Instructions for *ACL Proceedings

Anonymous ACL submission

Abstract

This document is a supplement to the general instructions for *ACL authors. It contains instructions for using the LaTeX style files for ACL conferences. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. These instructions should be used both for papers submitted for review and for final versions of accepted papers.

1 Introduction

Large Language Models (LLMs) have recently shown significant potential in enhancing traditional tabular classification tasks, especially in few-shot learning scenarios (Hegselmann et al., 2023; Yang et al., 2024; Han et al., 2024; van Breugel and van der Schaar, 2024). The effectiveness of LLMs can be primarily attributed to two key factors: their extensive pre-training or fine-tuning on vast amounts of web-scale data, which provides a rich repository of prior knowledge, and their inherent reasoning capabilities(Wang et al., 2023; Fang et al., 2024; Badaro et al., 2023). Given the growing use of LLMs in this domain, some studies have also explored how in-context learning can be leveraged to maintain algorithmic fairness. By incorporating diverse and representative samples into pretrained or fine-tuned LLMs, these approaches aim to mitigate bias across different sensitive groups (Hu et al., 2024; Liu et al., 2023; Chhikara et al., 2024).

Current research in this domain encounters several constraints. One key limitation is that many recent methods focus on serializing tabular data and inputting both the data and task descriptions into LLMs through in-context learning or fine-tuning techniques. This approach requires at least one inference per sample, resulting in high computational costs(Han et al., 2024). Moreover, the API-level access to certain advanced LLMs often renders fine-tuning impractical(OpenAI, 2023). Additionally, tabular data generally consist of numerous features, and despite recent improvements in extending context window lengths, the need for lengthy descriptions and multiple demonstrations can lead to inefficiency and performance degradation(Li et al., 2024; Chen et al., 2023). These challenges significantly impede the effective application of LLM-based models for tabular data predictions.

In addition, the need to ensure algorithmic fairness introduces further limitations. Simply adjusting the proportion of demonstrations between groups and classes is insufficient to fully mitigate implicit biases (Hu et al., 2024), and model performance is often sensitive to the selection of demonstrations (Dong et al., 2022). For example, even if all the demonstrations are sampled from the underrepresented group, fairness criteria may still be violated.(Hu et al., 2024)Those limitations above demand us to enhance the algorithmic fairness in a more controllable and effective way.

Thus, the research problem can be framed as follows: *How can we achieve and maintain algorithmic fairness in a more controlled manner while minimizing computational costs?*

Inspired by FeatLLM (Han et al., 2024), we introduce Fair-FeatLLM. We treat LLMs as encoders and prompt them to generate new features and derive 'rules' for transforming input samples, leveraging their prior knowledge and task descriptions. Our approach can be summarized into 3 steps. Firstly, the LLMs are prompted to filter out irrelevant or discriminative features, synthesize new features that capture the relationships among the existing features, and select the most important features for downstream tasks. This approach mitigates explicit bias towards underrepresented groups while addressing the challenge of excessive prompt size. Secondly, the LLMs can infer and generate rules corresponding to the most likely feature conditions. For example, in predicting the likelihood of rearrest (e.g., using the COMPAS dataset (Angwin et al., 2022)), a potential rule might be "df['length_of_stay'] > 10", where feature 'length_of_stay' refers to the number of days spent in jail. Once these rules are established, LLMs are no longer required in downstream tasks, as the newly generated features, derived from the rules, replace the original ones. Finally, by ensembling these features and incorporating fairness constraints, we can enhance model performance while ensuring fairness. In addition to the methodological contribution, We also adapt the High School Longitudinal Study (HSLS) dataset (Ingels et al., 2011), a commonly used dataset in traditional fair machine learning, to make it suitable for research on LLM-based fair tabular predictions. Furthermore, we put forth a rencently published dataset for thyroid cancer prediction (Borzooei et al., 2024) to address potential issues of data leakage. These two datasets are intended to encourage researchers to evaluate LLM-based fair tabular prediction methods beyond the commonly used Adult(Asuncion et al., 2007) and COMPAS(Angwin et al., 2022) datasets. The workflow can be seen in....

Our contributions are as follows:

- We introduce Fair-FeatLLM, a novel framework that leverages large language models (LLMs) for fair tabular classification, offering enhanced control and reduced computational expense.
- We demonstrate that Fair-FeatLLM outperforms existing fair in-context learning methods and achieves superior performance compared to traditional approaches, particularly in few-shot learning scenarios.
- We adapt the HSLS dataset, a widely used dataset in traditional fair machine learning, to be applicable for studies on LLM-based fair tabular prediction. Additionally, we put forth a new thyroid cancer prediction dataset to avoid potential data leakage. We hope these two datasets will motivate researchers to evaluate LLM-based fair tabular prediction methods beyond the commonly used Adult and COMPAS datasets.

2 Related Work

There are three research directions closely related to our work: (1) LLMs for tabular classification and (2) Fair Machine Learning. For simplicity, we include the discussion of LLMs for fair tabular classification in the second subsection.

2.1 LLMs for Tabular Prediction

Tabular prediction has emerged as a prominent topic across various fields and has been extensively studied in recent years (Yan et al., 2024; Zhang et al., 2023; Qin et al., 2021; Dong and Wang, 2024). Traditional approaches to tabular classification primarily center on the debate between neural networks (NNs) and gradient-boosted trees in fully supervised settings. Some researches suggests that the skewness of datasets and the distribution shift between training and test sets are key factors influencing this debate(Ye et al., 2024; McElfresh et al., 2024). Recently, LLMs have demonstrated remarkable performance across a variety of tasks, often requiring little to no labeled data(Hegselmann et al., 2023; Wang et al., 2023). Several studies have explored the serialization of tabular data into natural language and its subsequent input into LLMs, along with task descriptions, for tabular prediction. However, in-context learning methods generally require at least one LLM inference per sample, which incurs significant computational costs. Furthermore, the black-box nature of proprietary LLMs often renders fine-tuning impractical.

Inspired by FeatLLM (Han et al., 2024), we aim to leverage LLMs as feature engineers while minimizing their usage. This methods can not only use the prior knowledge of LLMs to mitigate the shift of the training set distribution and test set distribution caused by the few-shot training samples, but also avoid the repeated calls of LLMs during inference phase. Our approach differs from FeatLLM in two key ways: (1) While FeatLLM focuses on general tabular prediction, our work emphasizes maintaining algorithmic fairness; (2) Instead of generating rules based on fixed features, as in FeatLLM, we prompt LLMs to create new features through the combination of existing features and select the top-K most important features. This method captures feature relationships while excluding noise, resulting in more robust predictions.

2.2 Fair Machine Learning

Fair machine learning (FairML) methods have substantially developed in recent years (Shui et al., 2022; Zhang et al., 2022b; Deng et al., 2022; Kang et al., 2022; Qi et al., 2022; Zhang et al., 2022a). Some recent works on LLm-based fair prediction demonstrate the power of LLMs(Hu et al., 2024;

Liu et al., 2023; Chhikara et al., 2024). However, simply resampling the data does not fully address the underlying bias(Zietlow et al., 2022; Corbett-Davies et al., 2023; Sühr et al., 2021). The reasoning behind this is straightforward: assume we aim for nearly equal performance across sensitive groups. Even if we fully train the classifier on data from the underprivileged groups and achieve optimal performance for them, any remaining performance gap compared to the privileged groups may still violate fairness requirements. In such cases, the only way to achieve fairness would be to degrade the performance of both sensitive groups, a goal that cannot be accomplished solely by altering the proportions of the training dataset. This limitation has been highlighted in several studies (Pinzón et al., 2022; Liang et al., 2021).

Additionally, some research emphasizes that in-context learning can be decomposed into two stages: task recognition and task learning(Dai et al., 2022; Pan, 2023). Initially, LLMs familiarize themselves with the task, and as the number of demonstrations increases, they gradually internalize the underlying patterns of the data, effectively performing an implicit form of gradient descent. Consequently, relying solely on this implicit learning process to establish a fair mapping between input features and labels may not sufficiently eliminate bias. Therefore, there is a pressing need to develop novel methods that allow for a more controlled and effective approach to ensuring algorithmic fairness.

3 Preliminary

In this section, we introduce the notations, metrics and the datasets used throughout this paper.

3.1 Notation

Each individual is characterized by a covariate vector X drawn from the set $\mathcal{X}$, a label Y from the finite set $\mathcal{Y}$, and a sensitive attribute A from the finite set $\mathcal{A}$. The predicted value $\hat{Y}$ also belongs to $\mathcal{Y}$. In our investigation, we focus on binary classification with binary sensitive attributes. However, our analysis can be generalized to accommodate multiple sensitive attributes, which we defer to the future work. Consequently, $\mathcal{Y} = \{0, 1\}$ and $\mathcal{A} = \{0, 1\}$. For example, in the context of job hiring, $Y = 1$ denotes a high-quality job applicant, $A = 1$ denotes a male applicant, and $\hat{Y} = 1$ denotes that the applicant is hired. The vector X encompasses non-sensitive attributes such as resumes and references.

3.2 Metrics

Throughout this paper, we report the following three metrics:

3.2.1 Accuracy

Accuracy is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

3.2.2 Demographic Parity

Demographic Parity (DP) is a fairness criterion that ensures the probability of a positive outcome is the same across different demographic groups. It seeks to guarantee that each group has an equal chance of receiving a favorable outcome, regardless of other factors. Given this criterion, we quantify the unfairness as follows:

$$\Delta_{DP} = |P(\hat{Y} = 1 \mid A = 0) - P(\hat{Y} = 1 \mid A = 1)|$$

3.2.3 Equalized Odds

Equalized Odds (EOdd) is a more stringent fairness metric that ensures fairness across both true positive and false positive rates. It requires that individuals from different demographic groups have equal true positive rates and equal false positive rates. Based on this, we quantify the unfairness as follows:

$$\Delta_{EOdd} = \frac{1}{2} * (|\Delta_{TPR}| + |\Delta_{FPR}|)$$

where $\Delta_{TPR} = P(\hat{Y} = 1 \mid Y = 1, A = 0) - P(\hat{Y} = 1 \mid Y = 1, A = 1)$ and $\Delta_{FPR} = P(\hat{Y} = 1 \mid Y = 0, A = 0) - P(\hat{Y} = 1 \mid Y = 0, A = 1)$, respectively.

3.3 Datasets

As much as possible, fonts in figures should conform to the document fonts. See Figure 1 for an example of a figure and its caption.

Using the `graphicx` package graphics files can be included within figure environment at an appropriate point within the text. The `graphicx` package supports various optional arguments to control the appearance of the figure. You must include it explicitly in the LaTeX preamble (after the `\documentclass` declaration and before `\begin{document}`) using `\usepackage{graphicx}`.

Command	Output	Command	Output
`{\"a}`	ä	`{\c c}`	ç
`{\^e}`	ê	`{\u g}`	ğ
``{\`i}``	ì	`{\l}`	ł
`{\.I}`	İ	`{\~n}`	ñ
`{\o}`	ø	`{\H o}`	ő
`{\'u}`	ú	`{\v r}`	ř
`{\aa}`	å	`{\ss}`	ß

Table 1: Example commands for accented characters, to be used in, *e.g.*, BibTEX entries.

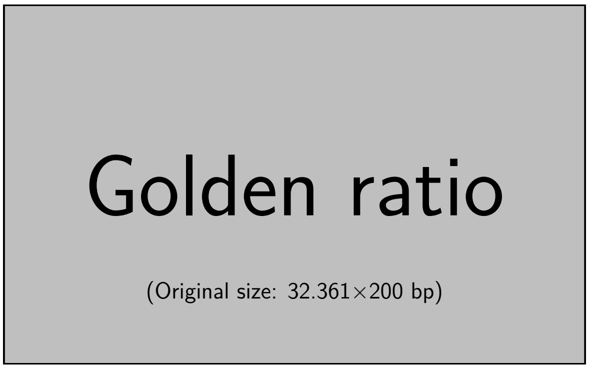

Figure 1: A figure with a caption that runs for more than one line. Example image is usually available through the `mwe` package without even mentioning it in the preamble.

3.4 Hyperlinks

Users of older versions of LATEX may encounter the following error during compilation:

> `\pdfendlink` ended up in different nesting level than `\pdfstartlink`.

This happens when pdfLATEX is used and a citation splits across a page boundary. The best way to fix this is to upgrade LATEX to 2018-12-01 or later.

3.5 Citations

Table 2 shows the syntax supported by the style files. We encourage you to use the natbib styles. You can use the command `\citet` (cite in text) to get "author (year)" citations, like this citation to a paper by **?**. You can use the command `\citep` (cite in parentheses) to get "(author, year)" citations (**?**). You can use the command `\citealp` (alternative cite without parentheses) to get "author, year" citations, which is useful for using citations within parentheses (e.g. **?**).

A possessive citation can be made with the command `\citeposs`. This is not a standard natbib command, so it is generally not compatible with other style files.

3.6 References

The LATEX and BibTEX style files provided roughly follow the American Psychological Association format. If your own bib file is named `custom.bib`, then placing the following before any appendices in your LATEX file will generate the references section for you:

```
\bibliography{custom}
```

You can obtain the complete ACL Anthology as a BibTEX file from `https://aclweb.org/anthology/anthology.bib.gz`. To include both the Anthology and your own .bib file, use the following instead of the above.

```
\bibliography{anthology,custom}
```

Please see Section 4 for information on preparing BibTEX files.

3.7 Equations

An example equation is shown below:

$$A = \pi r^2 \tag{1}$$

Labels for equation numbers, sections, subsections, figures and tables are all defined with the `\label{label}` command and cross references to them are made with the `\ref{label}` command.

This an example cross-reference to Equation 1.

3.8 Appendices

Use `\appendix` before any appendix section to switch the section numbering over to letters. See Appendix A for an example.

4 BibTEX Files

Unicode cannot be used in BibTEX entries, and some ways of typing special characters can disrupt BibTEX's alphabetization. The recommended way of typing special characters is shown in Table 1.

Please ensure that BibTEX records contain DOIs or URLs when possible, and for all the ACL materials that you reference. Use the `doi` field for DOIs and the `url` field for URLs. If a BibTEX entry has a URL or DOI field, the paper title in the references section will appear as a hyperlink to the paper, using the hyperref LATEX package.

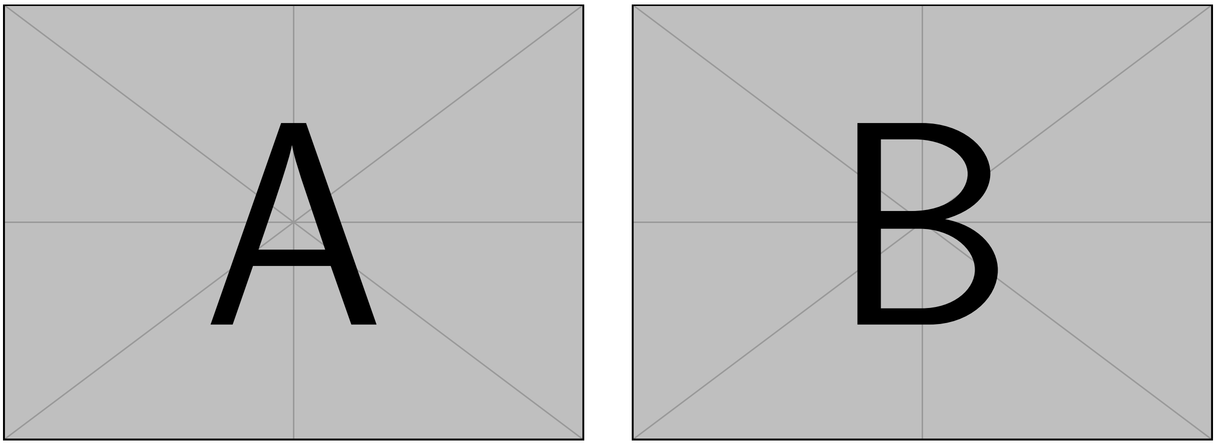

Figure 2: A minimal working example to demonstrate how to place two images side-by-side.

Output	natbib command	ACL only command
(**?**)	`\citep`	
?	`\citealp`	
?	`\citet`	
(**?**)	`\citeyearpar`	
?'s (**?**)		`\citeposs`

Table 2: Citation commands supported by the style file. The style is based on the natbib package and supports all natbib citation commands. It also supports commands defined in previous ACL style files for compatibility.

5 Limitations

References

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

Arthur Asuncion, David Newman, et al. 2007. Uci machine learning repository.

Gilbert Badaro, Mohammed Saeed, and Paolo Papotti. 2023. Transformers for tabular data representation: A survey of models and applications. *Transactions of the Association for Computational Linguistics*, 11:227–249.

Shiva Borzooei, Giovanni Briganti, Mitra Golparian, Jerome R Lechien, and Aidin Tarokhian. 2024. Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study. *European Archives of Oto-Rhino-Laryngology*, 281(4):2095–2104.

Jiuhai Chen, Lichang Chen, Chen Zhu, and Tianyi Zhou. 2023. How many demonstrations do you need for in-context learning? *arXiv preprint arXiv:2303.08119*.

Garima Chhikara, Anurag Sharma, Kripabandhu Ghosh, and Abhijnan Chakraborty. 2024. Few-shot fairness: Unveiling llm's potential for fairness-aware classification. *arXiv preprint arXiv:2402.18502*.

Sam Corbett-Davies, Johann D Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. 2023. The measure and mismeasure of fairness. *The Journal of Machine Learning Research*, 24(1):14730–14846.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2022. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. *arXiv preprint arXiv:2212.10559*.

Zhun Deng, Jiayao Zhang, Linjun Zhang, Ting Ye, Yates Coley, Weijie J Su, and James Zou. 2022. Fifa: Making fairness more generalizable in classifiers trained on imbalanced data. *arXiv preprint arXiv:2206.02792*.

Haoyu Dong and Zhiruo Wang. 2024. Large language models for tabular data: Progresses and future directions. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2997–3000.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, and Christos Faloutsos. 2024. Large language models on tabular data–a survey. *arXiv preprint arXiv:2402.17944*.

Sungwon Han, Jinsung Yoon, Sercan O Arik, and Tomas Pfister. 2024. Large language models can automatically engineer features for few-shot tabular learning. *arXiv preprint arXiv:2404.09491*.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR.

Jingyu Hu, Weiru Liu, and Mengnan Du. 2024. Strategic demonstration selection for improved fairness in llm in-context learning. *arXiv preprint arXiv:2408.09757*.

Steven J Ingels, Daniel J Pratt, Deborah R Herget, Laura J Burns, Jill A Dever, Randolph Ottem, James E Rogers, Ying Jin, and Steve Leinwand. 2011. High school longitudinal study of 2009 (hsls: 09): Base-year data file documentation. nces 2011-328. *National Center for Education Statistics*.

Mintong Kang, Linyi Li, Maurice Weber, Yang Liu, Ce Zhang, and Bo Li. 2022. Certifying some distributional fairness with subpopulation decomposition. *Advances in Neural Information Processing Systems*, 35:31045–31058.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*.

Annie Liang, Jay Lu, and Xiaosheng Mu. 2021. Algorithm design: A fairness-accuracy frontier. *arXiv preprint arXiv:2112.09975*.

Yanchen Liu, Srishti Gautam, Jiaqi Ma, and Himabindu Lakkaraju. 2023. Investigating the fairness of large language models for predictions on tabular data. *arXiv preprint arXiv:2310.14607*.

Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. 2024. When do neural nets outperform boosted trees on tabular data? *Advances in Neural Information Processing Systems*, 36.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5).

Jane Pan. 2023. What in-context learning "learns" in-context: Disentangling task recognition and task learning. Master's thesis, Princeton University.

Carlos Pinzón, Catuscia Palamidessi, Pablo Piantanida, and Frank Valencia. 2022. On the impossibility of non-trivial accuracy in presence of fairness constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7993–8000.

Tao Qi, Fangzhao Wu, Chuhan Wu, Lingjuan Lyu, Tong Xu, Hao Liao, Zhongliang Yang, Yongfeng Huang, and Xing Xie. 2022. Fairvfl: A fair vertical federated learning framework with contrastive adversarial learning. *Advances in Neural Information Processing Systems*, 35:7852–7865.

Jiarui Qin, Weinan Zhang, Rong Su, Zhirong Liu, Weiwen Liu, Ruiming Tang, Xiuqiang He, and Yong Yu. 2021. Retrieval & interaction machine for tabular data prediction. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1379–1389.

Changjian Shui, Gezheng Xu, Qi Chen, Jiaqi Li, Charles X Ling, Tal Arbel, Boyu Wang, and Christian Gagné. 2022. On learning fairness and accuracy on multiple subgroups. *Advances in Neural Information Processing Systems*, 35:34121–34135.

Tom Sühr, Sophie Hilgard, and Himabindu Lakkaraju. 2021. Does fair ranking improve minority outcomes? understanding the interplay of human and algorithmic biases in online hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 989–999.

Boris van Breugel and Mihaela van der Schaar. 2024. Why tabular foundation models should be a research priority. *arXiv preprint arXiv:2405.01147*.

Ruiyu Wang, Zifeng Wang, and Jimeng Sun. 2023. Unipredict: Large language models are universal tabular predictors. *arXiv preprint arXiv:2310.03266*.

Jiahuan Yan, Bo Zheng, Hongxia Xu, Yiheng Zhu, Danny Chen, Jimeng Sun, Jian Wu, and Jintai Chen. 2024. Making pre-trained language models great on tabular prediction. *arXiv preprint arXiv:2403.01841*.

Yazheng Yang, Yuqi Wang, Sankalok Sen, Lei Li, and Qi Liu. 2024. Unleashing the potential of large language models for predictive tabular tasks in data science. *arXiv preprint arXiv:2403.20208*.

Han-Jia Ye, Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, and De-Chuan Zhan. 2024. A closer look at deep learning on tabular data. *arXiv preprint arXiv:2407.00956*.

Fengda Zhang, Kun Kuang, Long Chen, Yuxuan Liu, Chao Wu, and Jun Xiao. 2022a. Fairness-aware contrastive learning with partially annotated sensitive attributes. In *The Eleventh International Conference on Learning Representations*.

Guanhua Zhang, Yihua Zhang, Yang Zhang, Wenqi Fan, Qing Li, Sijia Liu, and Shiyu Chang. 2022b. Fairness reprogramming. *Advances in Neural Information Processing Systems*, 35:34347–34362.

Tianping Zhang, Shaowen Wang, Shuicheng Yan, Jian Li, and Qian Liu. 2023. Generative table pre-training empowers models for tabular prediction. *arXiv preprint arXiv:2305.09696*.

Dominik Zietlow, Michael Lohaus, Guha Balakrishnan, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Chris Russell. 2022. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. *Preprint*, arXiv:2203.04913.

A Example Appendix

This is an appendix.