Towards Fair In-Context Learning with Tabular Foundation Models

Patrik Kenfack [1 2] **Samira Ebrahimi Kahou** [2 3 4] **Ulrich Aïvodji** [1 2]

Abstract

Tabular foundational models have shown promising in-context learning capabilities on structured data by using training examples as context without further parameter adjustments. This emerging approach positions itself as a competitive alternative to traditional gradient-boosted tree methods. However, while biases in conventional machine learning models are well documented, it remains unclear how these biases manifest in Tabular ICL. The paper investigates the fairness implications of Tabular ICL and explores three preprocessing strategies—correlation removal, group-balanced demonstration selection, and uncertainty-based demonstration selection—to address bias. Comprehensive experiments indicate that uncertainty-based demonstration selection consistently enhances group fairness in the predictions. The source code for reproducing the results of this work can be found at https://github.com/tisl-lab/FairTabICL.git. The longer version of this paper can be found at https://arxiv.org/pdf/2505.09503

1. Introduction

Tabular data, represented in rows and columns, is a data modality widely used for prediction tasks in domains such as finance and healthcare (Asuncion et al., 2007). Tree-based models such as XGboost (Chen et al., 2015) and Gradient-Boosted Trees (Ke et al., 2017) have shown the strongest generalization performance on tabular data. Recently, with the emergence of foundation models, Deep Learning (DL) based models have challenged the dominance of tree-based models (Hollmann et al., 2025). Foundation models are models pretrained on vast datasets, without a specific task in mind, and they can be adapted across a wide range of downstream tasks. Large language models (LLMs) such as GPT-3 are common examples of foundation models, and they have demonstrated emerging capabilities such as in-context learning (ICL) with few labelled data (Brown et al., 2020). In-context learning has mainly been used with LLMs for natural language tasks, where for a given text classification task, the labelled samples are formatted into textual demonstration examples and provided as context information to a language model (Radford et al., 2019). Given the demonstrations as context, the language model can effectively predict the label of a test example without any model update or finetuning (Brown et al., 2020). Attempts have been made to perform ICL using LLMs on tabular data, where the rows of the table are serialized into text or sentences (Hegselmann et al., 2023). However, LLMs are not pretrained to fully capture the complex relations between rows and columns of tabular data, and their performance still lags behind tree-based models in large-scale data (Hegselmann et al., 2023). On the other hand, Hollmann et al. (2022) proposed TabPFN (short for Tabular Prior-Fitted Network), a transformer-based tabular foundation model fully pretrained on synthetic datasets. Follow-up versions of TabPFN—TabPFNv2 (Hollmann et al., 2025) and TabICL (Qu et al., 2025)— demonstrated competitive or better performance compared to the traditional tree-based model on a variety of tasks under ICL (Hollmann et al., 2025; Qu et al., 2025).

With their state-of-the-art performance and ICL capabilities, transformer-based tabular foundation models will likely be widely adopted for decision-making and trigger a shift in the learning paradigm. However, using ICL-based models in high-stakes decision-making scenarios requires a thorough investigation of the negative social impact they might have. It has been shown that traditional machine-learning models can perpetuate bias in the data (Mehrabi et al., 2022). Recent studies have demonstrated that ICL can also provide biased predictions (Hu et al., 2024; Bhaila et al., 2024). However, these works use serialized tabular datasets and inherit the drawbacks of foundational language models on tabular data (Hollmann et al., 2025).

This paper investigates the fairness of ICL prediction using transformer-based tabular foundation models. First, our study reveals, perhaps unsurprisingly, that while these models focus on improving prediction accuracy, they can also

[*]Equal contribution [1]ÉTS Montréal [2]Mila - Quebec AI Institute [3]University of Calgary [4]CIFAR. Correspondence to: Patrik Kenfack <patrik-joslin.kenfack.1@ens.etsmtl.ca>.

Proceedings of the 1st ICML Workshop on Foundation Models for Structured Data, Vancouver, Canada. 2025. Copyright 2025 by the author(s).

amplify bias. Motivated by recent studies on the sensitivity of ICL performance—in terms of fairness and accuracy—to demonstration selection, we aim to address the following research question: *What in-context selection/transformation method can improve the fairness of ICL predictions?* In the fairness literature, fairness-enhancing methods are generally grouped into three categories: pre-processing, in-processing, and post-processing (Mehrabi et al., 2022). Projecting these categories into the ICL paradigm, pre-processing methods perform demonstration transformation or selection before predicting in context (Hu et al., 2024). In-processing methods would fine-tune or retrain the foundation model with fairness constraints (Robertson et al., 2024). Post-processing methods would alter the ICL predictions to improve a given fairness metric (Hardt et al., 2016). Pre- and Post-processing methods are more computationally friendly since they do not require model updates. This motivates our choice to focus on the pre-processing techniques and leave post-processing interventions for future exploration. More specifically, we propose and investigate three pre-processing fairness interventions: (i) Correlation Remover (Feldman et al., 2015), a method that alters each input feature to reduce their correlation with the sensitive attribute; (ii) group-balanced[1] in-context selection, ensures that the in-context set is group-balanced; (iii) Uncertainty-based in-context selection, estimates the uncertainty of predicting the sensitive attribute of in-context samples and only selects samples with uncertain predictions. We performed intensive experiments on eight fairness benchmark datasets to investigate the effectiveness of each method in terms of fairness and accuracy. Our results reveal that the uncertainty-based method can provide better fairness performance across datasets, fairness metrics, and foundational models, with marginal impact on accuracy. Our contribution can be summarized as follows:

- While most existing studies focus on fair ICL with serialized tabular data, we provide, to our knowledge, the first investigation into preprocessing methods for fair prediction in ICL using transformer-based tabular foundation models.
- We propose and investigate three pre-processing intervention methods to enforce fair ICL predictions. These methods aim to reduce the information about the sensitive attributes of in-context samples. We demonstrate that uncertainty-based in-context sample selection can significantly improve the fairness of ICL predictions with a slight drop in accuracy.
- We perform extensive experiments on a broad range of start-of-the-art fairness benchmarks and provide insights into contexts where a given fairness intervention performs best in terms of fairness accuracy tradeoff.
- We release the code to ease reproduction of the results and help researchers and practitioners integrate the proposed methods.

[1]Underlined represent the method's name throughout the paper and in the results.

2. Methodology

2.1. Problem Setup

We consider a classification task where giving the training data $\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^{N}$ where x_i is an input feature vector, y_i is the corresponding class label, and s_i the corresponding demographic group. The goal is to learn a classifier f to accurately predict the target y given a sample x while being *fair* w.r.t. demographic information s (e.g., gender and race). This work focuses on group fairness metrics, which measure performance disparity across demographic groups (See Appendix C.1 for more details).

2.2. Preprocessing Fair-ICL interventions

Pre-processing fairness interventions often involve applying a transformation to the training dataset to reduce the influence of the sensitive attribute. The intervention is model-agnostic as the intervened data is fitted to any downstream model with the hope of better fairness performance. We propose three interventions on in-context samples to improve the fairness of ICL predictions.

`Correlation Remover` In-context transformation with correlation remover (Feldman et al., 2015) where each non-sensitive feature is transformed to reduce correlation with the sensitive ones. We use the fairlearn toolkit (Bird et al., 2020) implementation and fixed the parameter α controlling the fairness-accuracy tradeoff to one, meaning maximal fairness.

Balanced Group-balanced in-context sample selection where context examples are randomly drawn with equal demographic group ratio. The majority group is uniformly downsampled across k-fold evaluations and independent runs.

`Uncertain` Using the uncertainty of the sensitive attribute prediction to select in-context examples. We use the Mapie (Taquet et al., 2022) implementation of conformal prediction to estimate the uncertainty of the training data. We fix the coverage parameter ϵ, of conformal prediction to 0.05 and only select as in-context example samples with prediction sets equal to two, i.e., samples with uncertain sensitive attributes. We consider two variants of the method under different model classes used to train the sensitive attribute classifier for uncertainty estimation: a variant that uses the traditional logistic regression model (Uncertain+LR) and a variant that uses a foundation model (Uncertain+TabPFN).

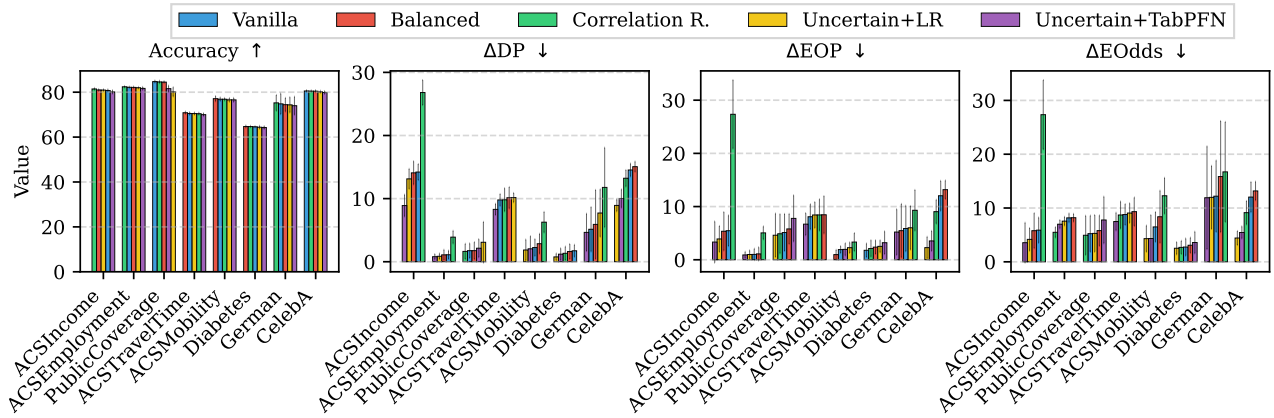


Figure 1. **Fairness-accuracy performance of different ICL methods**. ↑ indicates higher is better (accuracy) and ↓ lower is better (unfairness). Uncertainty-based instance selection tends to provide better fairness while preserving accuracy. Table 9 in the Appendix supplements the figure with the actual values.

3. Experiments

This section describes the experimental setup and provides the results and discussion.

3.1. Experimental Setup

Datasets We experiment on tasks from the recently proposed folktables (Ding et al., 2022), which contains data extracted from the American Community Survey (ACS). More details about each dataset, including the sensitive attributes used for fairness evaluation, the number of samples, and the number of features, can be found in the Appendix B.

Metrics In addition to the classic accuracy, we consider three popular group fairness metrics, i.e., Demographic Parity (ΔDP), Equal Opportunity (ΔEOP), and Equalized Odds (ΔEOD), more details can be found in C.1.

Evaluation For evaluation, we split each dataset into 80%-20%, where the 20% is used for training the sensitive attribute classifier for uncertainty quantification. For the remaining 80% we use 5-fold cross-validation with random shuffle across ten independent runs. This ensures our evaluation is robust and reliable since every data point is used in the in-example or query example across k-fold evaluation. We report the average and standard deviation of fairness and accuracy performance across the k-fold test sets and the ten random independent seeds. As aforementioned, we use TabPFN and TabICL as foundation models for ICL prediction.

Baselines In addition to fairness pre-processing methods presented in section C.3, we consider as a baseline for comparison the *vanilla* method, which performs ICL using randomly selected in-context examples data without any fairness consideration.

3.2. Results

We evaluate several aspects of fairness in ICL prediction on various tabular datasets. First, we compare the different baselines considered in terms of fairness and accuracy. Then, for the methods with controllable tradeoff between fairness and accuracy, we vary the hyperparameter controlling the fairness-accuracy tradeoff and compare their Pareto fronts.

3.2.1. Baseline comparison

Figure 2 summarizes the accuracy & fairness of the different baselines considered across the eight datasets with TabPFN as the foundational model. The results show that uncertainty-based methods significantly improve the fairness of the ICL prediction compared to all baselines.

Surprisingly, `Correlation Remover` exacerbates unfairness in most datasets. For example, on the `ACSIncome` dataset, demographic parity increases from 0.14 to 0.26. To further investigate the failure case of `Correlation Remover`, we measure and compare the reconstruction of sensitive attributes by the foundation model before and after applying the preprocessing intervention. Specifically, we perform ICL using the sensitive attribute as the target variable and use the accuracy of the ICL prediction of the sensitive attribute in the test set as a measure of how the foundation model can reconstruct the sensitive attribute after the fairness intervention. We observe that TabPFN and TabICL can fully reconstruct (up to 100% accuracy) the sensitive attribute based on the feature transformation ap-

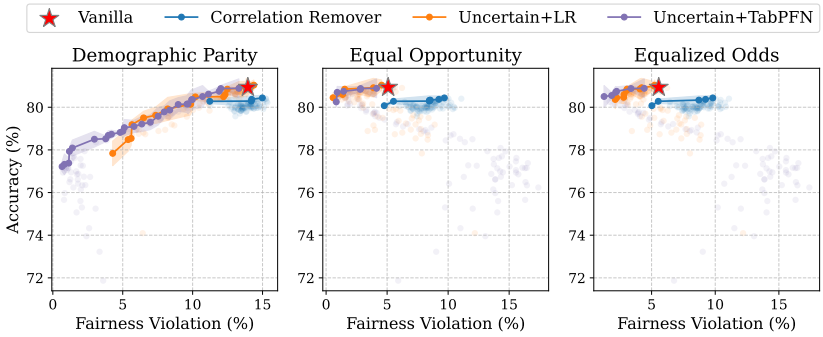


Figure 2. **Accuracy and fairness performance for ICL prediction with TabPFN as foundation model**. Comparing the fairness-accuracy Pareto-front of different fairness interventions using TabPFN. Results with other datasets can be found in the Appendix (Figure 5) and results using TabICL in Figure 7.

plied by `Correlation Remover` (see Table 3 in Appendix). This demonstrates that transformer-based models like TabPFN, pretrained on synthetic data with different sample and feature interactions, can uncover the transformation applied to the non-sensitive features at test-time. By transforming every non-sensitive feature based on the sensitive one, `Correlation Remover` introduces a sensitive information leakage that the foundational model uncovers and relies upon for predicting the target variable, thereby exacerbating unfairness. Further discussion and results can be found in Appendix D, where we also demonstrate that applying the feature transformation only to the training set and leaving the test set unchanged yields better fairness outcomes compared to the `Vanilla` approach. On the other hand, Table 3 in the Appendix shows that `Uncertain` methods yield the smallest accuracy of ICL predictions of the sensitive attribute across datasets. This indicates that the selected in-context samples do not encode sufficient information about the sensitive attribute that the foundation model can rely upon for inference, thereby reducing unfairness.

3.2.2. Fairness Accuracy Tradeoff

In the previous experiment, we fixed the parameters controlling the tradeoff between fairness and accuracy for the `Correlation Remover` and the `Uncertain` methods, α and ϵ respectively. For a better comparison of the fairness accuracy performances, we consider a range of values between [0, 1] for parameters controlling the tradeoff and plot the Pareto front. Figure 2 shows the Pareto front of the two variants of the `Uncertain`, `Correlation Remover`, and the `Vanilla` method. Each scatter point in the figure represents a different value of α and ϵ for the corresponding ICL method. As can be seen, both variants of the `Uncertain` consistently have better Pareto dominant points while `Correlation Remover` can perform worse than the `Vanilla` in terms of fairness, even with different values of α. This further provides evidence of bias amplification of `Correlation Remover` when used with foundation models. We also see that the `Uncertain` method can consistently control the fairness accuracy tradeoff with higher values of ϵ enforcing better fairness at the expense of accuracy. The slight decrease in accuracy when the `Uncertain` is applied is mainly due to the reduced size of in-context examples since more and more samples with certain sensitive attributes are not included in the in-context set. Comparing the fairness-accuracy tradeoffs of TabICL and TabPFN, Figure 9 and 10 in the Appendix show that both foundation models tend to have similar fairness performances while TabPFN often provides better accuracy across datasets.

4. Conclusion and future works

In this study, we proposed and analyzed the effectiveness of three preprocessing methods to enhance the fairness of in-context learning (ICL) predictions. Our empirical results, performed on eight fairness benchmarks, posit the uncertainty-based in-context selection method as a strong baseline for improving the fairness of tabular ICL. The key advantages of this method are threefold: (1) it does not require fine-tuning or retraining the foundation model to enforce the desired fairness metrics; (2) it can consistently improve three widely used group fairness metrics; (3) it offers a parameter to control the fairness-accuracy tradeoff. To our knowledge, this is the first work that explores pre-processing fairness intervention on tabular foundation models. We hope this work will trigger more investigations into fair tabular ICL, since in-context learning as a new learning paradigm will be increasingly adopted into decision-making tools. Interesting future research directions include investigating in-processing and post-processing methods and analyzing the effect of distribution shift between in-context and test examples on fairness and accuracy.

Ethics Statement

This paper explores ways to reduce unfairness in tabular foundation models, emphasizing fair treatment for various groups. We recognize the significance of fairness in machine learning, especially regarding sensitive attributes like race, gender, and socio-economic status. Our research seeks to uncover and tackle potential biases in these models, thereby enhancing transparency, accountability, and inclusivity. While the proposed method uses a sensitive attributes predictor, which could be unlawful in some countries, we emphasized that predicted sensitive values are not used either for training or measuring unfairness. We use the attribute classifier only to quantify uncertainty, and emphasize that this method should not be used for any purpose other than bias measuring or mitigation.

References

Aïvodji, U., Bidet, F., Gambs, S., Ngueveu, R. C., and Tapp, A. Local data debiasing for fairness based on generative adversarial training. *Algorithms*, 14(3):87, 2021.

Angelopoulos, A. N., Bates, S., et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

Asuncion, A., Newman, D., et al. Uci machine learning repository, 2007.

Bhaila, K., Van, M.-H., Edemacu, K., Zhao, C., Chen, F., and Wu, X. Fair in-context learning via latent concept variables. *arXiv preprint arXiv:2411.02671*, 2024.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Celis, L. E., Keswani, V., and Vishnoi, N. K. Data preprocessing to mitigate bias: A maximum entropy based approach. pp. 11.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring Adult: New Datasets for Fair Machine Learning, January 2022. URL http://arxiv.org/abs/2108.04884. arXiv:2108.04884 [cs, stat].

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*, pp. 214–226, Cambridge, Massachusetts, 2012. ACM Press. ISBN 978-1-4503-1115-1. doi: 10.1145/2090236.2090255. URL http://dl.acm.org/citation.cfm?doid=2090236.2090255.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Frank, A. Uci machine learning repository. *http://archive. ics. uci. edu/ml*, 2010.

Gardner, J., Popovic, Z., and Schmidt, L. Benchmarking distribution shift in tabular data with tableshift. *Advances in Neural Information Processing Systems*, 2023.

Hardt, M., Price, E., and Srebro, N. Equality of Opportunity in Supervised Learning, October 2016. URL http://arxiv.org/abs/1610.02413. Number: arXiv:1610.02413 arXiv:1610.02413 [cs].

Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., and Sontag, D. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pp. 5549–5581. PMLR, 2023.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., and Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Hu, J., Liu, W., and Du, M. Strategic demonstration selection for improved fairness in llm in-context learning. *arXiv preprint arXiv:2408.09757*, 2024.

Kamiran, F. and Calders, T. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33 (1):1–33, October 2012. ISSN 0219-1377, 0219-3116. doi: 10.1007/s10115-011-0463-8. URL http://link.springer.com/10.1007/s10115-011-0463-8.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Kenfack, P. J., Kahou, S. E., and Aïvodji, U. Fairness under demographic scarce regime. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=TB18G0w6Ld.

Liu, Z., Luo, P., Wang, X., and Tang, X. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15 (2018):11, 2018.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A Survey on Bias and Fairness in Machine Learning, January 2022. URL http://arxiv.org/abs/1908.09635. Number: arXiv:1908.09635 arXiv:1908.09635 [cs].

Qu, J., Holzmüller, D., Varoquaux, G., and Morvan, M. L. Tabicl: A tabular foundation model for in-context learning on large data. *arXiv preprint arXiv:2502.05564*, 2025.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Robertson, J., Hollmann, N., Awad, N., and Hutter, F. Fairpfn: Transformers can do counterfactual fairness. *arXiv preprint arXiv:2407.05732*, 2024.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Taquet, V., Blot, V., Morzadec, T., Lacombe, L., and Brunel, N. Mapie: an open-source library for distribution-free uncertainty quantification. *arXiv preprint arXiv:2207.12274*, 2022.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Appendix

A. Methodology

This section presents three pre-processing techniques proposed in this work to ensure fairer outcomes in in-context learning on tabular data. In particular, we consider *correlation remover* (Feldman et al., 2015), group-balanced demonstration selection, and uncertainty-based demonstration selection.

A.1. In-context Samples Transformation

Correlation remover (Feldman et al., 2015; Bird et al., 2020) is a preprocessing method that reduces the correlation between the sensitive and non-sensitive attributes before fitting the model. More specifically, each non-sensitive feature vector is *transform* to reduce its correlation with the sensitive feature. We apply the correlation remover as a preprocessing step over the entire demonstration set before performing in-context prediction. Ultimately, transforming input features to reduce their linear dependency on the sensitive feature can reduce the reliance on sensitive features in the downstream models. However, nonlinear and complex downstream models can infer the nonlinear dependencies over the sensitive feature and provide unfair results.

A.2. In-context Samples Selection

In this work, we posit that in-context sample selection can have a significant impact on the fairness of ICL prediction. We analyze several demonstration selection methods that can improve the fairness of ICL predictions without any fairness finetuning.

Group-balanced demonstration set selection. Representation bias is a common source of bias in machine learning models (Mehrabi et al., 2022). It occurs when the collected training data does not reflect the demographic diversity of the population. As a result, some demographic subgroups are under-represented, if not represented at all. Recent studies have demonstrated the benefits of group-balanced training data on the fairness properties of the downstream model. Several methods have been proposed to mitigate representation bias in the data, including *subsampling the majority group* or *reweighting the training data* based on group proportions (Kamiran & Calders, 2012; Celis et al.). In this paper, we focus on *subsampling* since current tabular foundation models do not handle sample weights (Hollmann et al., 2025; Qu et al., 2025). Specifically, we perform ICL with a group-balanced demonstration set sampling from each group uniformly at random. When the demonstration set size does allow equal group representation, we subsample the majority group at random. A similar strategy is employed by (Hu et al., 2024) to select demonstrations for few-shot ICL prediction with LLMs. In this paper, we focus on fair ICL with tabular foundation models instead of using LLMs on serialized tabular data.

Uncertainty-based demonstration set selection Kenfack et al. (2024) demonstrated that models trained without fairness constraints can have better fairness properties when the training data consists of samples with uncertain sensitive attributes. Building on this, we hypothesize that *the uncertainty of the sensitive attribute prediction can be a good measure to select demonstrations that improve the fairness in-context predictions*. To validate this, we measure the uncertainty of predicting the sensitive attribute in the demonstration examples set and use samples with high uncertainty for in-context learning. We focus on conformal prediction (Shafer & Vovk, 2008; Vovk et al., 2005) as uncertainty measure since it provides strong theoretical guarantees for the coverage. Instead of returning a single label, a conformal predictor returns a prediction set containing the true label with a probability of at least $1 - \epsilon$, with ϵ being a user-defined coverage parameter of the conformal prediction (Angelopoulos et al., 2023). For example, setting $\epsilon = 0.1$ ensures the prediction set contains the true sensitive attribute value with at least 90% probability. Specifically, we consider samples with prediction set sizes containing more than one value as uncertain. Intuitively, the coverage parameter ϵ controls the fairness-accuracy tradeoff, with $\epsilon \approx 1$ meaning no fairness intervention where all the datapoints are used and $\epsilon \approx 0$ meaning maximal fairness intervention where only uncertain samples are included in the in-context examples.

Since conformal prediction is model agnostic, we considered both classical methods, e.g., Logistic Regression (LR), and foundation models, e.g., TabPFN for training the sensitive attribute classifier to measure the prediction uncertainties. Note that the method could be applied using other uncertainty measures, such as Monte Carlo dropout and confidence interval (Kenfack et al., 2024). We focus on conformal prediction due to its rigorous theoretical guarantees, and it does not require a hyperparameter to threshold the level from which a prediction is considered uncertain (Angelopoulos et al., 2023).

A.3. In-context prediction

After performing demonstration selection or transformation using the fairness intervention methods presented previously, we pass them through the tabular foundation model as in-context examples for predicting class labels on the test set. In this paper, we consider TabICL (Hollmann et al., 2025) and TabPFN (Qu et al., 2025) as tabular foundation models. These models are transformers trained on extensive synthetic datasets to perform tabular predictions with a single forward pass without parameter update.

B. Datasets

Table 1. Summary of datasets used in our experiments. For each dataset, we report the number of features (including the sensitive attribute), the number of samples available, and the sensitive attribute used for fairness evaluation.

Dataset	# Features	# Samples	Sensitive Feature	Prediction Task
ACSIncome	10	22,268	Gender	Income $\geq$ \$50,000
ACSEmployment	16	47,777	Gender	Employment status
ACSTravelTime	16	19,492	Gender	Commute time over 20 minutes
ACSMobility	21	8,625	Gender	Residential mobility
PublicCoverage	19	18,525	Gender	Public health insurance coverage
CelebA	39	202,599	Gender	Attractiveness
Diabetes	183	38,575	Race	Prior diabetes diagnose
German	58	990	Age	Credit risk

We experiment on tasks from the recently proposed folktables (Ding et al., 2022), which contains data extracted from the American Community Survey (ACS) Public Use Microdata Sample (PUMS) (Ding et al., 2022). More specifically, we experiment with the following ACS PUMS tasks:

- **ACSIncome**: The task involves predicting whether an individual's income exceeds \$50,000. The dataset is filtered to include only individuals over the age of 16 who reported working at least 1 hour per week during the past year and earned a minimum of \$100.
- **ACSMobility**: This task involves predicting whether an individual had the same residential address one year ago. The dataset is filtered to include individuals aged between 18 and 35. This filtering increases the difficulty of the task, as more than 90% of the general population tends to stay at the same address year-to-year.
- **ACSTravelTime**: This task predicts whether an individual has a commute longer than 20 minutes. The dataset is filtered to include only employed individuals above the age of 16. The 20-minute threshold corresponds to the median commute time in the US, according to the 2018 ACS PUMS data.
- **ACSEmployment**: The objective is to predict whether an individual is employed, using a dataset filtered to include individuals aged between 16 and 90.
- **ACSPublicCoverage**: The goal is to predict whether an individual has public health insurance. The dataset is filtered to include individuals under 65 years of age and those with an income below \$30,000, focusing on low-income individuals who are ineligible for Medicare.

These tasks were selected to reflect a range of real-world predictive challenges with fairness concerns. We use the data of the year 2018 from the state of Alabama (AL), which is one of the states with the largest fairness violation (Ding et al., 2022)[2].

[2]We also perform experiments on data from other states, and observed that the results presented in the paper remain consistent.

A limitation of the ACS PUMS datasets is that they are US-centric; we diversify the experimental setup by including other tasks and datasets. Specifically, we also experiment on the following tabular datasets and tasks:

- **Diabetes** (Gardner et al., 2023): The diabetes prediction task uses features related to physical health, lifestyle factors, and chronic conditions, derived from the BRFSS questionnaires. Demographic attributes like race, sex, state, and income are also included. The target is a binary indicator of whether the respondent has ever been diagnosed with diabetes.
- **German Credit** (Frank, 2010): The German Credit dataset contains 20 attributes of 1,000 individuals. We create the task of classifying people according to whether they have a good or bad credit risk using age (over or below 25 years old) as the sensitive attribute.
- **CelebA** (Liu et al., 2018): The dataset contains 202,599 samples described with 40 facial attributes of human annotated images, we create the task of predicting *attractiveness* with facial attributes using gender as the sensitive attribute (Kenfack et al., 2024). Note that we do not train the model with images and consider this task to diversify the experimental tasks.

C. Background

C.1. Fairness Metrics

In this work, we focus on group fairness notions measuring the performance disparity across different demographic groups. More specifically, we consider the following three widely used group fairness metrics:

- **Demographic parity (DP)**: DP enforces equal positive outcome rate for different groups (Dwork et al., 2012) and is defined as follows:

$$P(f(X) = 1|S = s) = P(f(X) = 1) \tag{1}$$

- **Equalized Odds (EOD)**: EOdds is satisfied when the model makes correct and incorrect predictions at the same rate for different demographic groups (Hardt et al., 2016). The metric enforces equal true positive and false positive rates across groups and is measured as follows;

$$P(f(X) = 1|S = 0, Y = y) = P(f(X) = 1|S = 1, Y = y),\ \forall y \in \{0, 1\} \tag{2}$$

- **Equalized Opportunity (EOP)**: In some settings, one can care more about assessing unfairness when the model makes correct predictions. EOP enforces equal true positive rates across groups, i.e., we only consider $y = 1$ in Eq. 2, i.e.,

$$P(f(X) = 1|S = 0, Y = 1) = P(f(X) = 1|S = 1, Y = 1) \tag{3}$$

Empirically, we measure each fairness considered, i.e., Demographic Parity (ΔDP), Equal Opportunity (ΔEOP), and Equalized Odds (ΔEOD) as follows.

$$\Delta\text{DP} = \left| \mathop{\mathbb{E}}_{x|A=0}[\mathbb{I}\{f(x) = 1\}] - \mathop{\mathbb{E}}_{x|A=1}[\mathbb{I}\{f(x) = 1\}] \right| \tag{4}$$

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

$$\Delta\text{EOD} = \alpha_0 + \alpha_1 \tag{5}$$

$$\Delta\text{EOP} = \alpha_1 \tag{6}$$

Where α_0 and α_1 measure the difference between the false positive and the true positive rates across groups, respectively, and are empirically measured as follows.

Where α_0 and α_1 measure the difference between the false positive and the true positive rates across groups, respectively, and are empirically measured as follows.

$$\alpha_j = \left| \mathop{\mathbb{E}}_{x|A=0,Y=j}[\mathbb{I}\{f(x)=1\}] - \mathop{\mathbb{E}}_{x|A=1,Y=j}[\mathbb{I}\{f(x)=1\}] \right| \quad j \in \{0,1\} \tag{7}$$

C.2. Correlation Remover

The `Correlation Remover` (Feldman et al., 2015) is a preprocessing technique designed to eliminate linear correlations between sensitive attributes and non-sensitive features in a dataset. This method is particularly useful in mitigating biases that may arise due to such correlations, especially when employing linear models.

Considering a classification task with the given training data $\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^n$ where x_i is an input feature vector, y_i is the corresponding class label, and s_i the corresponding demographic group.

To apply `Correlation Remover`, we assume the training data is formulated as follows:

- $\mathbf{X} \in \mathbb{R}^{n \times d}$ represents the training data matrix containing sensitive and non-sensitive features.
- $\mathbf{S} \in \mathbb{R}^{n \times m_s}$ a matrix of the sensitive features. For simplicity, we assumed in this work $m_s = 1$, which corresponds to a single binary sensitive attribute.
- $\mathbf{Z} \in \mathbb{R}^{n \times m_z}$ a matrix of non-senstive features such that $X = [S\ Z]$

The goal of `Correlation Remover` is to transform Z into Z^* such that Z^* is uncorrelated with S, while retaining as much information from the original Z as possible.

For each non-sensitive feature vector $\mathbf{z}^j \in \mathbb{R}^n$ (the j-th column of $\mathbf{Z}$), the algorithm solves the following least squares problem:

$$\min_{\mathbf{w}_j} \left\| \mathbf{z}^j - (\mathbf{S} - \mathbf{1}_n \bar{\mathbf{s}}^\top)\mathbf{w}_j \right\|_2^2 \tag{8}$$

where:

- $\bar{\mathbf{s}} = [\bar{\mathbf{s}}_1, \bar{\mathbf{s}}_2, \ldots, \bar{\mathbf{s}}_{m_s}]$ is the mean vector of the sensitive features, i.e., $\bar{\mathbf{s}}_j$ is the mean of j-th the sensitive feature.
- $\mathbf{1}_n$ is an n-dimensional column vector of ones.
- $\mathbf{w}_j \in \mathbb{R}^{m_z}$ is the weight vector that projects the centered sensitive features onto $\mathbf{z}_j$.

After computing the optimal weight vectors $\mathbf{w}_j^*$ for all $j \in \{1, \ldots, m_z\}$, they are assembled into a weight matrix $\mathbf{W}^* = [\mathbf{w}_1^*, \ldots, \mathbf{w}_{m_z}^*]$. The transformed non-sensitive features are then obtained by:

$$\mathbf{Z}^* = \mathbf{Z} - (\mathbf{S} - \mathbf{1}_n \bar{\mathbf{s}}^\top)\mathbf{W}^* \tag{9}$$

This operation effectively removes the linear correlations between $\mathbf{S}$ and $\mathbf{Z}$, resulting in $\mathbf{Z}^*$ that is uncorrelated with the sensitive features.

`Correlation Remover` introduces a tunable parameter $\alpha \in [0, 1]$ that controls the extent of correlation removal, i.e., (i) $\alpha = 1$ corresponds to full removal of linear correlations, thus best possible fairness; (ii) $\alpha = 0$ corresponds no transformation, the original data is used; (iii) $0 < \alpha < 1$ corresponds to partial removal, balancing between the original and transformed data, thus controlling the fairness accuracy tradeoff. More specifically, the final transformed dataset $\mathbf{X}'$ is computed as:

$$\mathbf{X}' = \alpha \mathbf{Z}^* + (1 - \alpha)\mathbf{Z} \tag{10}$$

Note that $\mathbf{X}'$ is derived using $\mathbf{Z}^*$, since $\mathbf{S}$ is dropped after transformation. The convex combination 10 allows practitioners to adjust the fairness accuracy tradeoff based on specific requirements of their application.

Equation 8 is optimized on the training dataset, and the optimal weight vectors w_j^* are used to apply the transformation 10 on the test dataset.

C.3. Tabular Foundation Models

In-context learning with tabular foundation models presents a notable advantage over traditional machine learning approaches by enabling models to adapt dynamically to new data without the need for retraining (Hollmann et al., 2022; Qu et al., 2025; Hollmann et al., 2025). Conventional ML methods typically depend on predefined training datasets, meaning that any alteration in the data or task necessitates a time-consuming and resource-intensive retraining process. In contrast, tabular foundation models utilize in-context learning to execute tasks based on the specific context of the data provided at inference time. This allows these models to interpret and process new tabular data with minimal prior preparation, facilitating more flexible and efficient decision-making (Hollmann et al., 2022). The advantages of this approach are particularly apparent in scenarios where data distributions change over time or when models must quickly adjust to various data tasks without undergoing retraining. Thus, as in-context learning emerges as a powerful tool for real-time, adaptive predictions in complex and dynamic environments, assessing and mitigating biases in the prediction can make its use more socially acceptable. Most studies on fairness in ICL use large language models with tabular datasets serialized into text or sentences (Bhaila et al., 2024). However, these large language models are not trained to handle tabular data, and their performance is generally suboptimal compared to tree-based models. Robertson et al. (2024) proposed FairPFN, a pretrained model similar to PFNs, but trained to suppress the causal effects of the sensitive attributes without requiring a causal model at test time. FairPFN model is pretrained using *fair* synthetic dataset generated from structural causal models (SCM). Fair datasets are obtained by removing outgoing edges from the sensitive attribute (assumed exogenous variable) before sampling synthetic data using the SCM. During the model pre-training, fairness is enforced by predicting the class label of the fair dataset using the biased data as context. In contrast, this work focuses on achieving group fairness instead of counterfactual fairness.

D. On the failure of Correlation Remover

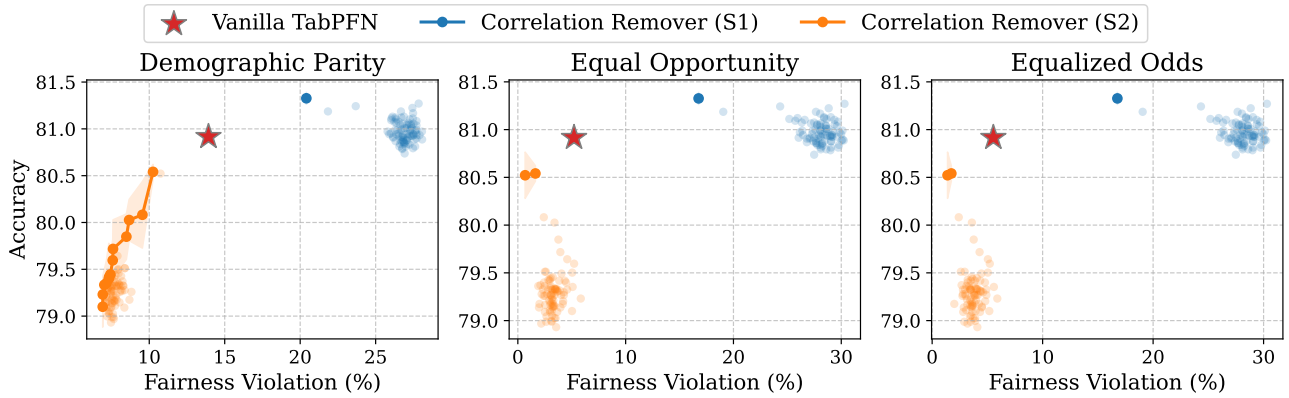
In the main paper, we observed that the transformation 10 on both the training and the testing data can exacerbate unfairness in ICL predictions. We hypothesized that the foundation model inferred the sensitive attribute from the linear transformation applied to each non-sensitive feature, resulting in higher unfairness in the ICL predictions. To validate this hypothesis, we perform ICL prediction of the sensitive attribute after the fairness interventions are applied. As can be seen in Table 3, ICL prediction of the sensitive attribute results in 100% accuracy after `Correlation Remover` is applied. These results suggest the foundation model still relies on the sensitive attribute after correlation removal is performed.

In the main paper, we observed that the transformation 10 on both the training and the testing data can exacerbate unfairness in ICL predictions. We hypothesized that the foundation model inferred the sensitive attribute from the linear transformation applied to each non-sensitive feature, resulting in higher unfairness in the ICL predictions. To validate this hypothesis, we perform ICL prediction of the sensitive attribute after the fairness interventions are applied. As can be seen in Table 3, ICL prediction of the sensitive attribute results in 100% accuracy after `Correlation Remover` is applied. These results suggest the foundation model still relies on the sensitive attribute after correlation removal is performed.

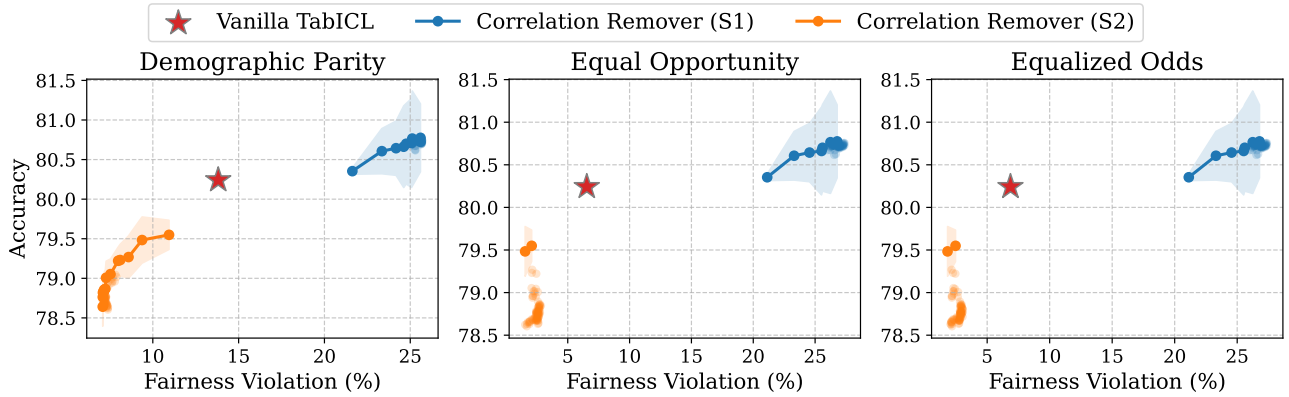
For further verification, we consider a variant of `Correlation Remover` where we apply the feature transformation (Eq. 10) only to the training dataset (S2), leaving the test data unchanged. Table 4 shows this variant (Eq. 10) significantly reduces the accuracy of ICL predictions. This demonstrates that the foundation model uses the transformation applied to the testing set as a proxy to fully reconstruct the sensitive attribute. (Aïvodji et al., 2021) discuss several scenarios where one can apply a data transformation to reduce correlation with the sensitive attribute. More specifically, in addition to the classical setting of correlation remover, where the transformation is applied to the training and testing dataset (S1), we considered another variant where the transformation is applied only to the training set (S2). We evaluated the fairness accuracy tradeoff of applying feature transformation only to the training set (S2). The results in Figure 3 show that the variant S2 of correlation removal can significantly improve fairness compared to applying the transformation to both the training and testing datasets. This further illustrates the ability of the foundation models to reconstruct the sensitive attributes when the transformation is applied to the test set. We recommend practitioners apply the S2 variant of Correlation Remover when used in in-context learning to avoid bias amplification due to sensitive attribute leakage in the test set.

E. Supplementary resutls

Ablation on impact in-context sample size. All the previous experiments used the entire training as in-context examples when possible. Note that the current version of TabPFN cannot handle more than 10000 samples. As an ablation study, we vary the in-context sample size ([100, 300, 500, 700, 1500, 2000, 2500, 3000, 4000, 5000]) and measure fairness and



(a) TabPFN



(b) TabICL

Figure 3. **Evaluating variants of the correlation remover on the `ACSIncome` dataset with TabPFN and TabICL**. Applying correlation remover to the training and testing data exacerbates unfairness, while applying the transformation only to the training set improves fairness.

accuracy using the same experimental setup as in Section 3.2.1. Figure 4 shows that accuracy significantly increases when more samples are added to the in-context set. Unfainness increases slightly before remaining almost constant when the in-context set size exceeds 700. On the other hand, uncertain consistently has the lowest fairness violation across in-context sizes.

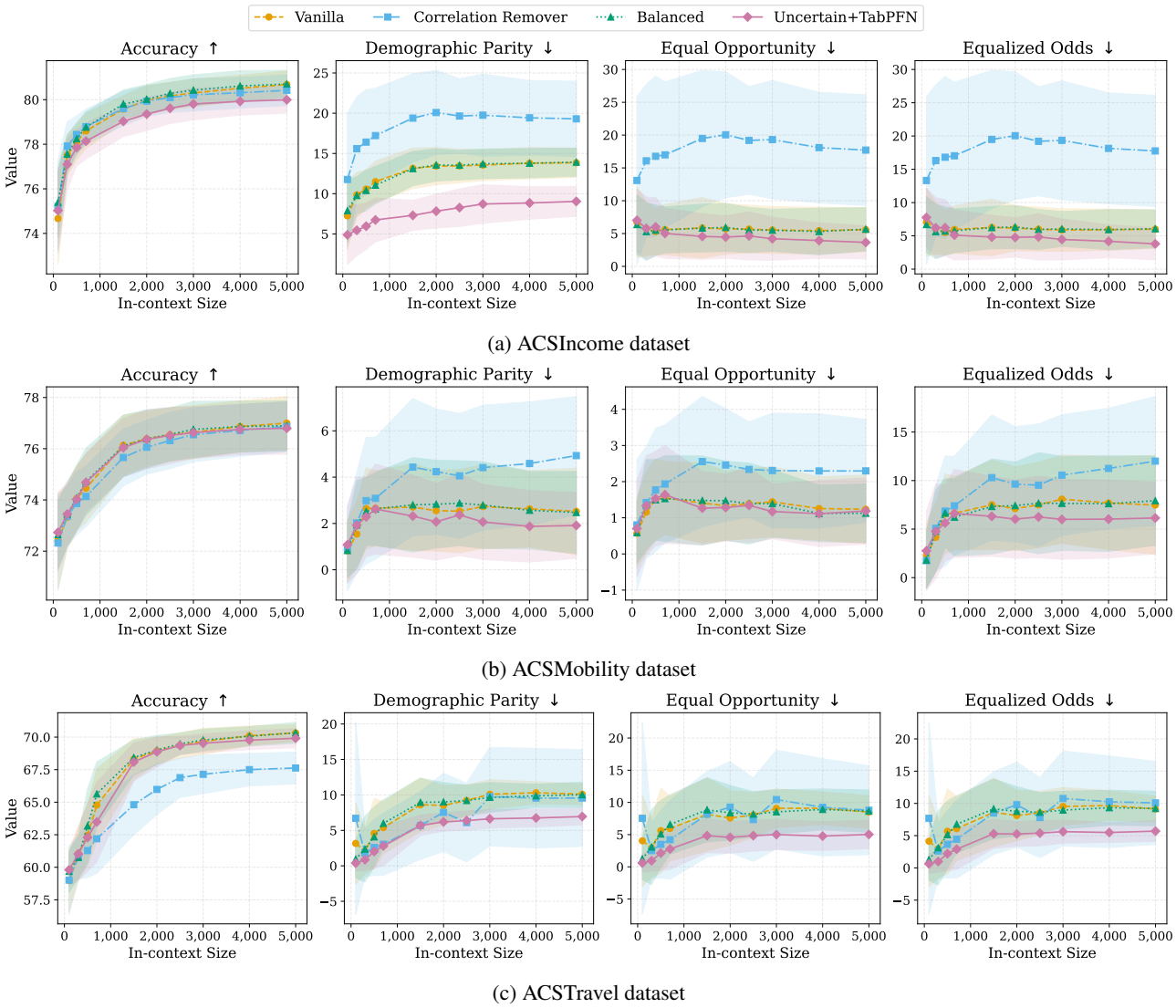


Figure 4. **Ablation on the in-context example set size**. Analyzing the impact of the in-context set size on the fairness and accuracy of ICL prediction.

Dataset	ICL Method	TabPFN		TabICL	
		Accuracy ↓	F1 Score ↓	Accuracy ↓	F1 Score ↓
ACSIncome	Vanilla	$77.2_{\pm 0.5}$	$78.4_{\pm 0.3}$	$75.0_{\pm 0.5}$	$76.2_{\pm 0.3}$
	Balanced	$77.1_{\pm 0.5}$	$77.8_{\pm 0.2}$	$75.0_{\pm 0.4}$	$75.5_{\pm 0.1}$
	Correlation R.	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$99.9_{\pm 0.0}$	$99.9_{\pm 0.0}$
	Uncertain+LR	$74.7_{\pm 1.3}$	$76.7_{\pm 0.6}$	$74.9_{\pm 0.5}$	$76.0_{\pm 0.4}$
	Uncertain+TabPFN	$\mathbf{51.3}_{\pm 2.3}$	$\mathbf{66.1}_{\pm 4.4}$	$\mathbf{71.7}_{\pm 0.4}$	$\mathbf{73.6}_{\pm 0.6}$
ACSTravelTime	Vanilla	$75.9_{\pm 0.4}$	$77.5_{\pm 0.5}$	$72.8_{\pm 0.4}$	$73.8_{\pm 0.5}$
	Balanced	$75.9_{\pm 0.5}$	$77.0_{\pm 0.5}$	$72.5_{\pm 0.6}$	$72.6_{\pm 0.6}$
	Correlation R.	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	Uncertain+LR	$75.9_{\pm 0.6}$	$77.8_{\pm 0.5}$	$72.8_{\pm 0.5}$	$74.3_{\pm 0.5}$
	Uncertain+TabPFN	$\mathbf{74.6}_{\pm 1.1}$	$\mathbf{75.7}_{\pm 1.7}$	$\mathbf{67.4}_{\pm 1.6}$	$\mathbf{66.2}_{\pm 2.4}$
ACSPublicCoverage	Vanilla	$91.4_{\pm 0.2}$	$88.9_{\pm 0.2}$	$91.5_{\pm 0.1}$	$89.2_{\pm 0.2}$
	Balanced	$91.1_{\pm 0.4}$	$89.0_{\pm 0.3}$	$90.9_{\pm 0.3}$	$88.9_{\pm 0.4}$
	Correlation R.	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	Uncertain+LR	$56.4_{\pm 10.0}$	$\mathbf{55.3}_{\pm 4.0}$	$57.7_{\pm 11.9}$	$60.0_{\pm 3.6}$
	Uncertain+TabPFN	$\mathbf{42.5}_{\pm 0.9}$	$58.7_{\pm 0.6}$	$\mathbf{42.7}_{\pm 1.1}$	$\mathbf{58.6}_{\pm 0.6}$
ACSEmployment	Vanilla	$64.0_{\pm 0.4}$	$62.0_{\pm 1.8}$	$65.0_{\pm 0.3}$	$62.2_{\pm 1.4}$
	Balanced	$64.0_{\pm 0.5}$	$65.0_{\pm 1.0}$	$64.8_{\pm 0.4}$	$65.3_{\pm 1.1}$
	Correlation R.	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	Uncertain+LR	$64.3_{\pm 0.4}$	$62.4_{\pm 2.0}$	$64.9_{\pm 0.3}$	$62.8_{\pm 0.9}$
	Uncertain+TabPFN	$\mathbf{57.5}_{\pm 3.3}$	$\mathbf{47.0}_{\pm 8.6}$	$\mathbf{61.1}_{\pm 3.0}$	$\mathbf{53.9}_{\pm 6.1}$
ACSMobility	Vanilla	$68.3_{\pm 0.8}$	$67.8_{\pm 1.0}$	$67.6_{\pm 1.0}$	$67.4_{\pm 1.2}$
	Balanced	$68.1_{\pm 0.7}$	$67.9_{\pm 1.4}$	$67.6_{\pm 1.1}$	$67.6_{\pm 1.5}$
	Correlation R.	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	Uncertain+LR	$68.1_{\pm 0.9}$	$67.8_{\pm 0.9}$	$67.4_{\pm 0.8}$	$66.9_{\pm 0.8}$
	Uncertain+TabPFN	$\mathbf{68.1}_{\pm 0.8}$	$\mathbf{67.7}_{\pm 1.0}$	$\mathbf{67.2}_{\pm 0.8}$	$\mathbf{66.8}_{\pm 0.9}$
German Credit	Vanilla	$72.5_{\pm 3.1}$	$70.3_{\pm 3.1}$	$71.8_{\pm 3.1}$	$71.0_{\pm 3.0}$
	Balanced	$72.7_{\pm 2.3}$	$70.7_{\pm 2.3}$	$71.9_{\pm 3.0}$	$71.2_{\pm 2.6}$
	Correlation R.	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	Uncertain+LR	$64.6_{\pm 5.1}$	$62.3_{\pm 8.5}$	$68.4_{\pm 4.8}$	$67.8_{\pm 5.4}$
	Uncertain+TabPFN	$\mathbf{60.4}_{\pm 5.5}$	$\mathbf{51.3}_{\pm 26.5}$	$\mathbf{63.8}_{\pm 3.9}$	$\mathbf{60.8}_{\pm 10.9}$
Diabetes	Vanilla	$80.2_{\pm 0.1}$	$89.0_{\pm 0.1}$	$80.4_{\pm 0.1}$	$89.0_{\pm 0.1}$
	Balanced	$\mathbf{66.2}_{\pm 0.9}$	$\mathbf{75.9}_{\pm 0.8}$	$\mathbf{65.1}_{\pm 0.4}$	$\mathbf{74.6}_{\pm 0.4}$
	Correlation R.	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	Uncertain+LR	$70.4_{\pm 20.7}$	$77.1_{\pm 26.8}$	$80.3_{\pm 0.1}$	$89.0_{\pm 0.1}$
	Uncertain+TabPFN	$74.9_{\pm 10.0}$	$84.8_{\pm 8.0}$	$80.2_{\pm 0.1}$	$89.0_{\pm 0.1}$
CelebA	Vanilla	$84.7_{\pm 0.2}$	$83.2_{\pm 0.3}$	$85.0_{\pm 0.2}$	$83.2_{\pm 0.3}$
	Balanced	$84.6_{\pm 0.2}$	$83.2_{\pm 0.3}$	$84.9_{\pm 0.2}$	$83.3_{\pm 0.3}$
	Correlation R.	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	Uncertain+LR	$\mathbf{72.4}_{\pm 11.7}$	$\mathbf{61.2}_{\pm 21.7}$	$84.9_{\pm 0.2}$	$83.1_{\pm 0.3}$
	Uncertain+TabPFN	$74.5_{\pm 8.4}$	$70.8_{\pm 10.1}$	$\mathbf{81.2}_{\pm 7.2}$	$\mathbf{77.2}_{\pm 11.9}$

Table 2. **ICL prediction performance of sensitive attributes after applying different fairness interventions**. Smaller accuracy is better since it indicates how well the foundation model can reconstruct the sensitive attribute after the pre-processing fairness interventions. `Uncertain` methods yield the smallest accuracy, which justifies the improved fairness performance.

Dataset	ICL Method	TabPFN		TabICL	
		Accuracy ↓	F1 Score ↓	Accuracy ↓	F1 Score ↓
ACSIncome	Vanilla	$77.2_{\pm 0.5}$	$78.4_{\pm 0.3}$	$75.0_{\pm 0.5}$	$76.2_{\pm 0.3}$
	Balanced	$77.1_{\pm 0.5}$	$77.8_{\pm 0.2}$	$75.0_{\pm 0.4}$	$75.5_{\pm 0.1}$
	Correlation R.	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$99.9_{\pm 0.0}$	$99.9_{\pm 0.0}$
	Uncertain+LR	$74.7_{\pm 1.3}$	$76.7_{\pm 0.6}$	$74.9_{\pm 0.5}$	$76.0_{\pm 0.4}$
	Uncertain+TabPFN	$\mathbf{51.3}_{\pm 2.3}$	$\mathbf{66.1}_{\pm 4.4}$	$\mathbf{71.7}_{\pm 0.4}$	$\mathbf{73.6}_{\pm 0.6}$
ACSTravelTime	Vanilla	$75.9_{\pm 0.4}$	$77.5_{\pm 0.5}$	$72.8_{\pm 0.4}$	$73.8_{\pm 0.5}$
	Balanced	$75.9_{\pm 0.5}$	$77.0_{\pm 0.5}$	$72.5_{\pm 0.6}$	$72.6_{\pm 0.6}$
	Correlation R.	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	Uncertain+LR	$75.9_{\pm 0.6}$	$77.8_{\pm 0.5}$	$72.8_{\pm 0.5}$	$74.3_{\pm 0.5}$
	Uncertain+TabPFN	$\mathbf{74.6}_{\pm 1.1}$	$\mathbf{75.7}_{\pm 1.7}$	$\mathbf{67.4}_{\pm 1.6}$	$\mathbf{66.2}_{\pm 2.4}$
ACSPublicCoverage	Vanilla	$91.4_{\pm 0.2}$	$88.9_{\pm 0.2}$	$91.5_{\pm 0.1}$	$89.2_{\pm 0.2}$
	Balanced	$91.1_{\pm 0.4}$	$89.0_{\pm 0.3}$	$90.9_{\pm 0.3}$	$88.9_{\pm 0.4}$
	Correlation R.	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	Uncertain+LR	$56.4_{\pm 10.0}$	$\mathbf{55.3}_{\pm 4.0}$	$57.7_{\pm 11.9}$	$60.0_{\pm 3.6}$
	Uncertain+TabPFN	$\mathbf{42.5}_{\pm 0.9}$	$58.7_{\pm 0.6}$	$\mathbf{42.7}_{\pm 1.1}$	$\mathbf{58.6}_{\pm 0.6}$
ACSEmployment	Vanilla	$64.0_{\pm 0.4}$	$62.0_{\pm 1.8}$	$65.0_{\pm 0.3}$	$62.2_{\pm 1.4}$
	Balanced	$64.0_{\pm 0.5}$	$65.0_{\pm 1.0}$	$64.8_{\pm 0.4}$	$65.3_{\pm 1.1}$
	Correlation R.	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	Uncertain+LR	$64.3_{\pm 0.4}$	$62.4_{\pm 2.0}$	$64.9_{\pm 0.3}$	$62.8_{\pm 0.9}$
	Uncertain+TabPFN	$\mathbf{57.5}_{\pm 3.3}$	$\mathbf{47.0}_{\pm 8.6}$	$\mathbf{61.1}_{\pm 3.0}$	$\mathbf{53.9}_{\pm 6.1}$
ACSMobility	Vanilla	$68.3_{\pm 0.8}$	$67.8_{\pm 1.0}$	$67.6_{\pm 1.0}$	$67.4_{\pm 1.2}$
	Balanced	$68.1_{\pm 0.7}$	$67.9_{\pm 1.4}$	$67.6_{\pm 1.1}$	$67.6_{\pm 1.5}$
	Correlation R.	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	Uncertain+LR	$68.1_{\pm 0.9}$	$67.8_{\pm 0.9}$	$67.4_{\pm 0.8}$	$66.9_{\pm 0.8}$
	Uncertain+TabPFN	$\mathbf{68.1}_{\pm 0.8}$	$\mathbf{67.7}_{\pm 1.0}$	$\mathbf{67.2}_{\pm 0.8}$	$\mathbf{66.8}_{\pm 0.9}$
German Credit	Vanilla	$72.5_{\pm 3.1}$	$70.3_{\pm 3.1}$	$71.8_{\pm 3.1}$	$71.0_{\pm 3.0}$
	Balanced	$72.7_{\pm 2.3}$	$70.7_{\pm 2.3}$	$71.9_{\pm 3.0}$	$71.2_{\pm 2.6}$
	Correlation R.	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	Uncertain+LR	$64.6_{\pm 5.1}$	$62.3_{\pm 8.5}$	$68.4_{\pm 4.8}$	$67.8_{\pm 5.4}$
	Uncertain+TabPFN	$\mathbf{60.4}_{\pm 5.5}$	$\mathbf{51.3}_{\pm 26.5}$	$\mathbf{63.8}_{\pm 3.9}$	$\mathbf{60.8}_{\pm 10.9}$
Diabetes	Vanilla	$80.2_{\pm 0.1}$	$89.0_{\pm 0.1}$	$80.4_{\pm 0.1}$	$89.0_{\pm 0.1}$
	Balanced	$\mathbf{66.2}_{\pm 0.9}$	$\mathbf{75.9}_{\pm 0.8}$	$\mathbf{65.1}_{\pm 0.4}$	$\mathbf{74.6}_{\pm 0.4}$
	Correlation R.	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	Uncertain+LR	$70.4_{\pm 20.7}$	$77.1_{\pm 26.8}$	$80.3_{\pm 0.1}$	$89.0_{\pm 0.1}$
	Uncertain+TabPFN	$74.9_{\pm 10.0}$	$84.8_{\pm 8.0}$	$80.2_{\pm 0.1}$	$89.0_{\pm 0.1}$
CelebA	Vanilla	$84.7_{\pm 0.2}$	$83.2_{\pm 0.3}$	$85.0_{\pm 0.2}$	$83.2_{\pm 0.3}$
	Balanced	$84.6_{\pm 0.2}$	$83.2_{\pm 0.3}$	$84.9_{\pm 0.2}$	$83.3_{\pm 0.3}$
	Correlation R.	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	Uncertain+LR	$\mathbf{72.4}_{\pm 11.7}$	$\mathbf{61.2}_{\pm 21.7}$	$84.9_{\pm 0.2}$	$83.1_{\pm 0.3}$
	Uncertain+TabPFN	$74.5_{\pm 8.4}$	$70.8_{\pm 10.1}$	$\mathbf{81.2}_{\pm 7.2}$	$\mathbf{77.2}_{\pm 11.9}$

Table 3. **ICL prediction performance of sensitive attributes after applying different fairness interventions**. Smaller accuracy is better since it indicates how well the foundation model can reconstruct the sensitive attribute after the pre-processing fairness interventions. `Uncertain` methods yield the smallest accuracy, which justifies the improved fairness performance.

Dataset	Fairness Intervention	TabPFN		TabICL	
		Accuracy ↓	F1 Score ↓	Accuracy ↓	F1 Score ↓
ACSIncome	None	$77.2_{\pm 0.5}$	$78.4_{\pm 0.3}$	$75.0_{\pm 0.5}$	$76.18_{\pm 0.3}$
	`Correlation R.` (S1)	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$99.9_{\pm 0.0}$	$99.93_{\pm 0.0}$
	`Correlation R.` (S2)	$\mathbf{53.8}_{\pm 0.4}$	$\mathbf{67.0}_{\pm 0.9}$	$\mathbf{52.9}_{\pm 0.3}$	$\mathbf{68.9}_{\pm 0.3}$
ACSTravelTime	None	$75.9_{\pm 0.4}$	$77.5_{\pm 0.5}$	$72.8_{\pm 0.4}$	$73.75_{\pm 0.5}$
	`Correlation R.` (S1)	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.00_{\pm 0.0}$
	`Correlation R.` (S2)	$\mathbf{53.6}_{\pm 1.4}$	$\mathbf{54.2}_{\pm 14.5}$	$\mathbf{51.8}_{\pm 1.4}$	$\mathbf{66.4}_{\pm 3.4}$
ACSPublicCoverage	None	$91.4_{\pm 0.2}$	$88.9_{\pm 0.2}$	$91.5_{\pm 0.1}$	$89.23_{\pm 0.2}$
	`Correlation R.` (S1)	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.00_{\pm 0.0}$
	`Correlation R.` (S2)	$\mathbf{57.8}_{\pm 0.4}$	$\mathbf{0.1}_{\pm 0.1}$	$\mathbf{56.5}_{\pm 1.0}$	$\mathbf{0.1}_{\pm 0.1}$
ACSEmployment	None	$64.0_{\pm 0.4}$	$62.0_{\pm 1.8}$	$65.0_{\pm 0.3}$	$62.23_{\pm 1.4}$
	`Correlation R.` (S1)	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.00_{\pm 0.0}$
	`Correlation R.` (S2)	$\mathbf{52.6}_{\pm 0.6}$	$\mathbf{27.7}_{\pm 4.4}$	$\mathbf{53.7}_{\pm 5.1}$	$\mathbf{53.0}_{\pm 10.6}$
ACSMobility	None	$68.3_{\pm 0.8}$	$67.8_{\pm 1.0}$	$67.6_{\pm 1.0}$	$67.40_{\pm 1.2}$
	`Correlation R.` (S1)	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.00_{\pm 0.0}$
	`Correlation R.` (S2)	$\mathbf{49.2}_{\pm 1.5}$	$\mathbf{49.2}_{\pm 13.2}$	$\mathbf{49.2}_{\pm 0.8}$	$\mathbf{40.2}_{\pm 31.2}$
Diabetes	None	$80.2_{\pm 0.1}$	$89.0_{\pm 0.1}$	$80.4_{\pm 0.1}$	$89.04_{\pm 0.1}$
	`Correlation R.` (S1)	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.00_{\pm 0.0}$
	`Correlation R.` (S2)	$\mathbf{67.8}_{\pm 13.8}$	$\mathbf{78.9}_{\pm 12.2}$	$\mathbf{79.3}_{\pm 0.9}$	$\mathbf{88.4}_{\pm 0.6}$
German Credit	None	$72.5_{\pm 3.1}$	$70.3_{\pm 3.1}$	$71.8_{\pm 3.1}$	$71.02_{\pm 3.0}$
	`Correlation R.` (S1)	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.00_{\pm 0.0}$
	`Correlation R.` (S2)	$\mathbf{37.8}_{\pm 6.7}$	$\mathbf{44.0}_{\pm 14.5}$	$\mathbf{46.3}_{\pm 4.2}$	$\mathbf{38.8}_{\pm 14.6}$
CelebA	None	$84.7_{\pm 0.2}$	$83.2_{\pm 0.3}$	$85.0_{\pm 0.2}$	$83.16_{\pm 0.3}$
	`Correlation R.` (S1)	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.00_{\pm 0.0}$
	`Correlation R.` (S2)	$\mathbf{54.8}_{\pm 0.3}$	$\mathbf{1.0}_{\pm 0.8}$	$\mathbf{45.2}_{\pm 0.2}$	$\mathbf{62.2}_{\pm 0.2}$

Table 4. Accuracy ICL prediction of sensitive attribute after applying `Correlation Remover` on the training and testing datasets (S1) or only to train dataset (S2). Applying the transformation only to the train dataset significantly reduces the accuracy of predicting the sensitive attribute.

Dataset	ICL Method	Accuracy ↑	Δ DP ↓	Δ EOP ↓	Δ EOD ↓
ACSIncome	Vanilla	$80.76_{\pm 0.5}$	$14.21_{\pm 1.3}$	$5.46_{\pm 2.9}$	$5.89_{\pm 2.4}$
	Balanced	$80.99_{\pm 0.4}$	$14.08_{\pm 1.8}$	$5.36_{\pm 3.6}$	$5.79_{\pm 3.3}$
	Correlation R.	$81.44_{\pm 0.5}$	$26.81_{\pm 2.0}$	$27.35_{\pm 6.4}$	$27.35_{\pm 6.4}$
	Uncertain+LR	$80.94_{\pm 0.4}$	$13.13_{\pm 1.6}$	$3.91_{\pm 2.4}$	$4.17_{\pm 2.1}$
	Uncertain+TabPFN	$80.13_{\pm 0.8}$	$8.90_{\pm 1.7}$	$3.33_{\pm 3.9}$	$3.56_{\pm 3.7}$
ACSEmployment	Vanilla	$82.18_{\pm 0.3}$	$1.11_{\pm 0.7}$	$1.01_{\pm 1.1}$	$8.17_{\pm 0.8}$
	Balanced	$82.13_{\pm 0.4}$	$1.09_{\pm 0.8}$	$1.10_{\pm 1.2}$	$8.20_{\pm 0.7}$
	Correlation R.	$82.41_{\pm 0.4}$	$3.90_{\pm 1.0}$	$5.05_{\pm 1.2}$	$5.49_{\pm 1.0}$
	Uncertain+LR	$81.98_{\pm 0.4}$	$0.86_{\pm 0.4}$	$0.99_{\pm 0.8}$	$7.56_{\pm 0.8}$
	Uncertain+TabPFN	$81.69_{\pm 0.7}$	$0.80_{\pm 0.4}$	$0.91_{\pm 0.6}$	$6.99_{\pm 0.8}$
ACSPublicCoverage	Vanilla	$84.71_{\pm 0.5}$	$1.75_{\pm 1.2}$	$5.13_{\pm 3.5}$	$5.23_{\pm 3.4}$
	Balanced	$84.49_{\pm 0.5}$	$1.75_{\pm 1.4}$	$5.80_{\pm 2.8}$	$5.80_{\pm 2.8}$
	Correlation R.	$84.57_{\pm 0.6}$	$1.64_{\pm 1.2}$	$4.89_{\pm 3.7}$	$4.94_{\pm 3.6}$
	Uncertain+LR	$80.15_{\pm 2.0}$	$3.08_{\pm 3.2}$	$4.61_{\pm 4.1}$	$5.28_{\pm 3.4}$
	Uncertain+TabPFN	$81.58_{\pm 1.4}$	$2.14_{\pm 1.4}$	$7.78_{\pm 4.3}$	$7.78_{\pm 4.3}$
ACSTravelTime	Vanilla	$70.52_{\pm 0.6}$	$9.79_{\pm 0.9}$	$8.09_{\pm 2.4}$	$8.79_{\pm 1.9}$
	Balanced	$70.85_{\pm 0.6}$	$10.14_{\pm 1.7}$	$8.46_{\pm 3.5}$	$9.32_{\pm 2.6}$
	Correlation R.	$70.42_{\pm 0.5}$	$9.82_{\pm 1.8}$	$8.43_{\pm 3.0}$	$8.70_{\pm 2.5}$
	Uncertain+LR	$70.48_{\pm 0.5}$	$10.16_{\pm 0.7}$	$8.42_{\pm 2.4}$	$9.07_{\pm 1.8}$
	Uncertain+TabPFN	$70.00_{\pm 0.6}$	$8.30_{\pm 0.9}$	$6.69_{\pm 2.1}$	$7.49_{\pm 1.6}$
ACSMobility	Vanilla	$76.86_{\pm 0.8}$	$2.25_{\pm 1.3}$	$1.91_{\pm 0.6}$	$6.49_{\pm 2.8}$
	Balanced	$77.11_{\pm 1.1}$	$2.86_{\pm 1.6}$	$0.97_{\pm 0.9}$	$8.38_{\pm 4.9}$
	Correlation R.	$76.86_{\pm 0.6}$	$6.27_{\pm 1.6}$	$3.33_{\pm 1.7}$	$12.28_{\pm 3.3}$
	Uncertain+LR	$76.59_{\pm 0.8}$	$1.86_{\pm 1.7}$	$2.27_{\pm 1.0}$	$4.31_{\pm 2.4}$
	Uncertain+TabPFN	$76.58_{\pm 0.9}$	$2.05_{\pm 2.0}$	$1.95_{\pm 1.2}$	$4.34_{\pm 4.4}$
Diabetes	Vanilla	$64.59_{\pm 0.3}$	$1.74_{\pm 1.0}$	$1.78_{\pm 1.3}$	$2.74_{\pm 1.6}$
	Balanced	$64.70_{\pm 0.5}$	$1.62_{\pm 1.2}$	$2.37_{\pm 1.3}$	$3.08_{\pm 1.4}$
	Correlation R.	$64.69_{\pm 0.3}$	$1.33_{\pm 1.1}$	$2.13_{\pm 1.5}$	$2.68_{\pm 1.2}$
	Uncertain+LR	$64.39_{\pm 0.6}$	$0.77_{\pm 0.5}$	$2.52_{\pm 1.1}$	$2.52_{\pm 1.1}$
	Uncertain+TabPFN	$64.24_{\pm 0.6}$	$1.20_{\pm 0.9}$	$3.21_{\pm 2.1}$	$3.59_{\pm 2.0}$
German Credit	Vanilla	$74.80_{\pm 4.6}$	$5.14_{\pm 3.5}$	$5.88_{\pm 4.3}$	$12.20_{\pm 6.6}$
	Balanced	$74.43_{\pm 3.0}$	$5.92_{\pm 5.5}$	$5.47_{\pm 5.0}$	$15.85_{\pm 10.3}$
	Correlation R.	$75.25_{\pm 3.5}$	$11.78_{\pm 6.3}$	$9.30_{\pm 3.8}$	$16.72_{\pm 9.3}$
	Uncertain+LR	$74.36_{\pm 3.4}$	$7.72_{\pm 3.8}$	$6.01_{\pm 4.1}$	$11.98_{\pm 5.9}$
	Uncertain+TabPFN	$73.98_{\pm 4.0}$	$4.65_{\pm 3.0}$	$5.21_{\pm 4.3}$	$11.91_{\pm 9.6}$
CelebA	Vanilla	$80.55_{\pm 0.4}$	$14.54_{\pm 1.0}$	$12.03_{\pm 2.8}$	$12.03_{\pm 2.8}$
	Balanced	$80.45_{\pm 0.5}$	$15.06_{\pm 0.8}$	$13.18_{\pm 1.7}$	$13.18_{\pm 1.7}$
	Correlation R.	$80.47_{\pm 0.4}$	$13.23_{\pm 1.3}$	$9.04_{\pm 2.2}$	$9.14_{\pm 2.2}$
	Uncertain+LR	$80.16_{\pm 0.5}$	$8.92_{\pm 0.9}$	$2.28_{\pm 2.0}$	$4.42_{\pm 1.3}$
	Uncertain+TabPFN	$79.86_{\pm 0.7}$	$10.01_{\pm 1.4}$	$3.54_{\pm 1.9}$	$5.46_{\pm 1.1}$

Table 5. This table supplements Figure **??** in main paper. It shows the accuracy and fairness performance of ICL predictions with TabPFN as foundation model under different preprocessing methods. The color range highlights the best ( ) to the worst-performing method ( ) for fairness accuracy. ↑ indicates higher is better (accuracy) and ↓ lower is better (unfairness).

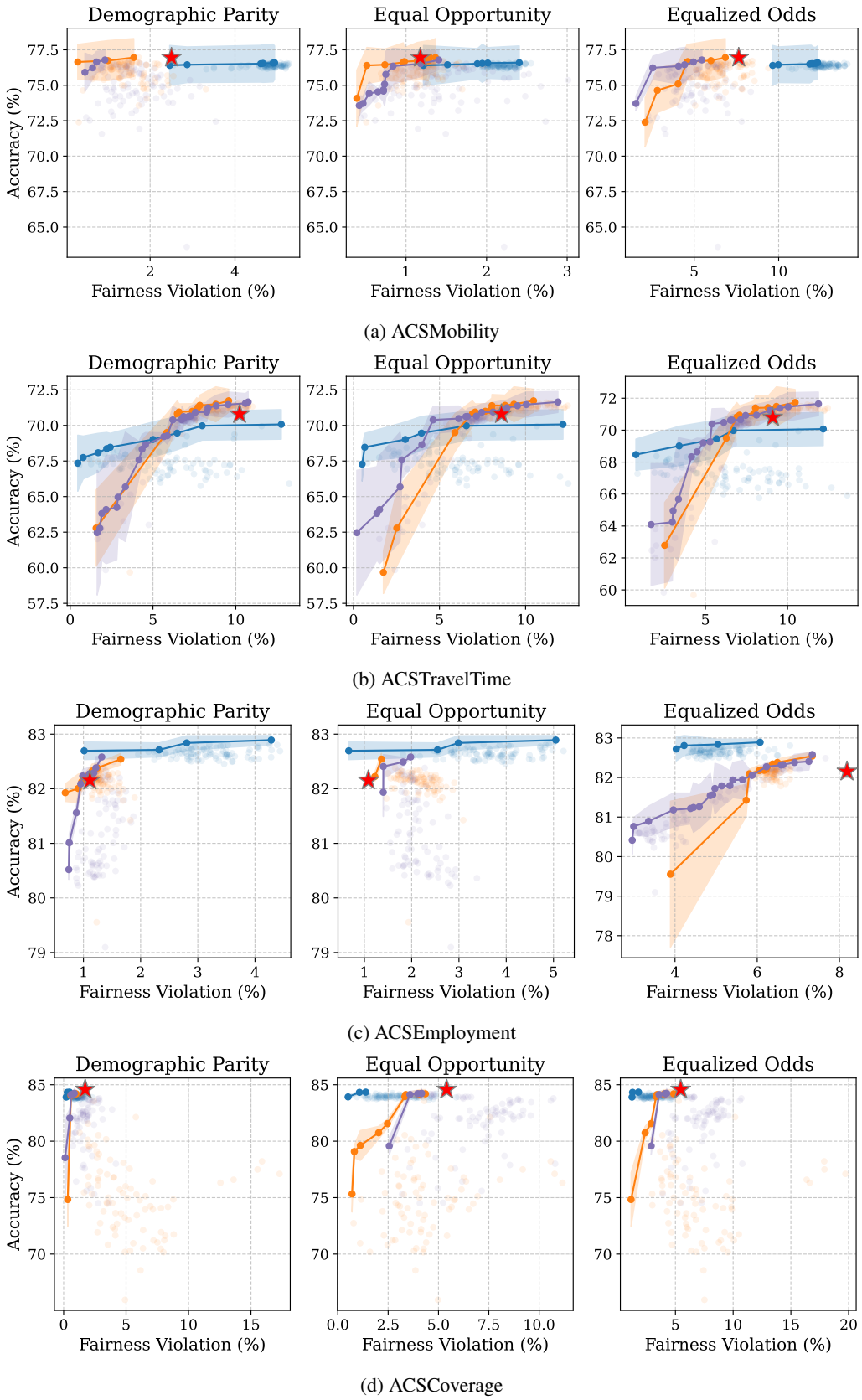
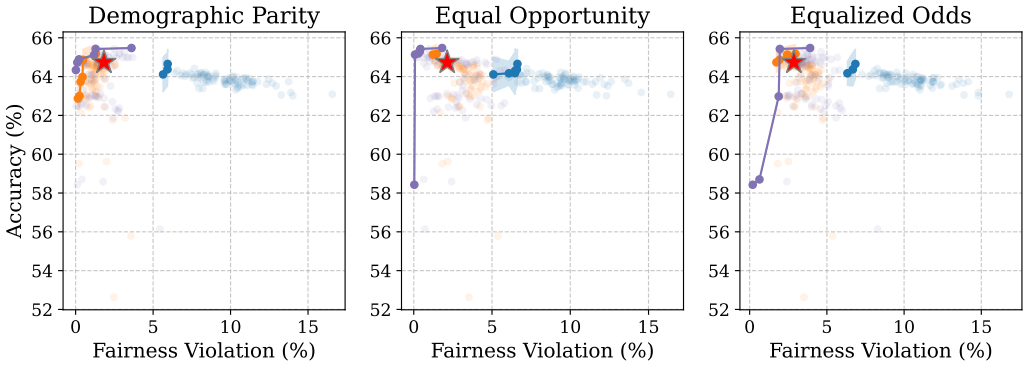
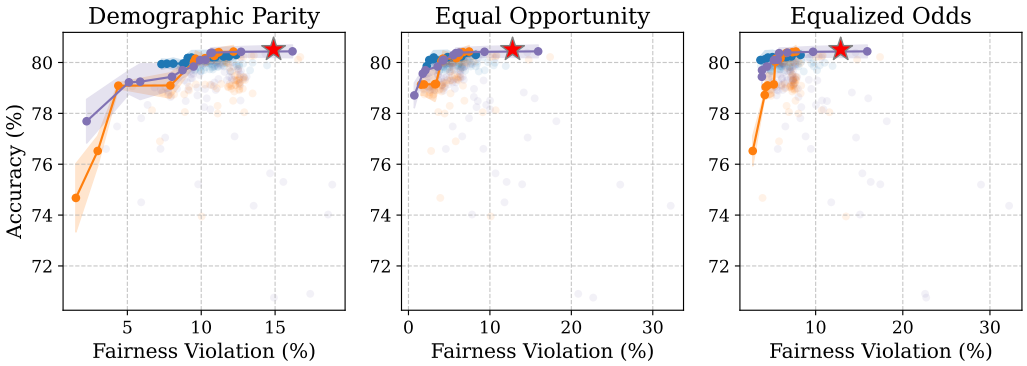


Figure 5. **Fairness-accuracy tradeoffs on ACS dataset**. Comparing the fairness-accuracy Pareto-front of different fairness interventions using TabPFN as foundation model. Results with TabICL can be found in Fig 7.



(a) Diabete



(b) CelebA

Figure 6. **Fairness-accuracy tradeoffs on Diabete and CelebA datasets**. Comparing the fairness-accuracy Pareto-front of different fairness interventions using TabPFN as foundation model. Results with TabICL can be found in Fig. 7 and 8.

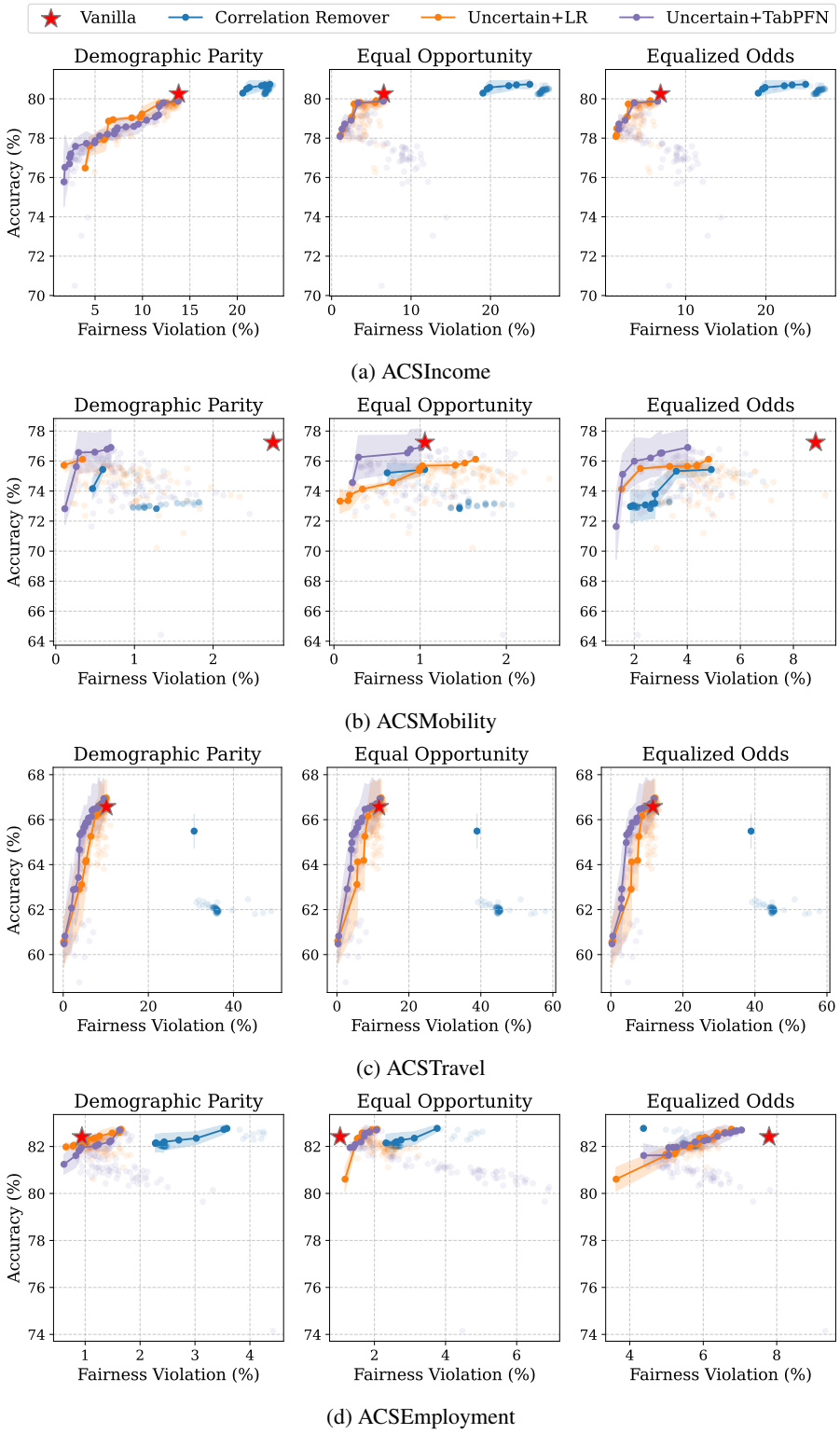


Figure 7. Fairness-accuracy tradeoffs on the ACS datasets using TabICL as foundation model.

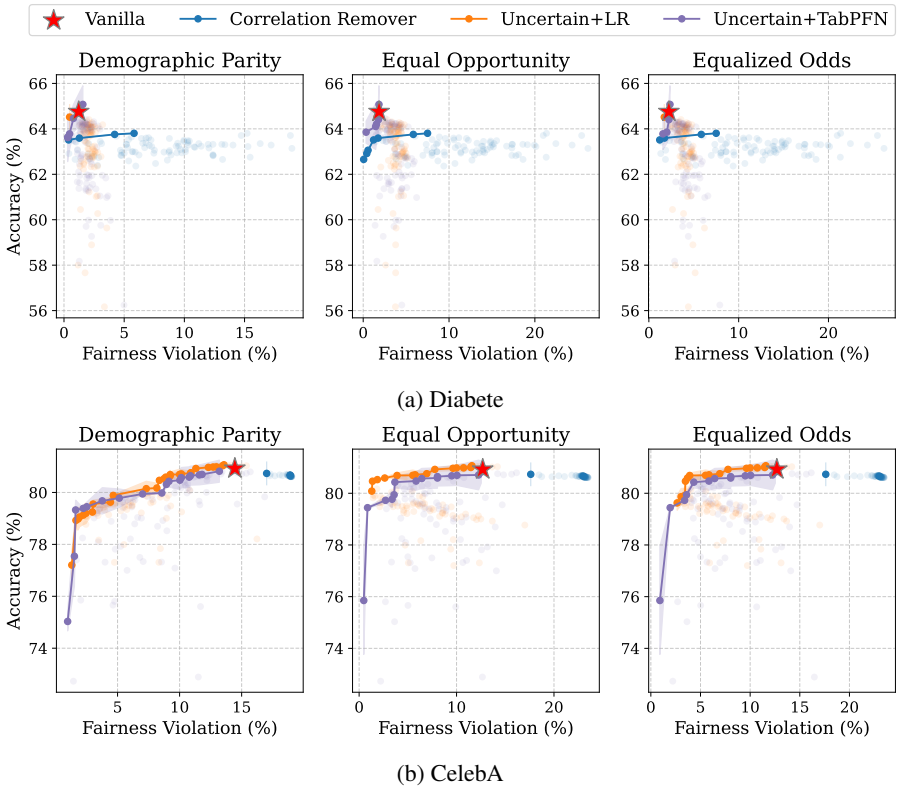


Figure 8. Fairness-accuracy tradeoffs on the Diabete and CelebA datasets using TabICL as foundation model

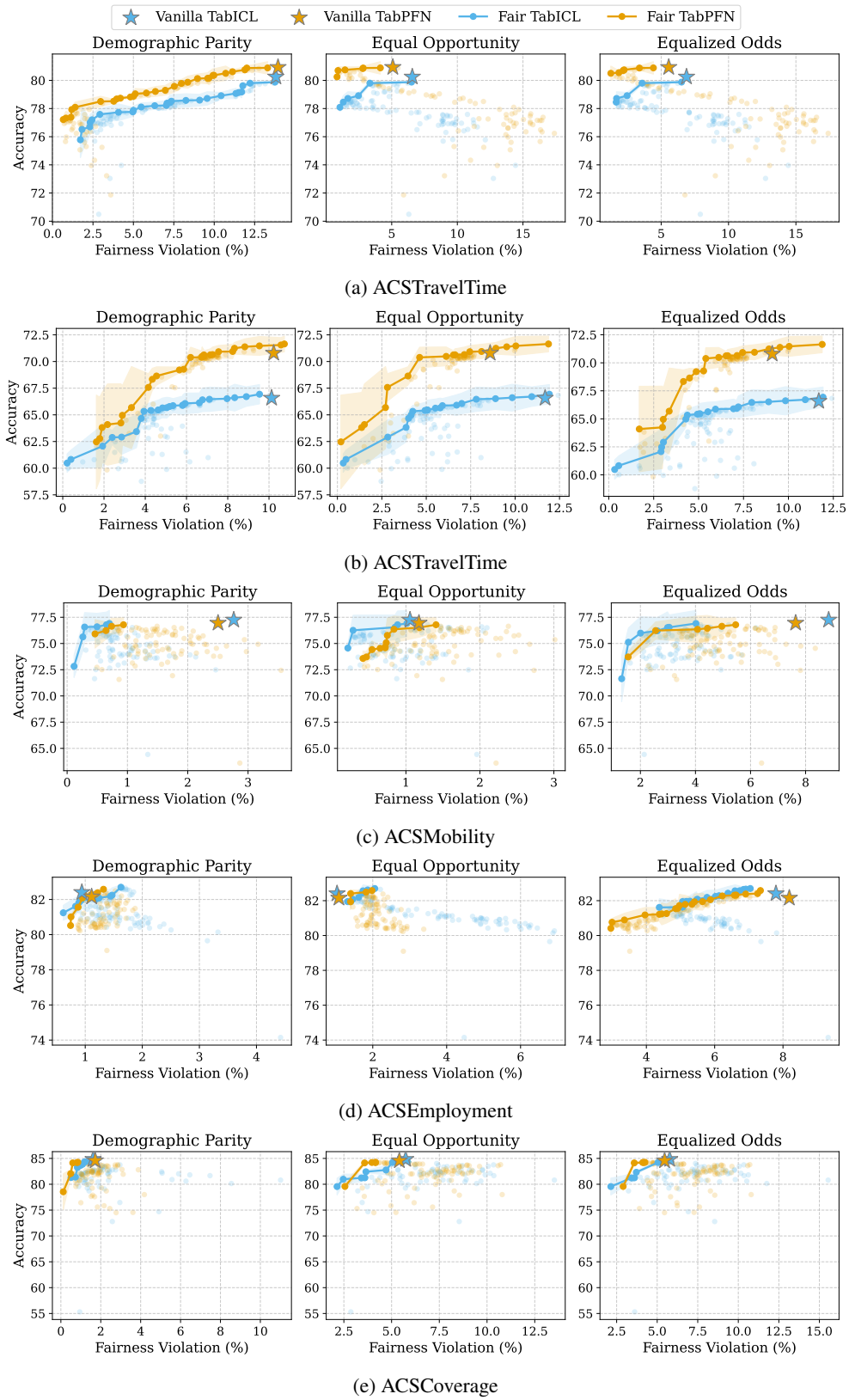
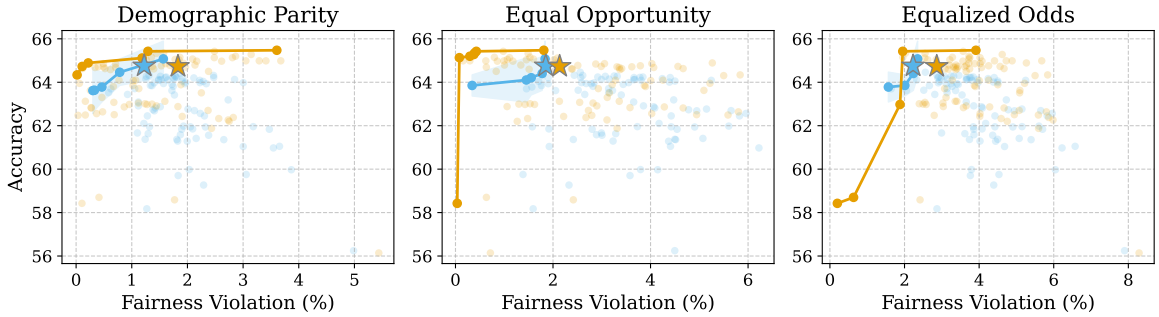
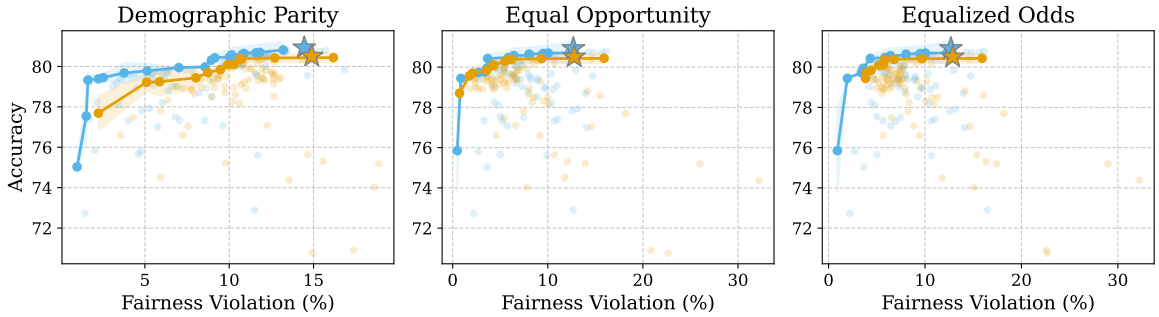


Figure 9. **TabPFN vs. TabICL on ACS datasets**. Comparing the fairness-accuracy tradeoffs of tabular foundation models under different fairness interventions. TabPFN generally provides better fairness accuracy tradeoffs.



(a) Diabetes



(b) CelebA

Figure 10. **TabPFN vs. TabICL on Diabetes and CelebA**. Comparing the fairness-accuracy tradeoffs of tabular foundation models under different fairness interventions. TabPFN generally provides better fairness accuracy tradeoffs.