Appendix D). The transformer, shown in Figure 1b, produces an embedding that captures row- and feature-wise interactions among the context instances, and the two MLP heads output logits for the query label ($\hat{y}_{\text{qy}}$) and corresponding sensitive attribute ($\hat{s}_{\text{qy}}$). We optimize the model using the following joint cross-entropy objective:

$$\mathcal{L} = \text{CE}(\hat{y}_{\text{qy}}, y_{\text{qy}}) + \lambda \cdot \text{CE}(\hat{s}_{\text{qy}}, s_{\text{qy}}). \tag{1}$$

Minimizing $\mathcal{L}$ encourages accurate prediction of both the target label y and the sensitive attribute s, while the GRL before the s head ensures the encoder learns features that are not predictive of s. This objective naturally induces a fairness-accuracy trade-off, especially when the target label y and sensitive attribute s are correlated. The parameter λ in Eq. 1 controls the strength of this trade-off. As summarized in Algorithm 1, each pretraining step instantiates a new fairness task, forms context and query sets for ICL, masks the query label and sensitive attribute to preserve the intended inference setting, and then updates the shared transformer encoder and other components through the coupled label-prediction and sensitive-attribute losses.

4 Results

In this section, we describe the experimental setup, including the evaluation tasks, baseline models, and metrics, before discussing the main empirical results. To facilitate reproducibility, we

Algorithm 1 `FairTFM` pretraining

Require: Data prior $p(\mathcal{D})$, `FairTFM` parameters θ, fairness weight λ

while not converged **do**
  Sample a dataset $(X', y) \sim p(\mathcal{D})$
  Produce X by randomly removing one feature from X' as the sensitive attribute s
  if s is continuous **then**
    Discretize s into categorical groups via "stick-breaking" quantile bins
  end if
  Split (X, y, s) into context $(X_{\text{ctx}}, y_{\text{ctx}}, s_{\text{ctx}})$ and query $(X_{\text{qy}}, y_{\text{qy}}, s_{\text{qy}})$
  Mask y_{qy} and replace s_{qy} with the learnable sensitive attribute parameter
  Encode X_i, y_i, and s_i into d-dimensional vectors using their respective MLP encoders
  Encode the d-dimensional triplets (X_i, y_i, s_i) with the shared transformer encoder
  Predict query labels $\hat{y}_{\text{qy}}$ with the main head
  Predict query sensitive attributes $\hat{s}_{\text{qy}}$ with the adversarial head through a GRL
  Compute $\mathcal{L} = \text{CE}(\hat{y}_{\text{qy}}, y_{\text{qy}}) + \lambda\, \text{CE}(\hat{s}_{\text{qy}}, s_{\text{qy}})$
  Update θ by backpropagation; the GRL reverses gradients from the sensitive-attribute head into the encoder
end while

provide the inference code, pretrained checkpoints, and scripts for reproducing the main results at `anonymous.4open.science/r/FairTFM-inference`.

4.1 Experimental setup

Datasets We evaluate our model on 120 fairness tasks derived from the 2018 1-Year American Community Surveys (Ding et al., 2021), accessed through the `folktables` library, which is released under the MIT License.[1] We generate these tasks by varying three factors: (i) the prediction problem (ACSIncome, ACSPublicCoverage, ACSMobility, ACSEmployment, and ACSTravelTime), (ii) the sensitive attribute (Gender, Age, and Race), and (iii) the state from which the data is sampled (eight US states). This yields a total of $5 \times 3 \times 8 = 120$ real-world fairness tasks not used during training. Additional details are provided in Appendix B. To test generalization beyond a single data source, we further evaluate on 12 fairness tasks built from six widely used non-ACS benchmarks (Appendix E.6).

Base models We compare our method against standard machine learning baselines, including logistic regression (LR), random forest (RF), XGBoost (XGB), and k-nearest neighbours (KNN) in their default scikit-learn configuration (Pedregosa et al., 2011). We also include recent TFMs, namely TabPFNv2.5 and TabICLv2. Because these models do not incorporate any fairness intervention, they provide strong baselines for assessing the fairness limitations of existing approaches on our benchmark. These models do not receive s_{qy} at inference, similar to `FairTFM`. We additionally compare against FairPFN (Robertson et al., 2025), the only prior TFM that incorporates fairness during pretraining. FairPFN targets counterfactual fairness rather than the statistical notions we optimize, so it is not designed to minimize DP, EOD, or EOP; we include it as the closest existing approach to ours, to test whether pretraining for a causal fairness notion also yields group-fairness benefits. We use the publicly released FairPFN checkpoint, so its backbone size and pretraining prior differ from ours; this comparison should therefore be read as a comparison between published methods rather than as a controlled ablation of the fairness objective.

Metrics In addition to accuracy and AUCROC, we report three fairness metrics: demographic parity difference (DP), equalized odds difference (EOD), and equal opportunity difference (EOP). Definitions and implementation details are provided in Appendix C. Unless otherwise noted, we report metric values averaged across all 120 tasks to summarize overall fairness–accuracy trends.

4.2 Experimental Results

In this section, we present a comprehensive empirical evaluation of `FairTFM` across three complementary settings. First, we benchmark its performance in terms of both accuracy and fairness against classical methods and recent tabular foundation models. Second, we compare against fairness-aware variants of classical approaches to assess trade-offs under explicit fairness constraints. Third, we

[1] https://github.com/socialfoundations/folktables

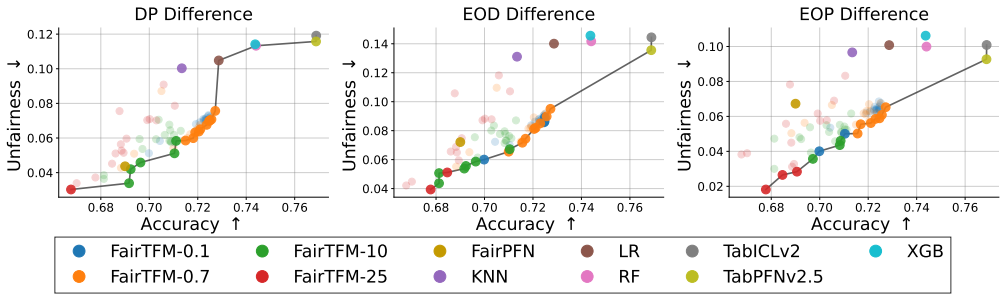


Figure 2: Pareto front between accuracy and fairness for various models. ↑ indicates higher is better (accuracy) and ↓ indicates smaller is better (unfairness).

evaluate FairTFM as a fairness-enhancing representation learner, examining its ability to produce embeddings that support fairer downstream predictions.

Fairness-accuracy trade-off across tasks Figure 2 shows our main results: the Pareto front between accuracy and fairness for various models. For FairTFM, we pretrain variants with $\lambda \in \{0.1, 0.7, 10, 25\}$ to control the fairness–accuracy trade-off, and display that trade-off for several pretraining checkpoints to more fully trace out the frontier. TabICLv2 and TabPFNv2.5 achieve the highest average accuracy across tasks, but they also occupy the most unfair region of the Pareto frontier, suggesting that their predictive gains come with a substantial fairness cost. In contrast, the classical machine learning baselines are generally less accurate, yet they remain similarly unfair.

FairTFM consistently improves fairness metrics, and for a range of λ parameters maintains competitive—and often stronger—accuracy compared to classical baselines such as LR and KNN. Table 2 in Appendix E.1 complements Figure 2 by reporting the corresponding aggregated metrics over all 120 tasks, including standard deviations across three random seeds. Whereas Figure 2 emphasizes the Pareto frontier traced out by different checkpoints and fairness weights, Table 2 makes the overall pattern explicit at the level of final checkpoints: the strongest unconstrained TFMs achieve the highest accuracy but are also the least fair, while FairTFM variants deliver substantial reductions in DP, EOD, and EOP with only a moderate loss in predictive performance. Table 3 in Appendix E.2 reports the relative changes against the strongest TFM (TabPFNv2.5), the strongest classical model (XGB), and LR: FairTFM improves fairness by 32–75% for accuracy costs of 2–11%, and even dominates LR on all metrics at $\lambda = 0.7$. Tables 2 and 3 report the last pretraining checkpoint of each FairTFM variant, while the Pareto-dominant checkpoints visible in Figures 2 and 3 yield better trade-offs, but would require model selection on a validation set.

The results also highlight the role of λ in shaping the accuracy-fairness trade-off. Larger values of λ (e.g., $\lambda = 25$) move the model toward the fairer region of the Pareto frontier at the expense of predictive accuracy. Smaller values of λ, by contrast, prioritize accuracy and therefore place the model in less fair regions of the frontier. The Pareto front ends up tracing a nearly linear path through accuracy-unfairness space, with FairTFM filling out a large segment. These trends indicate that our training objective provides a simple and effective mechanism for navigating different operating points depending on the fairness requirements of the application. We observe the same trend when measuring predictive performance using AUCROC, as shown in Figure 7 of Appendix E.

FairPFN behaves differently from the unconstrained baselines: it is fairer than all of them on all three metrics, and in Table 2 attains the lowest DP of any final checkpoint (0.043). Pretraining for a causal fairness notion therefore does transfer part of its benefit to statistical group fairness, supporting our broader claim that fairness can be acquired during pretraining and translate to ICL predictions. However, for FairPFN this comes at a large predictive cost; FairPFN has the lowest AUCROC of all models we evaluate here (0.695). In every panel of Figures 2 and 7 FairTFM provides operating points that Pareto-dominate it, being simultaneously fairer and more predictive. Just as important, FairPFN exposes no mechanism for navigating the fairness–utility trade-off: it yields a single operating point, whereas λ lets a practitioner choose one that matches the application. Part of the predictive gap is expected, since FairPFN optimizes fair outcomes under interventions on structural causal models rather than accuracy on real labels; the comparison is thus evidence that the two fairness notions are not interchangeable, not that FairPFN fails at its own objective.

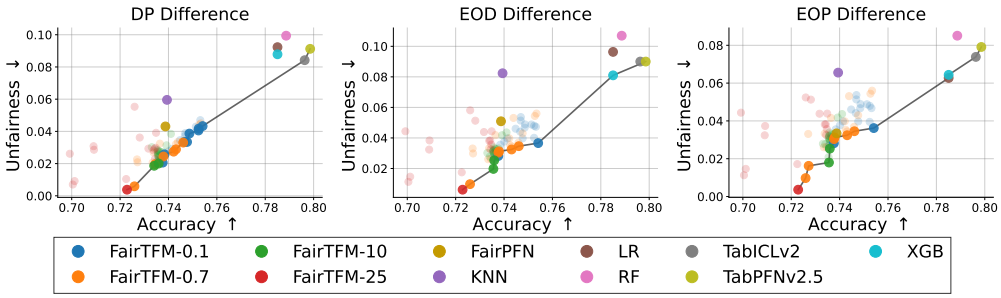


Figure 3: Pareto front between accuracy and fairness for various models on the 12 fairness tasks beyond ACS PUMS (Appendix E.6). ↑ indicates higher is better (accuracy) and ↓ indicates smaller is better (unfairness).

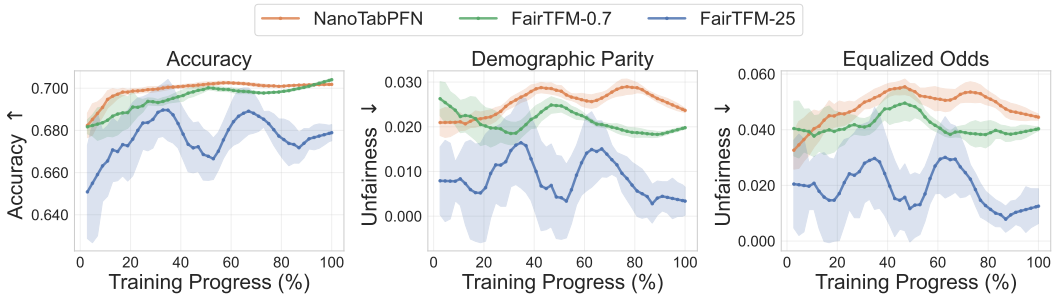


Figure 4: Training dynamics comparison between nanoTabPFN (without fairness constraint) and `FairTFM` with $\lambda \in \{0.7, 25\}$.

Generalization beyond ACS PUMS tasks Figure 3 shows the Pareto front on the 12 non-ACS tasks. The conclusions match those from Figure 2: TabICLv2 and TabPFNv2.5 are the most accurate but occupy the most unfair region, while `FairTFM` variants trace the low-unfairness end of the front, with λ controlling the trade-off. The AUCROC view (Figure 13 in Appendix E.6) further shows that `FairTFM` improves fairness without a commensurate loss in ranking performance. FairPFN behaves as on the ACS benchmark: it is fairer than every unconstrained baseline on all three metrics, but at a predictive cost that again leaves it Pareto-dominated by `FairTFM`, under both accuracy and AUCROC. Table 4 in Appendix E.6 reports the corresponding aggregated metrics and demonstrates the gap directly. At essentially the same accuracy (0.734 vs. 0.737), `FairTFM`-10 is fairer than FairPFN on all three metrics and 7.5 points higher in AUCROC.

Performance dynamics during training Figure 4 provides a finer-grained view of how fairness emerges over the course of pretraining, tracking accuracy and fairness (DP and EOD) for `FairTFM` with $\lambda \in \{0.7, 25\}$ and for nanoTabPFN across checkpoints. Both `FairTFM` variants achieve a fairness advantage over nanoTabPFN very early in training and retain this advantage throughout optimization. Since both models share the same data prior, backbone, and training budget, differing only in the fairness-specific components, this comparison isolates the effect of our intervention (full discussion in Appendix E.3).

Comparing `FairTFM` against classical fairness-aware models The results above show that `FairTFM` improves fairness substantially while remaining competitive with standard classical baselines such as LR and KNN. A stronger comparison, however, is against classical methods that are themselves explicitly optimized for fairness. To this end, we evaluate three tabular models—LR, RF, and XGB—augmented with the Exponentiated Gradient (EG) reduction of Agarwal et al. (2018), which enforces group-fairness constraints during training. We focus on ACSIncome in Alabama (AL) and instantiate three fairness tasks by varying the sensitive attribute over gender, race, and age. For each task, we use a random 80/20 train–test split and sweep the EG fairness-violation tolerance over $[0.01, 0.02, \ldots, 0.1, 0.2, \ldots, 1.0]$, with finer resolution in the low-violation regime to better characterize the high-fairness end of the trade-off. We average the results across three random seeds. This setup allows us to compare `FairTFM` not only to strong predictive baselines, but also to established in-processing fairness methods under a matched downstream evaluation protocol.

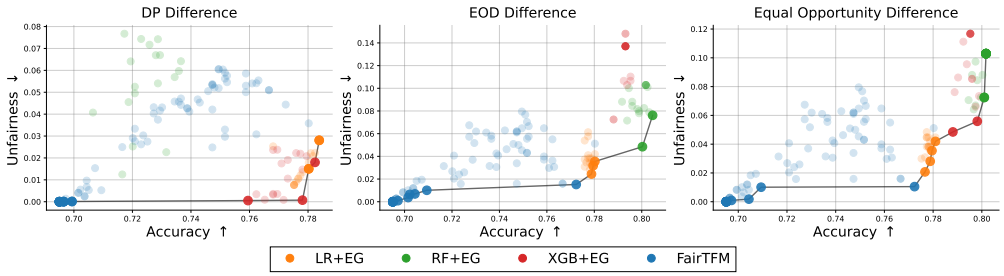


Figure 5: Fairness-accuracy trade-off comparison with task-specific classical models trained with fairness constraints using Exponentiated Gradient (EG), on ACSIncome (Alabama) with gender as the sensitive attribute. Results for race and age are shown in Figure 11 of Appendix E.

Figure 5 shows that `FairTFM` remains highly competitive even against these fairness-constrained baselines. Across all three sensitive attributes, its checkpoints trace out a broad Pareto frontier, indicating that a single pretrained model can realize multiple fairness–accuracy operating points without retraining. For gender, `FairTFM` attains especially strong trade-offs for EOD and EOP, matching or improving upon the frontier formed by EG-based baselines while remaining competitive on DP. For race (Figure 11 in Appendix E), where all methods incur larger fairness gaps, `FairTFM` still spans a wide and competitive portion of the frontier, particularly at the lower-unfairness end. For age, the comparison is more metric-dependent: EG-based classical models achieve stronger accuracy–fairness trade-offs for EOD and EOP, whereas `FairTFM` remains competitive on DP but exhibits a clearer trade-off between predictive performance and these stricter parity criteria. Overall, the main advantage of `FairTFM` is not that it dominates every baseline on every metric, but that it delivers competitive Pareto-efficient solutions across heterogeneous fairness tasks in a single forward pass, whereas the classical alternatives must be retrained with task-specific fairness constraints for each new setting. Moreover, as shown in Figure 12 of Appendix E, `FairTFM` provides the best overall Pareto front when predictive performance is measured through AUCROC.

We also evaluated the EG baselines on the 12 non-ACS tasks with results in Appendix E.7. Each `FairTFM` setting improves the metrics that are relevant to each EG variant—by 20–33% for `FairTFM`-0.7 and 79–82% for `FairTFM`-25—while achieving substantially higher AUCROC: EG reduces XGBoost's AUCROC from 0.803 to about 0.70, whereas `FairTFM`-0.7 remains at 0.803.

`FairTFM` as a fair representation learner We hypothesize that the `FairTFM` encoder produces representations that are useful for downstream fair prediction. Fair representation learning methods learn a map g of the input that suppresses information about the sensitive attribute while preserving task-relevant signal; formally:

Definition 1 (Fair Representation) *Let (X, Y, S) be random variables where $X \in \mathbb{R}^{m\times d}$ is the input, $Y \in \mathcal{Y}$ is the target label, and $S \in \mathcal{S}$ is a sensitive attribute. Let $g : \mathbb{R}^{m\times d} \to \mathbb{R}^{m\times d'}$ be a representation map and define $Z := g(X)$. We say that Z is a fair representation with respect to S if $Z \perp\!\!\!\perp S \mid Y$.*

Definition 1 is conceptually aligned with the objective of `FairTFM`, which encourages the learned query embedding to suppress sensitive-attribute information while preserving task-relevant signal. We compare `FairTFM` embeddings against two task-specific pre-processing methods from the `fairlearn` library (Bird et al., 2020): Correlation Remover (CR) and Learning Fair Representations (LFR) (Zemel et al., 2013), on ACSIncome (Alabama) with gender, race, and age as sensitive attributes. Notably, `FairTFM` embeddings are computed in a single forward pass without any task-specific gradient updates. For each type of representation as well as the raw features, we train downstream classifiers (LR, RF, and XGBoost), and compare their accuracy and fairness properties. Full protocol details are in Appendix E.4.

Figure 6 showcases the Pareto front of XGBoost models trained using data representation from different fair methods, with gender as the sensitive attribute. As can be seen, `FairTFM` can provide competitive Pareto points on downstream tasks compared to raw data and dominates LFR, whereas CR severely worsens fairness. The results show the ability of `FairTFM` to generate embeddings that encode less information about the given sensitive attributes while still being predictive of the target.

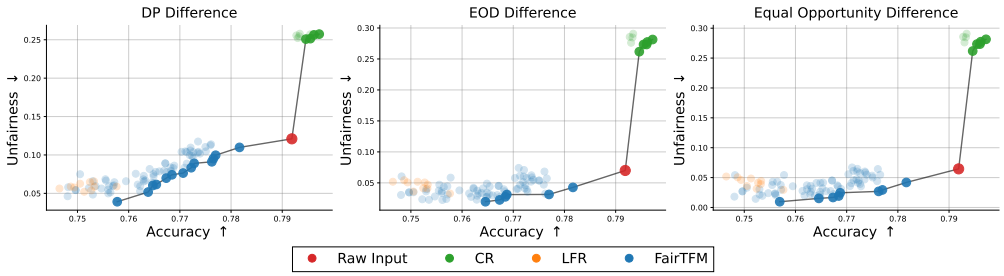


Figure 6: Pareto front of XGBoost models trained with different data representations, with gender as the sensitive attribute. Results for race and age, and for RF and LR as downstream models, are shown in Appendix E.4.

Table 1: Linear probe AUC for predicting the sensitive attribute s and the label y on different representations. We show AUC mean $\pm$ std over three seeds and the three sensitive attributes on ACSIncome, Alabama. `FairTFM` displays a controllable trade-off of reducing the predictability of the sensitive attribute with some obfuscation of the label.

Probe target (AUC)	Raw input	nanoTabPFN	`FairTFM`-0.7	`FairTFM`-25
Sensitive attribute s	0.783 ± 0.11	0.746 ± 0.10	0.646 ± 0.14	0.542 ± 0.11
Label y	0.834 ± 0.02	0.823 ± 0.03	0.805 ± 0.04	0.787 ± 0.03

`FairTFM` stands out due to its ability to adapt to new tasks via ICL without needing task-specific optimization. The remaining sensitive attributes (race and age) and downstream models (RF and LR) in Appendix E.4 show a similar trend.

Probing the learned representations To verify that the GRL removes information about the designated sensitive attribute rather than degrading the embedding broadly (Section 3), we train linear probes to predict (a) the sensitive attribute s and (b) the label y from three representations: the raw input features X, the query embeddings of nanoTabPFN, and the query embeddings of `FairTFM`. For the embedding-based representations, the query embeddings of the initial test set are treated as a new dataset for each probing task. For both sensitive-attribute and label prediction, this dataset is randomly split into 80% training and 20% testing subsets. The probing model (XGBoost) is trained on the training split, and performance is evaluated on the held-out test split using AUC.

Table 1 shows three things. First, the nanoTabPFN embedding retains most of the sensitive information present in the raw features (AUC 0.746 vs. 0.783) even though s_{qy} itself is not an input. Second, `FairTFM` substantially reduces the sensitive-attribute AUC, approaching random guessing (0.542 at $\lambda = 25$). Third, task-relevant information is largely preserved (AUC $0.834 \rightarrow 0.805$ at $\lambda = 0.7$). These probes support the idea that suppression of sensitive information is attributable to the adversarial objective and is selective.

5 Conclusion

We present `FairTFM`, the first pretraining framework for tabular foundation models that provides group fairness while preserving the single-pass inference setting of in-context learning. The key idea is to synthesize fairness tasks during pretraining and couple label prediction with adversarial sensitive-attribute prediction through a gradient reversal layer. On 120 ACS-based fairness tasks and 12 additional tasks from six non-ACS benchmarks, the resulting models consistently improve demographic parity, equal opportunity, and equalized odds while maintaining competitive accuracy. Beyond the aggregate benchmark results, we showed that these gains emerge throughout training, remain competitive against fairness-constrained classical baselines, and transfer to downstream fair-representation settings, with probing experiments confirming that the learned embeddings selectively remove sensitive information. Taken together, these results suggest that fairness can be incorporated directly into the pretraining of TFMs, rather than introduced only through task-specific post-hoc correction as done in prior work.

AI USE STATEMENT

In this work, we used generative AI tools for coding assistance, implementation of methods, and editing of the paper in select parts. We have not used generative AI tools for writing the bulk of the paper, generating synthetic data sets, cleaning data sets, supporting data analysis, interpreting results, or proposing or refining hypotheses, and tasks related to theory, mathematical claims, proofs, or translation are not applicable to this work. Additionally, we used generative AI tools for producing code to plot experimental data, conducting literature searches, and providing feedback on the overall paper. We have reviewed all AI-assisted work, for example by manually curating papers from the literature search into our bibliography and discussions of related work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to reduce unfair outcomes from TFMs by incorporating fairness objectives during pretraining. Statistical fairness criteria such as demographic parity, equal opportunity, and equalized odds are particular notions of fairness. Achieving them does not guarantee the absence of discrimination or other harms in real-world deployments. We recommend that practitioners define fairness in a way that is contextually appropriate, and evaluate any use of `FairTFM` with appropriate metrics before deployment.

REPRODUCIBILITY STATEMENT

We have made significant efforts to ensure that this work is reproducible. Our methodology is thoroughly explained in Sections 3 and 4. Additional experimental details have been included in Appendices B, C, and D. Finally, we have included source code for our implementation at `anonymous.4open.science/r/FairTFM-inference`, which includes scripts for reproducing important parts of our work. This codebase has been released publicly in a non-anonymized version.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 60–69, 2018.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pp. 254–264. Auerbach Publications, 2022.

Vahid Balazadeh, Hamidreza Kamkari, Valentin Thomas, Benson Li, Junwei Ma, Jesse C Cresswell, and Rahul G Krishnan. CausalPFN: Amortized Causal Effect Estimation via In-Context Learning. In *Advances in Neural Information Processing Systems*, 2025.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A Cookbook of Self-Supervised Learning . *arXiv:2304.12210*, 2023.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft, 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever,

and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

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*. Association for Computing Machinery, 2016. ISBN 9781450342322. doi: 10.1145/2939672.2939785.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 1597–1607, 2020.

Jesse C Cresswell. Trustworthy AI must account for interactions. *arXiv:2504.07170*, 2025.

Aaron Defazio, Xingyu Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The road less scheduled. In *Advances in Neural Information Processing Systems*, volume 37, pp. 9974–10007, 2024. doi: 10.52202/079017-0320.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring Adult: New Datasets for Fair Machine Learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

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*, 2012. ISBN 9781450311151. doi: 10.1145/2090236.2090255.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Desai, David Salinas, and Frank Hutter. TabArena: A Living Benchmark for Machine Learning on Tabular Data. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-0519.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1180–1189, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Anurag Garg, Muhammad Ali, Noah Hollmann, Lennart Purucker, Samuel Müller, and Frank Hutter. Real-TabPFN: Improving Tabular Foundation Models via Continued Pre-training With Real-World Data. *arXiv:2507.03971*, 2025.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18932–18943, 2021.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, Dominik Safaric, Simone Alessi, Adrian Hayler, Mihir Manium, Rosen Yu, Felix Jablonski, Shi Bin Hoo, Anurag Garg, Jake Robertson, Magnus Bühler, Vladyslav Moroshan, Lennart Purucker, Clara Cornu, Lilly Charlotte Wehrhahn, Alessandro Bonetto, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. TabPFN-2.5: Advancing the State of the Art in Tabular Foundation Models. *arXiv:2511.08667*, 2026.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. In *International Conference on Learning Representations*, 2023.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pp. 1–6, 2009. doi: 10.1109/IC4.2009.4909197.

Patrik Kenfack, Samira Ebrahimi Kahou, and Ulrich Aïvodji. Towards fair in-context learning with tabular foundation models. *Transactions on Machine Learning Research*, 2026, 2026.

Mohammad Khan, Shakir Mohamed, Benjamin Marlin, and Kevin Murphy. A stick-breaking likelihood for categorical data analysis with latent gaussian models. In *Artificial Intelligence and Statistics*, pp. 610–618. PMLR, 2012.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):e1452, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Hamidreza Kamkari, Jesse C. Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L. Caterini, and Maksims Volkovs. TabDPT: Scaling Tabular Foundation Models on Real Data. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pp. 3384–3393. PMLR, 2018.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, 54(6), 2021. ISSN 0360-0300. doi: 10.1145/3457607.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Alexander Pfefferle, Johannes Hog, Lennart Purucker, and Frank Hutter. nanoTabPFN: A Lightweight and Educational Reimplementation of TabPFN. *arXiv:2511.03634*, 2025.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A Tabular Foundation Model for In-Context Learning on Large Data. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 50817–50847, 2025.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. In *Forty-third International Conference on Machine Learning*, 2026.

Jake Robertson, Noah Hollmann, Samuel Müller, Noor Awad, and Frank Hutter. FairPFN: A Tabular Foundation Model for Causal Fairness. In *Forty-second International Conference on Machine Learning*, 2025.

Christopher Stith, Medha Barath, Vahid Balazadeh, Jesse C Cresswell, and Rahul G Krishnan. Causal Foundation Models with Continuous Treatments. *arXiv:2605.15133*, 2026.

Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014(1):781670, 2014.

Yi Sui, Tongzi Wu, Jesse Cresswell, Ga Wu, George Stein, Xiaoshi Huang, Xiaochen Zhang, and Maksims Volkovs. Self-supervised representation learning from random data projectors. In *International Conference on Learning Representations*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Linda F Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

A Additional Related Work

Fairness interventions Classical fairness-aware learning methods are often grouped by where the intervention occurs in the pipeline (Cresswell, 2025). Pre-processing methods attempt to reduce bias directly in the data, for example through label massaging (Kamiran & Calders, 2009) or fair representation learning (Zemel et al., 2013). In-processing methods instead modify the training objective to encourage fairer behavior during optimization (Zhang et al., 2018). Post-processing methods operate on a trained model's outputs to enforce fairness constraints after training (Hardt et al., 2016). These families of methods have been influential, but they typically assume task-specific model access, retraining, or output recalibration.

Fairness in in-context learning These assumptions become restrictive for foundation models used through ICL, where predictions are generated without task-specific parameter updates. In our setting, the goal is not to retrofit fairness onto a frozen predictor after deployment, but to pretrain a model whose internal representations already support fairer predictions at inference time. Our method therefore combines aspects of in-processing and representation learning: fairness is encouraged during training, yet inference remains a single forward pass without additional optimization.

Tabular foundation models Work on deep learning for tabular data has evolved from specialized supervised architectures (Gorishniy et al., 2021) to pretrained tabular foundation models that generalize across tasks through ICL. TabPFN (Hollmann et al., 2023) established this direction by showing that transformers trained on synthetic supervised tasks can perform strong zero-shot prediction. Subsequent models, including TabPFNv2.5 (Grinsztajn et al., 2026) and TabICLv2 (Qu et al., 2026), improved efficiency and scale, while models such as TabDPT (Ma et al., 2025) and related real-data pretraining approaches (Garg et al., 2025) sought to better align training with real-world tabular distributions.

Fairness-aware TFMs Among prior TFM work, FairPFN (Robertson et al., 2025) is the closest to our setting because it also incorporates fairness during pretraining. However, FairPFN is designed around a causal notion of fairness, using structurally generated datasets with biased and fair outcomes to target counterfactual fairness (Kusner et al., 2017). Our focus is different: we study statistical group fairness notions, namely demographic parity, equal opportunity, and equalized odds, which are more directly aligned with the evaluation metrics commonly used in fair classification benchmarks. Statistical fairness is data-driven and focuses on distributional fairness (outcome-based), while counterfactual fairness is more interventional and does not necessarily ensure outcome parity across groups

B Datasets

In this section, we provide more details about the datasets used for evaluation. We describe the prediction tasks and the dataset construction.

B.1 Prediction Task Details

We construct our benchmark from tasks provided by the `folktables` benchmark (Ding et al., 2021), which is derived from the American Community Survey (ACS) Public Use Microdata Sample (PUMS). In particular, we consider five binary prediction tasks commonly used in prior work for fairness evaluation:

- **ACSIncome**: predict whether an individual's annual income exceeds $50,000. Following the standard task definition, we restrict the data to individuals older than 16 who worked at least one hour per week during the previous year and earned at least $100.
- **ACSMobility**: predict whether an individual lived at the same address one year earlier. We focus on individuals between 18 and 35 years old, which makes the task less imbalanced than in the full population, where most individuals do not move within a year.
- **ACSTravelTime**: predict whether an individual's commute exceeds 20 minutes. The task is defined on employed individuals older than 16, and the 20-minute threshold roughly matches the median commute time in the 2018 ACS PUMS data.
- **ACSEmployment**: predict whether an individual is employed. For this task, we consider individuals between 16 and 90 years old.

- **ACSPublicCoverage**: predict whether an individual receives public health insurance coverage. We restrict the sample to individuals younger than 65 with income below $30,000, thereby focusing on lower-income individuals who are not eligible for Medicare.

B.2 Task Construction

Generating new learning signals from randomly selected features in pretraining data is a core concept in self-supervised learning (Sui et al., 2024). In our work, we specifically construct fairness tasks, which is necessary to scale up pretraining, due to insufficient existing datasets labeled with fairness attributes. For each base prediction task, we instantiate fairness evaluation settings by combining it with three sensitive attributes—Gender, Age, and Race—and with data drawn from eight states: Alabama (AL), California (CA), Hawaii (HI), Indiana (IN), Maine (ME), Michigan (MI), New Mexico (NM), and New York (NY). We choose these states to span a range of bias levels reported in the original `folktables` study (Ding et al., 2021), so that the benchmark includes settings with meaningfully different fairness profiles rather than a narrow slice of the ACS. This yields a total of $5 \times 3 \times 8 = 120$ tasks. In other words, each prediction problem contributes 24 tasks, each sensitive attribute appears in 40 tasks, and each state contributes 15 tasks to the full benchmark. For the sensitive attributes, Age is binarized using a 25-year-old threshold, and Race is restricted to White and Black Americans. Across all task instantiations, dataset sizes vary from roughly 3.5k to 12k samples, and we use a random 80/20 train–test split.

This evaluation design is important for two reasons. First, varying the prediction problem, sensitive attribute, and state allows us to assess fairness across substantially different label distributions, demographic compositions, and regional contexts, rather than tailoring conclusions to a single task configuration. Second, reporting results over the full Cartesian product reduces the risk that observed fairness improvements are driven by a small number of favorable settings. We therefore view performance aggregated over these 120 tasks as a stronger indicator of whether a method learns fairness-aware behavior that transfers across heterogeneous real-world tabular prediction problems.

C Fairness metrics

In this work, we focus on group fairness criteria that quantify disparities in model behavior across demographic groups. Let $\hat{Y} = f(X)$ denote the binary prediction of a classifier, let $Y \in \{0, 1\}$ be the ground-truth label, and let $S \in \{0, 1\}$ denote the sensitive attribute. We consider the following three standard fairness notions.

- **Demographic parity (DP)** requires the rate of positive predictions to be the same across groups (Dwork et al., 2012). Formally,
$$\mathbb{P}(\hat{Y} = 1 \mid S = 0) = \mathbb{P}(\hat{Y} = 1 \mid S = 1). \tag{2}$$
- **Equalized odds (EOD)** requires the predictor to have the same true positive and false positive rates across groups (Hardt et al., 2016). Equivalently, for each label value $y \in \{0, 1\}$,
$$\mathbb{P}(\hat{Y} = 1 \mid S = 0, Y = y) = \mathbb{P}(\hat{Y} = 1 \mid S = 1, Y = y). \tag{3}$$
- **Equal opportunity (EOP)** focuses only on parity of true positive rates across groups. It can be viewed as the $y = 1$ special case of equalized odds:
$$\mathbb{P}(\hat{Y} = 1 \mid S = 0, Y = 1) = \mathbb{P}(\hat{Y} = 1 \mid S = 1, Y = 1). \tag{4}$$

In the experiments, we report empirical disparity versions of these metrics. For demographic parity, we use the absolute difference in the expected positive prediction rate across groups:

$$\mathrm{DP} = \left| \mathop{\mathbb{E}}_{x|S=0} \left[\mathbb{I}\{\hat{Y} = 1\} \right] - \mathop{\mathbb{E}}_{x|S=1} \left[\mathbb{I}\{\hat{Y} = 1\} \right] \right|. \tag{5}$$

Where $\mathbb{I}(\cdot)$ denotes the indicator function.

For the equalized-odds-based metrics, we define the group gaps in false positive rate and true positive rate using the same expectation notation:

$$\delta_{\mathrm{FPR}} = \left| \mathop{\mathbb{E}}_{x|S=0,Y=0} \left[\mathbb{I}\{\hat{Y} = 1\} \right] - \mathop{\mathbb{E}}_{x|S=1,Y=0} \left[\mathbb{I}\{\hat{Y} = 1\} \right] \right|, \tag{6}$$

$$\delta_{\mathrm{TPR}} = \left| \underset{x|S=0,Y=1}{\mathbb{E}} \left[\mathbb{I}\{\hat{Y}=1\} \right] - \underset{x|S=1,Y=1}{\mathbb{E}} \left[\mathbb{I}\{\hat{Y}=1\} \right] \right|. \tag{7}$$

We then report

$$\mathrm{EOD} = \max\left(\delta_{\mathrm{FPR}}, \delta_{\mathrm{TPR}}\right) \tag{8}$$

$$\mathrm{EOP} = \delta_{\mathrm{TPR}}. \tag{9}$$

Smaller values of DP, EOD, and EOP indicate fairer behaviour, with zero corresponding to perfect parity under the respective criterion. We use these empirical gaps because they provide an interpretable summary of group-level disparities and are standard in fairness evaluations for binary classification.

D Model Architecture and Hyperparameters

Our model builds on the nanoTabPFN architecture from `TFM-Playground` [2], which we use as a lightweight transformer backbone for fairness-aware pretraining. Concretely, we use a model with 6 transformer layers, 6 attention heads, embedding dimension 192, and feed-forward hidden dimension 192. Consistent with the main architecture described in Section 3, we augment this backbone with three input encoders—for features, targets, and sensitive attributes—that map their respective inputs into the shared 192-dimensional token space. The label-prediction head and the sensitive-attribute head are both implemented as two-layer MLPs with hidden size 768.

Pretraining setup. We pretrain on 300,000 synthetically generated tabular datasets sampled from the TabICL's prior implementation (Qu et al., 2026). Each sampled task contains 150 datapoints, 6 features, and 2 classes, and training is performed with batch size 32. During pretraining, the model receives context and query sets as described in Section 3, with the query label masked and the query sensitive attribute replaced by the learned mask token. All reported `FairTFM` results are obtained from checkpoints trained under this shared setup, varying only the fairness weight λ in the joint objective.

Optimization details. We optimize the model using Schedule-Free AdamW (Defazio et al., 2024; Loshchilov & Hutter, 2019) with learning rate 1×10^{-4} and no weight decay. This choice provided stable optimization across the fairness weights considered in the main experiments. For the fairness-aware variants, the only task-level hyperparameter we vary is λ, which directly controls the strength of the adversarial sensitive-attribute objective and thereby the fairness–accuracy trade-off.

E Additional Results

E.1 Aggregated results on the 120 ACS tasks

Table 2 reports the aggregated accuracy and fairness metrics behind Figure 2, and Figure 7 shows the corresponding Pareto front when predictive performance is measured with AUCROC instead of accuracy. The two views agree: unconstrained TFMs sit in the most accurate but least fair region, while `FairTFM` variants trace the fair end of the front.

E.2 Percent improvement over the strongest baselines

To make the size of the fairness gains explicit, Table 3 reports the relative change of each `FairTFM` variant against the strongest TFM baseline (TabPFNv2.5), the strongest classical baseline (XGB), and LR, computed from the averages in Table 2. Positive fairness numbers mean fairer (lower DP, EOD, or EOP), and negative accuracy and AUCROC numbers mean lower predictive performance. Three points stand out. First, the fairness gains are large relative to the predictive costs: `FairTFM`-0.7 improves DP, EOD, and EOP by 32–36% for a 5.3% accuracy cost against TabPFNv2.5, and only a 2% accuracy and 0.8% AUCROC cost against XGB. Second, against LR—the classical baseline closest in accuracy—`FairTFM`-0.7 improves every metric at once: slightly higher accuracy (+0.1%), higher AUCROC (+2.9%), and 27–31% better fairness. Third, the gains grow monotonically with λ, up to 51–75% for `FairTFM`-25, confirming that λ provides direct control over the fairness–accuracy trade-off.

[2] https://github.com/automl/TFM-Playground

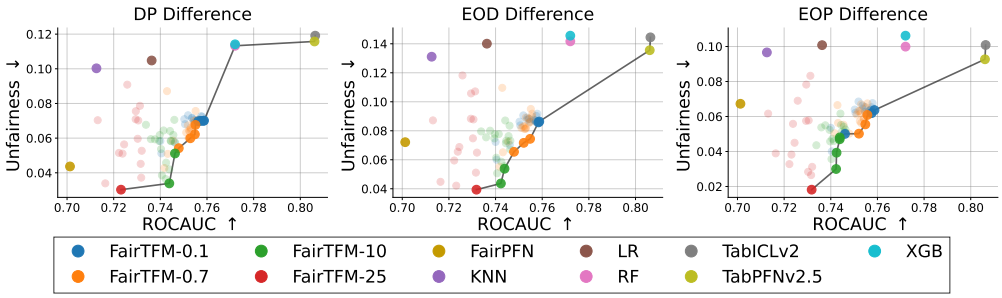


Figure 7: Pareto front between AUCROC and fairness for various models. ↑ indicates higher is better (accuracy) and ↓ indicates smaller is better (unfairness).

Table 2: Average accuracy, AUCROC, and fairness metrics across the 120 benchmark tasks, reported as mean ± standard deviation over three random seeds. This table complements Figure 2 by summarizing the final-checkpoint performance of `FairTFM` models. Recent unconstrained TFMs attain the highest average accuracy but also the largest fairness gaps, classical baselines are generally less accurate without being substantially fairer, and `FairTFM` variants provide better fairness–accuracy trade-offs, with larger λ yielding progressively lower unfairness at the cost of reduced accuracy.

Model	Accuracy	AUCROC	DP Diff	EOD Diff	EOP Diff
TabICLv2	$\mathbf{0.771 \pm 0.05}$	$\mathbf{0.802 \pm 0.07}$	0.120 ± 0.11	0.140 ± 0.10	0.096 ± 0.09
TabPFNv2.5	0.770 ± 0.05	0.801 ± 0.07	0.121 ± 0.12	0.145 ± 0.11	0.098 ± 0.10
XGB	0.744 ± 0.06	0.760 ± 0.09	0.116 ± 0.11	0.144 ± 0.11	0.097 ± 0.10
KNN	0.716 ± 0.07	0.707 ± 0.10	0.106 ± 0.11	0.139 ± 0.10	0.093 ± 0.09
RF	0.745 ± 0.06	0.767 ± 0.08	0.113 ± 0.11	0.147 ± 0.11	0.100 ± 0.10
LR	0.728 ± 0.06	0.733 ± 0.10	0.107 ± 0.12	0.134 ± 0.11	0.094 ± 0.09
FairPFN	0.690 ± 0.07	0.695 ± 0.11	$\mathbf{0.043 \pm 0.07}$	0.078 ± 0.11	0.069 ± 0.10
`FairTFM-0.7`	0.729 ± 0.07	0.754 ± 0.10	0.077 ± 0.13	0.098 ± 0.13	0.065 ± 0.10
`FairTFM-1.0`	0.726 ± 0.07	0.751 ± 0.10	0.075 ± 0.12	0.091 ± 0.12	0.064 ± 0.09
`FairTFM-10`	0.711 ± 0.07	0.743 ± 0.10	0.063 ± 0.12	0.076 ± 0.12	0.051 ± 0.08
`FairTFM-25`	0.688 ± 0.06	0.731 ± 0.10	0.059 ± 0.14	$\mathbf{0.072 \pm 0.15}$	$\mathbf{0.025 \pm 0.06}$

E.3 Performance dynamics during training

Figure 4 in the main text provides a finer-grained view of how fairness emerges over the course of pretraining. We track accuracy and fairness (DP and EOD) for `FairTFM` with $\lambda \in \{0.7, 25\}$ and for nanoTabPFN, evaluating checkpoints throughout training and averaging the resulting curves over 10 real-world fairness tasks from our benchmark. The shaded regions indicate one standard deviation across these tasks. Since lower DP and EOD correspond to fairer behavior, the middle and right panels show that both `FairTFM` variants achieve a fairness advantage over nanoTabPFN very early in training and retain this advantage throughout optimization. This gap is not confined to a narrow set of checkpoints: it persists across most of the training trajectory and becomes especially pronounced in the later stages. We note that the nanoTabPFN reference is trained under the same data prior, backbone, optimizer, and training budget as `FairTFM`; the two differ only in the fairness-specific components (the sensitive-feature designation in task sampling, the added encoder and adversarial head, and the fairness loss), so this comparison isolates the effect of our intervention.

At the same time, the figure makes the fairness–accuracy trade-off induced by λ visually explicit. Specifically, nanoTabPFN and `FairTFM` with $\lambda = 0.7$ converge to very similar final accuracies, but the latter does so while maintaining consistently lower DP and EOD gaps. By contrast, `FairTFM` with $\lambda = 25$ occupies a distinctly different operating regime: its accuracy remains below that of the other two models, but it attains by far the lowest unfairness on both metrics. The trajectories are also mildly non-monotonic, which is expected because fairness and accuracy are measured on heterogeneous downstream tasks rather than on the pretraining objective itself. Taken together, these dynamics show

Table 3: Fairness improvement (positive numbers mean fairer) and accuracy and AUCROC change of each FairTFM variant relative to (a) the best TFM (TabPFNv2.5), (b) the best classical model (XGB), and (c) LR, averaged over the 120 benchmark tasks and derived from Table 2.

(a) vs. TabPFNv2.5

Model	Acc	AUCROC	DP	EOD	EOP
FairTFM-0.7	-5.3%	-5.9%	$\mathbf{+36.4\%}$	$\mathbf{+32.4\%}$	$\mathbf{+33.7\%}$
FairTFM-1.0	-5.7%	-6.2%	$\mathbf{+38.0\%}$	$\mathbf{+37.2\%}$	$\mathbf{+34.7\%}$
FairTFM-10	-7.7%	-7.2%	$\mathbf{+47.9\%}$	$\mathbf{+47.6\%}$	$\mathbf{+48.0\%}$
FairTFM-25	-10.6%	-8.7%	$\mathbf{+51.2\%}$	$\mathbf{+50.3\%}$	$\mathbf{+74.5\%}$

(b) vs. XGB

Model	Acc	AUCROC	DP	EOD	EOP
FairTFM-0.7	-2.0%	-0.8%	$\mathbf{+33.6\%}$	$\mathbf{+31.9\%}$	$\mathbf{+33.0\%}$
FairTFM-1.0	-2.4%	-1.2%	$\mathbf{+35.3\%}$	$\mathbf{+36.8\%}$	$\mathbf{+34.0\%}$
FairTFM-10	-4.4%	-2.2%	$\mathbf{+45.7\%}$	$\mathbf{+47.2\%}$	$\mathbf{+47.4\%}$
FairTFM-25	-7.5%	-3.8%	$\mathbf{+49.1\%}$	$\mathbf{+50.0\%}$	$\mathbf{+74.2\%}$

(c) vs. LR

Model	Acc	AUCROC	DP	EOD	EOP
FairTFM-0.7	$+0.1\%$	$+2.9\%$	$\mathbf{+28.0\%}$	$\mathbf{+26.9\%}$	$\mathbf{+30.9\%}$
FairTFM-1.0	-0.3%	$+2.5\%$	$\mathbf{+29.9\%}$	$\mathbf{+32.1\%}$	$\mathbf{+31.9\%}$
FairTFM-10	-2.3%	$+1.4\%$	$\mathbf{+41.1\%}$	$\mathbf{+43.3\%}$	$\mathbf{+45.7\%}$
FairTFM-25	-5.5%	-0.3%	$\mathbf{+44.9\%}$	$\mathbf{+46.3\%}$	$\mathbf{+73.4\%}$

that our objective changes the optimization path, steering the encoder toward representations that are progressively less informative about the sensitive attribute, with λ providing a direct handle on the final fairness–accuracy operating point.

E.4 Fair representation learning: protocol details and additional results

The fairness condition in Definition 1 is equivalent to $\mathbb{P}(Z \mid Y, S) = \mathbb{P}(Z \mid Y)$ or, in information-theoretic terms, $I(Z; S \mid Y) = 0$. A downstream model trained on Z is therefore expected to exhibit improved fairness relative to a model trained directly on X.

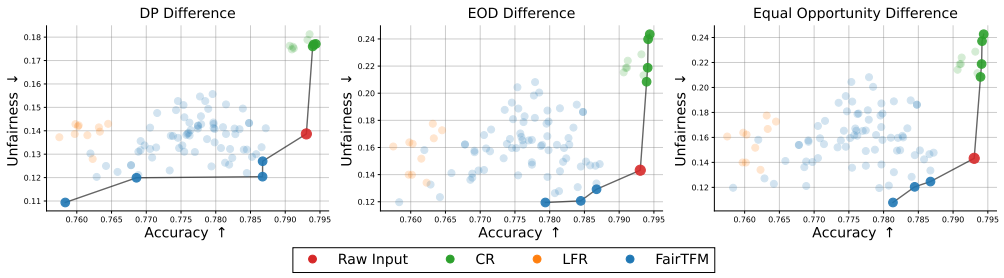
Existing approaches for learning g are typically task-specific, requiring a separate model to be fitted for each downstream task (Zemel et al., 2013; Madras et al., 2018). We compare FairTFM against two open-source, sklearn-compatible pre-processing methods from the fairlearn library (Bird et al., 2020). The first is Correlation Remover (CR), which reduces linear dependence on the sensitive attribute by applying a linear transformation to the non-sensitive features (Bird et al., 2020). The second is Learning Fair Representations (LFR), which maps inputs to latent prototypes while encouraging similar assignment behaviour across demographic groups (Zemel et al., 2013).

We consider ACSIncome for Alabama (AL) with three sensitive attributes—gender, race, and age—yielding three downstream fairness tasks. Each task is randomly split into 80% training data and 20% test data. For CR and LFR, the training split is used to fit the representation map g, which is then applied to both the training and test features. For FairTFM, we use the training split as context with the test as query, and compute embeddings for both training and test examples from the transformer's output. Importantly, these representations are obtained in a single forward pass, without any task-specific gradient updates.

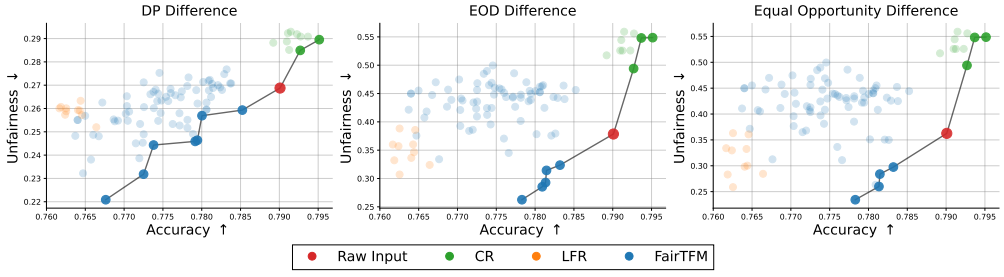
We train three downstream classifiers—LR, RF, and XGB—on four types of representations: the raw input features and the fair representations produced by CR, LFR, and FairTFM. For CR and LFR, we sweep the parameter controlling the fairness–accuracy trade-off over

$[0.01, 0.02, \ldots, 0.1, 0.2, \ldots, 1.0]$, with denser coverage in the higher-fairness regime. For `FairTFM`, we evaluate the checkpoints associated with the λ values used in Figure 2. We run this experiment across three random seeds and average the results.

Figure 8 shows the XGBoost results for the remaining sensitive attributes (race and age; gender is in Figure 6 of the main text), and Figures 9 and 10 show the Pareto fronts with Random Forest and Logistic Regression as downstream models. Across all three downstream models, `FairTFM` embeddings provide competitive fairness–accuracy trade-offs without any task-specific fitting of the representation map, while CR is most effective for the linear LR model, as it only removes linear correlation with the sensitive attribute.



(a) Race



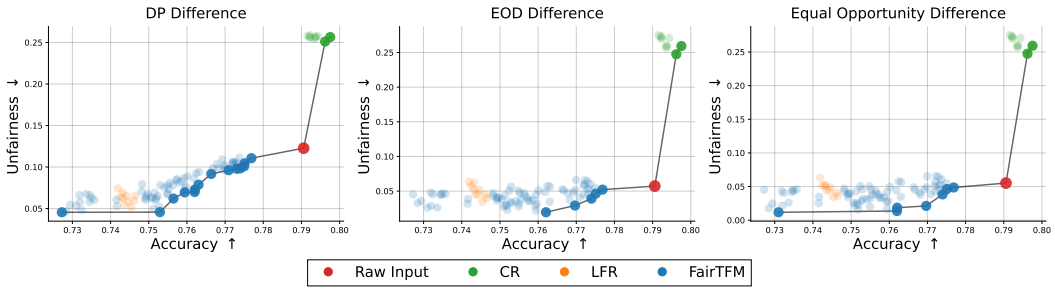
(b) Age

Figure 8: Pareto front of XGBoost models trained with different data representations, for race and age as sensitive attributes (gender is shown in Figure 6 of the main text).

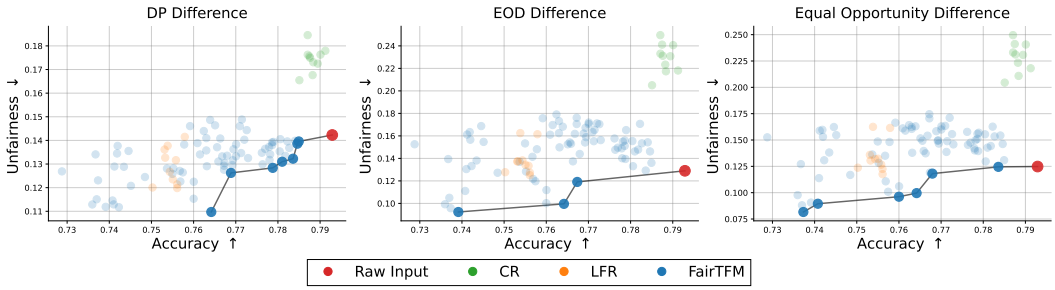
E.5 Additional Exponentiated Gradient results

Figure 11 complements Figure 5 in the main text by showing the fairness–accuracy trade-off against the EG-constrained baselines when race and age are used as sensitive attributes (same setup, ACSIncome in Alabama). For race, where all methods incur larger fairness gaps, `FairTFM` spans a wide and competitive portion of the frontier, particularly at the lower-unfairness end. For age, EG-based classical models achieve stronger accuracy–fairness trade-offs for EOD and EOP, whereas `FairTFM` remains competitive on DP.

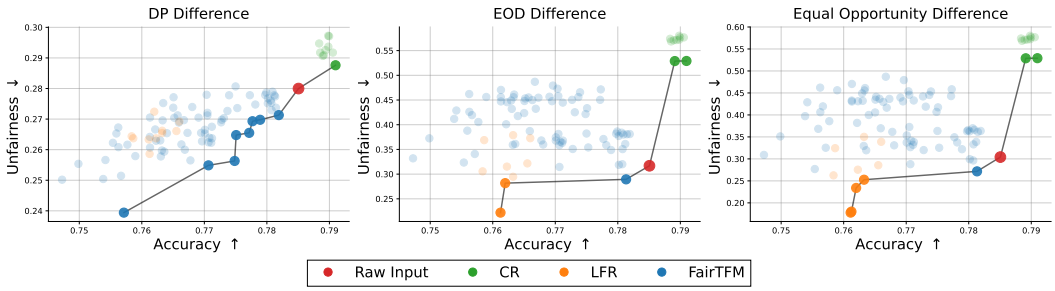
Figure 12 additionally measures predictive performance with AUCROC instead of accuracy. Because the EG reduction produces randomized ensembles that trade calibrated scores for constraint satisfaction, its AUCROC drops sharply, and `FairTFM` provides the best overall Pareto front across all three sensitive attributes under this view.



(a) Gender

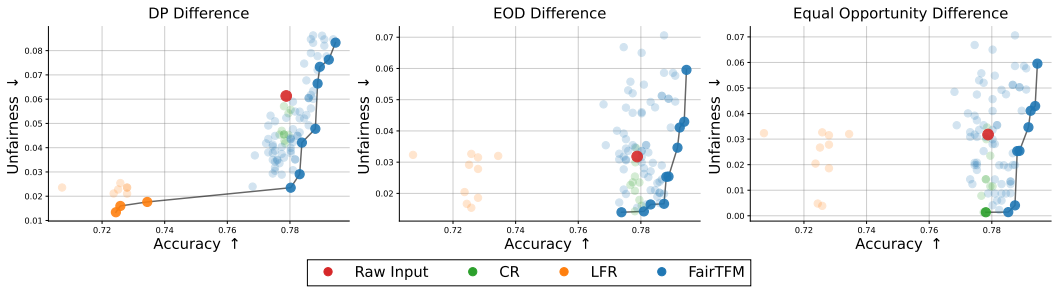


(b) Race

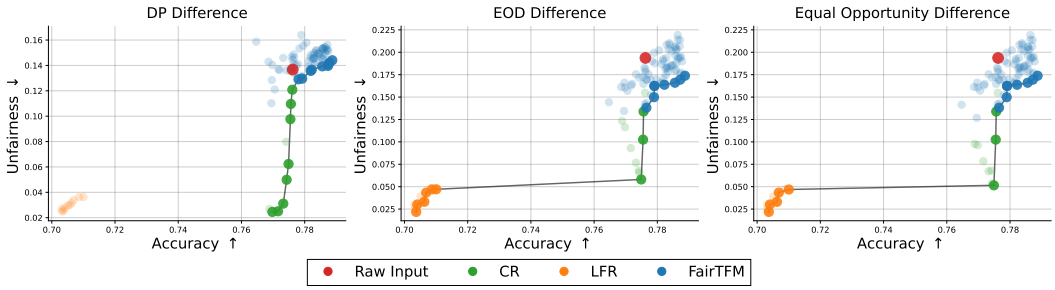


(c) Age

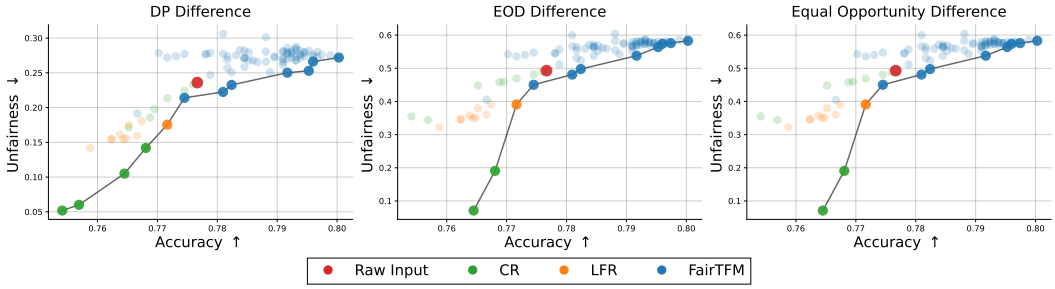
Figure 9: Pareto front of Random Forest models trained with different data representations.



(a) Gender

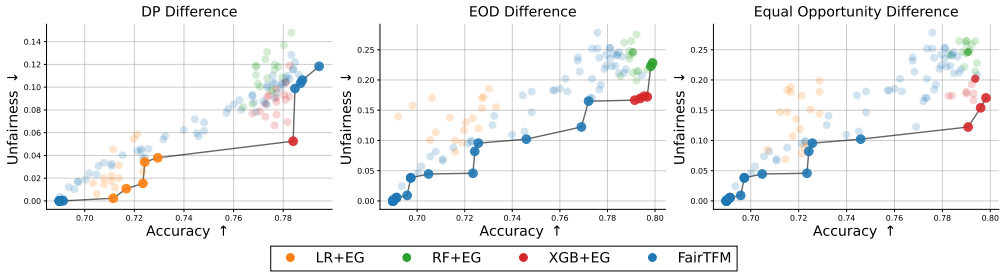


(b) Race

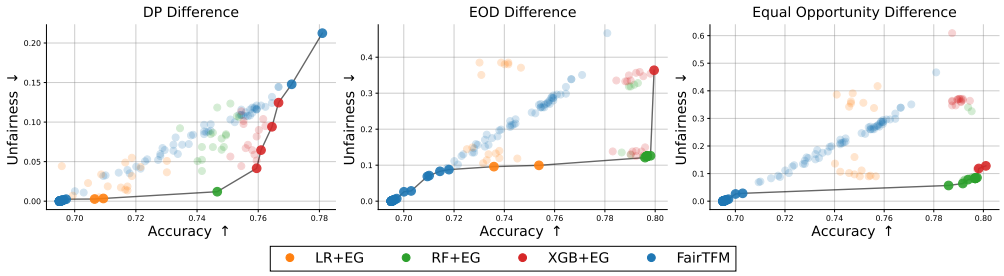


(c) Age

Figure 10: Pareto front of Logistic Regression models trained with different data representations.

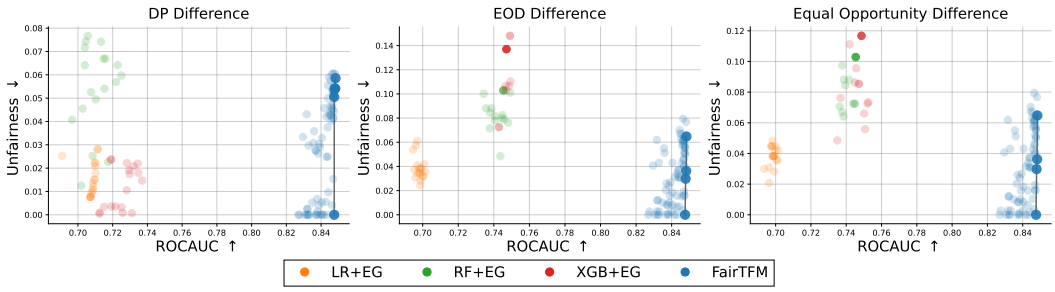


(a) Race

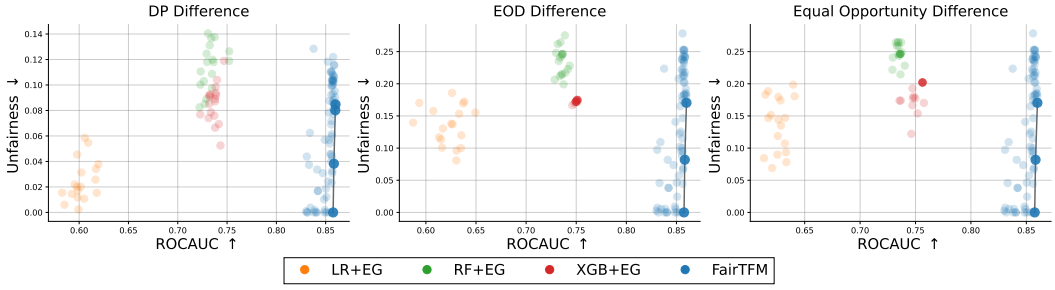


(b) Age

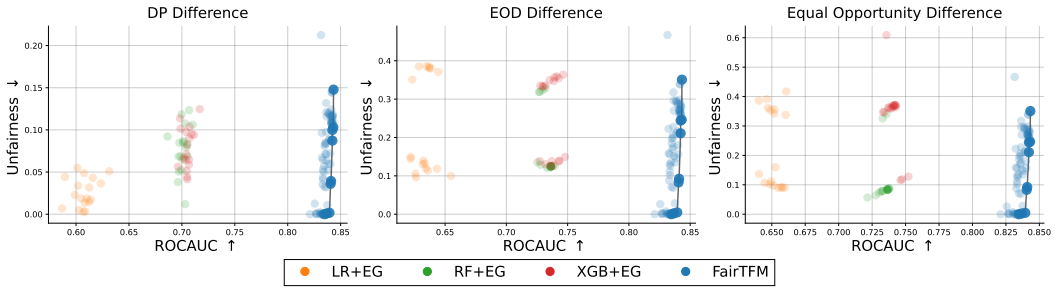
Figure 11: Fairness-accuracy trade-off comparison with task-specific classical models trained with fairness constraints using Exponentiated Gradient (EG), on ACSIncome (Alabama) with race and age as sensitive attributes (gender is shown in Figure 5 of the main text).



(a) Gender



(b) Race



(c) Age

Figure 12: Pareto front of classical models trained with fairness constraints using Exponentiated Gradient (EG). `FairTFM` show best overall pareto front when performance are measured with AUCROC.

E.6 Evaluation beyond ACS PUMS tasks

The main evaluation covers 120 tasks derived from ACS PUMS (Ding et al., 2021). To test whether the observed fairness–utility trends extend beyond this data source, we additionally evaluate on a collection of widely used fairness benchmarks spanning different domains, prediction targets, and sensitive attributes (Le Quy et al., 2022). Specifically, we consider the following datasets:

- The Adult (Census Income) dataset contains demographic and socioeconomic records from the U.S. Census (Becker & Kohavi, 1996). The task is to predict whether an individual's annual income exceeds $50,000, using gender, race, and age as sensitive attributes.
- The COMPAS dataset contains criminal justice screening records for defendants (Angwin et al., 2022). The task is to predict whether an individual will be rearrested within two years of their initial arrest, with race as the sensitive attribute.
- The German Credit dataset contains records of bank account holders and is commonly used for credit risk assessment. The task is to classify applicants as low-risk or high-risk, using gender and age ($\leq$25 years) as sensitive attributes.
- The Diabetes dataset contains clinical records from 130 U.S. hospitals collected between 1999 and 2008 (Strack et al., 2014). The task is to predict whether a patient will be readmitted within 30 days of discharge, using gender and race as sensitive attributes.
- The Law School dataset consists of admissions records collected by the Law School Admission Council (LSAC) from 163 U.S. law schools in 1991 (Wightman, 1998). The task is to predict whether a candidate will pass the bar exam, using race and gender as sensitive attributes.
- The CelebA dataset provides facial-attribute annotations for celebrity images. We consider two binary prediction tasks, *Blond Hair* and *Smiling*, using gender as the sensitive attribute in both cases.

Together, these benchmarks define 12 additional fairness tasks. We use the same evaluation protocol as in Section 4 and average results over three random seeds. Figure 3 in the main text shows the accuracy Pareto fronts, Figure 13 below shows the AUCROC view, and Table 4 reports the averaged metrics, playing the same role as Table 2 for the 120 ACS tasks. As there, the table reports the last pretraining checkpoint of each `FairTFM` variant, so it is a conservative summary: Pareto-dominant checkpoints in the figures could yield better trade-offs, at the cost of model selection on a real-world validation set. The trend from the main evaluation transfers fully to these tasks: `FairTFM`-0.7 reduces DP by 49% and EOD by 47% relative to TabPFNv2.5 while matching XGB and LR in AUCROC (0.803), and `FairTFM`-25 nearly eliminates group disparities (DP of 0.011, an 89% reduction). Notably, KNN—the fairest unconstrained baseline—achieves its lower disparities only through much weaker predictive performance (AUCROC 0.664), whereas `FairTFM` offers strictly better fairness at substantially higher AUCROC. Since `FairTFM` is pretrained on purely synthetic data, all real datasets are unseen during pretraining; these results additionally rule out any ACS-specific effect in the evaluation and show that the fairness behavior learned from synthetic pretraining generalizes across data provenances.

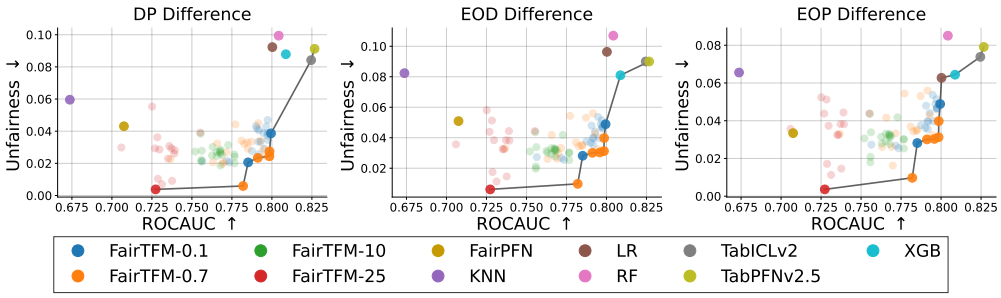


Figure 13: Pareto front between AUCROC and fairness for various models on the 12 fairness tasks beyond ACS PUMS. ↑ indicates higher is better and ↓ indicates smaller is better (unfairness).

As for the 120 ACS tasks, Table 5 reports the relative change of each `FairTFM` variant against TabPFNv2.5, XGB, and LR, derived from Table 4. The picture is even stronger than on the ACS benchmark: `FairTFM`-0.7 improves DP, EOD, and EOP by 39–50% against TabPFNv2.5 for a 5.2% accuracy and 3.0% AUCROC cost, and against XGB and LR it improves fairness by 34–51% with *no*

Table 4: Average fairness and accuracy metrics across the 12 fairness tasks beyond ACS PUMS, reported as mean ± standard deviation over three random seeds. The same pattern as on the 120 ACS tasks holds: unconstrained TFMs are the most accurate but least fair, while `FairTFM` variants sharply reduce disparities at a moderate predictive cost.

Model	**Accuracy**	**AUCROC**	**DP Diff**	**EOD Diff**	**EOP Diff**
TabICLv2	$\mathbf{0.796 \pm 0.09}$	$\mathbf{0.829 \pm 0.09}$	0.101 ± 0.09	0.097 ± 0.06	0.087 ± 0.06
TabPFNv2.5	0.793 ± 0.10	0.828 ± 0.09	0.097 ± 0.09	0.106 ± 0.07	0.092 ± 0.08
KNN	0.732 ± 0.10	0.664 ± 0.14	0.065 ± 0.06	0.078 ± 0.05	0.062 ± 0.06
XGB	0.780 ± 0.10	0.803 ± 0.10	0.100 ± 0.09	0.099 ± 0.07	0.085 ± 0.07
LR	0.784 ± 0.09	0.803 ± 0.10	0.099 ± 0.09	0.115 ± 0.08	0.087 ± 0.07
RF	0.786 ± 0.09	0.806 ± 0.09	0.102 ± 0.09	0.112 ± 0.06	0.084 ± 0.05
FairPFN	0.737 ± 0.10	0.705 ± 0.10	0.045 ± 0.08	0.062 ± 0.09	0.043 ± 0.08
`FairTFM-0.1`	0.753 ± 0.11	0.782 ± 0.13	0.052 ± 0.09	0.060 ± 0.10	0.060 ± 0.10
`FairTFM-0.7`	0.752 ± 0.10	0.803 ± 0.09	0.049 ± 0.08	0.056 ± 0.09	0.056 ± 0.09
`FairTFM-10`	0.734 ± 0.10	0.780 ± 0.10	0.021 ± 0.07	0.025 ± 0.07	0.025 ± 0.07
`FairTFM-25`	0.708 ± 0.14	0.737 ± 0.12	$\mathbf{0.011 \pm 0.10}$	$\mathbf{0.015 \pm 0.10}$	$\mathbf{0.015 \pm 0.10}$

AUCROC cost at all (0.803 for all three models). At the fair end, `FairTFM-25` improves fairness by 82–89% across all three baselines.

Table 5: Fairness improvement (positive numbers mean fairer) and accuracy and AUCROC change of each `FairTFM` variant relative to (a) TabPFNv2.5, (b) XGB, and (c) LR, averaged over the 12 fairness tasks beyond ACS PUMS and derived from Table 4.

(a) vs. TabPFNv2.5

Model	Acc	AUCROC	DP	EOD	EOP
`FairTFM-0.1`	-5.0%	-5.6%	$\mathbf{+46.4\%}$	$\mathbf{+43.4\%}$	$\mathbf{+34.8\%}$
`FairTFM-0.7`	-5.2%	-3.0%	$\mathbf{+49.5\%}$	$\mathbf{+47.2\%}$	$\mathbf{+39.1\%}$
`FairTFM-10`	-7.4%	-5.8%	$\mathbf{+78.4\%}$	$\mathbf{+76.4\%}$	$\mathbf{+72.8\%}$
`FairTFM-25`	-10.7%	-11.0%	$\mathbf{+88.7\%}$	$\mathbf{+85.8\%}$	$\mathbf{+83.7\%}$

(b) vs. XGB

Model	Acc	AUCROC	DP	EOD	EOP
`FairTFM-0.1`	-3.5%	-2.6%	$\mathbf{+48.0\%}$	$\mathbf{+39.4\%}$	$\mathbf{+29.4\%}$
`FairTFM-0.7`	-3.6%	$\pm 0.0\%$	$\mathbf{+51.0\%}$	$\mathbf{+43.4\%}$	$\mathbf{+34.1\%}$
`FairTFM-10`	-5.9%	-2.9%	$\mathbf{+79.0\%}$	$\mathbf{+74.7\%}$	$\mathbf{+70.6\%}$
`FairTFM-25`	-9.2%	-8.2%	$\mathbf{+89.0\%}$	$\mathbf{+84.8\%}$	$\mathbf{+82.4\%}$

(c) vs. LR

Model	Acc	AUCROC	DP	EOD	EOP
`FairTFM-0.1`	-4.0%	-2.6%	$\mathbf{+47.5\%}$	$\mathbf{+47.8\%}$	$\mathbf{+31.0\%}$
`FairTFM-0.7`	-4.1%	$\pm 0.0\%$	$\mathbf{+50.5\%}$	$\mathbf{+51.3\%}$	$\mathbf{+35.6\%}$
`FairTFM-10`	-6.4%	-2.9%	$\mathbf{+78.8\%}$	$\mathbf{+78.3\%}$	$\mathbf{+71.3\%}$
`FairTFM-25`	-9.7%	-8.2%	$\mathbf{+88.9\%}$	$\mathbf{+87.0\%}$	$\mathbf{+82.8\%}$

E.7 Fairness-aware baselines on the 12 non-ACS tasks

We next compare `FairTFM` against task-specific fairness-aware baselines on the same 12 non-ACS tasks. We wrap XGBoost with the Exponentiated Gradient (EG) reduction (Agarwal et al., 2018); we focus on XGBoost because it is among the strongest non-TFM baselines on these tasks (Table 4) and was the strongest EG-wrapped model in the ACS comparison (Figure 5). For each of the three fairness constraints (DP, EOD, EOP), we sweep the fairness-violation tolerance ε over $\{0.01, 0.02, 0.03, 0.05, 0.07, 0.10, 0.20, 0.30, 0.50, 0.70, 0.90\}$, yielding 33 configurations, each retrained per task and averaged over three seeds. Table 7 reports all configurations, and Table 6 summarizes the comparison using, for each constraint type, the best (lowest) value of its targeted metric across ε.

Three observations follow. First, each `FairTFM` variant improves on the fairness metric that each EG variant explicitly optimizes: relative to the best EG value of each targeted metric (DP 0.062, EOD 0.083, EOP 0.070), `FairTFM`-0.7 improves by 21% (DP), 33% (EOD), and 20% (EOP), `FairTFM`-10 by 66%, 70%, and 64%, and `FairTFM`-25 by 82%, 82%, and 79%. Second, this comes at substantially higher predictive performance: the EG randomized ensembles reduce XGBoost's AUCROC from 0.803 to roughly 0.69–0.70 regardless of ε, while `FairTFM`-0.7 keeps AUCROC at 0.803 with only a modest accuracy cost. Third, the comparison highlights a structural advantage: obtaining the EG numbers required retraining $33 \times 12 \times 3$ model configurations, whereas a single pretrained `FairTFM` checkpoint serves every task and constraint in one forward pass. This per-task, per-constraint retraining cost is itself part of the case for fairness-aware pretraining.

Table 6: Summary comparison on the 12 non-ACS tasks. For each EG constraint type, we report the configuration achieving the best (lowest) value of its targeted metric across the ε sweep (targeted metric in bold); unconstrained XGBoost is included as a reference. Each `FairTFM` variant is fairer on every targeted metric at substantially higher AUCROC.

Model	Accuracy	AUCROC	DP Diff	EOD Diff	EOP Diff
XGB (unconstrained)	0.780	0.803	0.100	0.099	0.085
XGB-EG-DP (best ε)	0.777	0.695	**0.062**	0.075	0.060
XGB-EG-EOD (best ε)	0.778	0.702	0.084	**0.083**	0.068
XGB-EG-EOP (best ε)	0.777	0.700	0.085	0.084	**0.070**
`FairTFM`-0.7	0.752	0.803	0.049	0.056	0.056
`FairTFM`-10	0.734	0.780	0.021	0.025	0.025
`FairTFM`-25	0.708	0.737	0.011	0.015	0.015

Table 7: Full Exponentiated Gradient results on the 12 non-ACS tasks: average fairness and accuracy metrics for each fairness constraint (DP, EOD, EOP) and tolerance ε, mean $\pm$ standard deviation over three random seeds.

Model	ε	Accuracy	AUCROC	DP Diff	EOD Diff	EOP Diff
XGB-EG-DP	0.01	0.778 ± 0.10	0.695 ± 0.07	0.068 ± 0.06	0.073 ± 0.03	0.061 ± 0.03
	0.02	0.778 ± 0.10	0.695 ± 0.07	0.064 ± 0.05	0.065 ± 0.03	0.054 ± 0.03
	0.03	0.778 ± 0.10	0.694 ± 0.07	0.064 ± 0.05	0.069 ± 0.03	0.058 ± 0.03
	0.05	0.778 ± 0.10	0.692 ± 0.06	0.067 ± 0.05	0.070 ± 0.03	0.059 ± 0.03
	0.07	0.777 ± 0.10	0.694 ± 0.06	0.062 ± 0.05	0.072 ± 0.03	0.062 ± 0.03
	0.10	0.777 ± 0.10	0.692 ± 0.06	0.068 ± 0.05	0.071 ± 0.02	0.059 ± 0.03
	0.20	0.778 ± 0.10	0.693 ± 0.06	0.067 ± 0.05	0.075 ± 0.04	0.061 ± 0.04
	0.30	0.776 ± 0.10	0.693 ± 0.07	0.068 ± 0.06	0.078 ± 0.03	0.065 ± 0.03
	0.50	0.777 ± 0.10	0.695 ± 0.07	0.062 ± 0.05	0.075 ± 0.04	0.060 ± 0.04
	0.70	0.778 ± 0.10	0.696 ± 0.07	0.072 ± 0.06	0.083 ± 0.04	0.068 ± 0.04
	0.90	0.777 ± 0.10	0.695 ± 0.06	0.066 ± 0.06	0.071 ± 0.04	0.059 ± 0.04
XGB-EG-EOD	0.01	0.778 ± 0.10	0.702 ± 0.07	0.084 ± 0.08	0.084 ± 0.05	0.069 ± 0.05
	0.02	0.778 ± 0.10	0.702 ± 0.07	0.084 ± 0.08	0.083 ± 0.05	0.068 ± 0.05
	0.03	0.778 ± 0.10	0.704 ± 0.07	0.084 ± 0.08	0.086 ± 0.05	0.071 ± 0.05
	0.05	0.778 ± 0.10	0.704 ± 0.07	0.086 ± 0.08	0.090 ± 0.05	0.075 ± 0.05
	0.07	0.778 ± 0.10	0.703 ± 0.07	0.086 ± 0.08	0.088 ± 0.05	0.074 ± 0.05
	0.10	0.778 ± 0.10	0.703 ± 0.07	0.085 ± 0.08	0.084 ± 0.05	0.067 ± 0.05
	0.20	0.778 ± 0.10	0.700 ± 0.07	0.086 ± 0.08	0.087 ± 0.06	0.072 ± 0.06
	0.30	0.779 ± 0.10	0.700 ± 0.07	0.085 ± 0.07	0.084 ± 0.06	0.069 ± 0.05
	0.50	0.777 ± 0.10	0.701 ± 0.07	0.084 ± 0.08	0.089 ± 0.06	0.076 ± 0.06
	0.70	0.777 ± 0.10	0.700 ± 0.07	0.086 ± 0.08	0.087 ± 0.06	0.074 ± 0.06
	0.90	0.778 ± 0.10	0.701 ± 0.07	0.088 ± 0.08	0.088 ± 0.06	0.074 ± 0.06
XGB-EG-EOP	0.01	0.776 ± 0.10	0.698 ± 0.07	0.087 ± 0.08	0.089 ± 0.06	0.074 ± 0.06
	0.02	0.776 ± 0.10	0.698 ± 0.07	0.085 ± 0.08	0.087 ± 0.06	0.072 ± 0.06
	0.03	0.776 ± 0.10	0.699 ± 0.07	0.086 ± 0.08	0.088 ± 0.06	0.072 ± 0.06
	0.05	0.777 ± 0.10	0.699 ± 0.07	0.084 ± 0.08	0.087 ± 0.06	0.073 ± 0.06
	0.07	0.776 ± 0.10	0.699 ± 0.07	0.085 ± 0.08	0.086 ± 0.06	0.073 ± 0.06
	0.10	0.777 ± 0.10	0.699 ± 0.07	0.087 ± 0.08	0.087 ± 0.06	0.074 ± 0.06
	0.20	0.778 ± 0.10	0.699 ± 0.07	0.086 ± 0.08	0.085 ± 0.06	0.071 ± 0.06
	0.30	0.777 ± 0.10	0.700 ± 0.07	0.085 ± 0.09	0.084 ± 0.06	0.070 ± 0.06
	0.50	0.777 ± 0.10	0.700 ± 0.07	0.088 ± 0.08	0.088 ± 0.06	0.075 ± 0.06
	0.70	0.778 ± 0.10	0.699 ± 0.07	0.091 ± 0.09	0.092 ± 0.06	0.077 ± 0.06
	0.90	0.777 ± 0.10	0.701 ± 0.07	0.086 ± 0.08	0.091 ± 0.06	0.080 ± 0.06