

**Optimizing Fairness Measures Through Neural Networks and Subdominance
Minimization**

BY

TRONG LINH VU

B.S., University of Illinois Chicago, 2023

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Chicago, 2024

Chicago, Illinois

Defense Committee:

Brian Ziebart, Chair and Advisor
Xiaoguang Wang
Abolfazl Asudeh

Copyright by
Trong Linh Vu
2024

To my mom,

For your unwavering support and encouragement,

Thank you for pushing me to pursue my studies in the United States.

Your belief in me has been my greatest motivation.

ACKNOWLEDGMENTS

Special thanks to Brian Ziebart, Omid Memarrast and Rushit Shah. Thank you, Professor Ziebart, for providing me with this opportunity to dive deep into research. It has been a pleasure working as your Teaching Assistant and Research Assistant. Thank you, Omid and Rushit, for helping me so much with this project. You guys have given me a ton of experience.

LV

CONTRIBUTION OF AUTHORS

In Chapter 1, Chapter 2, and a portion of Chapter 3, Portions of this work focused on superhuman fairness for logistic regression classifiers were in collaboration with Omid Memmarast and Brian Ziebart, appearing in ICML 2023 and within Omid Memmarast’s PhD thesis.

TABLE OF CONTENTS

<u>CHAPTER</u>	<u>PAGE</u>
1 INTRODUCTION	1
2 RELATED WORK	4
2.1 Group Fairness Measures	4
2.2 Performance-Fairness Trade-offs	6
2.3 Preference Elicitation & Imitation Learning	6
2.3.1 Subdominance Minimization for Improved Fairness-Aware Clas- sification	9
2.3.1.1 Superhumanness and Subdominance	9
2.3.1.2 Performance-Fairness Subdominance Minimization	10
3 APPROACH	14
3.1 Background of Neural Networks	14
3.1.1 Architecture of Neural Networks	14
3.1.1.1 Feedforward Neural Networks	14
3.1.1.2 Activation Functions	15
3.1.2 Training Neural Networks	15
3.1.3 Applications of Neural Networks	15
3.2 Subdominance Minimization with Neural Networks	16
3.2.1 Mathematical Formulation	16
3.2.2 Training Procedure	17
4 EXPERIMENTS AND RESULTS	19
4.1 Neural Network Architecture	19
4.2 Training and Testing Dataset Construction	20
4.3 Results	21
4.3.1 Comparision with Original Method	24
5 CONCLUSION	32
APPENDIX	33
CITED LITERATURE	36
VITA	40

LIST OF TABLES

<u>TABLE</u>		<u>PAGE</u>
I	Experimental results on noise-free datasets	23
II	Mean and Standard Deviation of Evaluation Methods	25

LIST OF FIGURES

<u>FIGURE</u>		<u>PAGE</u>
1	Trade-off between group disparity and predictive performance. The blue curve is the Pareto frontier. Shaded regions show "superhuman" performance levels, where algorithms outperform human decision-makers. Black dots represent specific algorithm configurations.	2
2	Visualization of the neural network structure	20
3	Experimental results on the <code>Adult</code> dataset.	22
4	Experimental results on the <code>Adult</code> dataset. EqOdds vs DP	26
5	Experimental results on the <code>Adult</code> dataset. DP vs Error	27
6	Experimental results on the <code>Adult</code> dataset. DP vs Error	28
7	Experimental results on the <code>Adult</code> dataset. DP vs Error	29
8	Experimental results on the <code>Adult</code> dataset. EqOdds vs Error	30
9	Experimental results on the <code>Adult</code> dataset. PRP vs Error	31

LIST OF ALGORITHMS

<u>ALGORITHM</u>	<u>PAGE</u>
1 Subdominance policy gradient optimization	12

SUMMARY

Machine learning methods increasingly have social impacts for which differences in performance for different groups can be undesirable. Many different measures of fairness have been defined to assess these differences and guide machine learning algorithms for classification towards better balances between unfairness and inaccuracy. Choosing which fairness measure(s) to employ can be challenging since each resulting classifier can have very different social impact. In this work, we avoid this choice of fairness measure(s) by reframing fair classification as an imitation learning problem. Rather than producing a classifier that balances predictive performance with a specific fairness measure, our approach seeks a classifier that is better than reference decisions across a set of fairness (and performance) measures. This thesis specifically investigates expanding the set of classifiers supported under this formulation from flat logistic regression classifiers to multilayer neural networks. This transition to neural networks enables a more robust and flexible model capable of handling complex datasets and achieving better performance in terms of fairness and accuracy.

CHAPTER 1

INTRODUCTION

(This chapter is based on a paper published as “Superhuman Fairness” (Memarrast et al., 2023b) in the International Conference on Machine Learning 40 (ICML 2023).

The social implications of algorithmic decisions driven by machine learning have pushed the development of various fairness criteria to ensure equitable outcomes [1, 2]. However, it is impossible to meet all common group fairness criteria simultaneously [3]. This means no decision-making process can be entirely fair to all groups and individuals under every fairness definition. As a result, specific weightings or trade-offs between different fairness criteria are often optimized [4]. Determining an appropriate trade-off for these fairness methods is challenging and can lead to philosophical and ideological debates, potentially hindering the adoption of algorithmic methods.

We focus on a scenario where well-meaning but inherently flawed human decision-makers currently make fairness-aware decisions. Instead of striving for optimal decisions based on specific performance-fairness trade-offs—which may be hard to define accurately—we propose a more feasible goal: to outperform human decisions in terms of both performance and fairness as frequently as possible. We assume that human decisions, though noisy and suboptimal, reflect desired performance-fairness trade-offs. This enables superhuman decisions that are superior to human decisions in terms of predictive performance and fairness metrics (Figure 1) without needing an explicit trade-off definition.

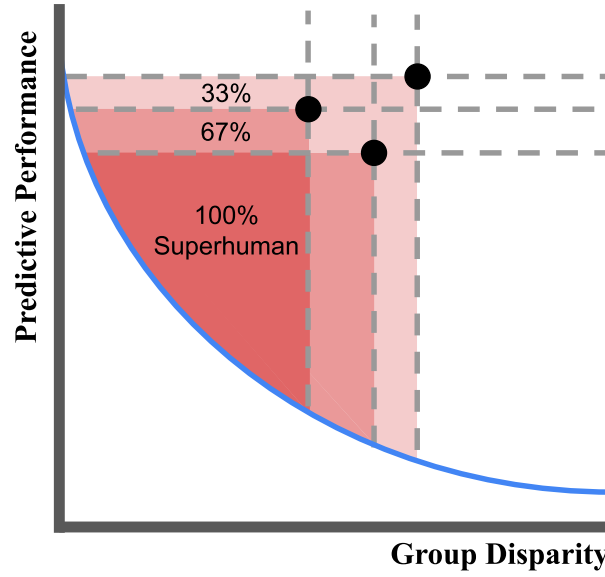


Figure 1: Trade-off between group disparity and predictive performance. The blue curve is the Pareto frontier. Shaded regions show "superhuman" performance levels, where algorithms outperform human decision-makers. Black dots represent specific algorithm configurations.

To our knowledge, this formulation based on minimizing the classifier’s subdominance [5] across different fairness/performance measures is the first to establish fairness objectives for supervised machine learning by comparing noisy human decisions rather than relying on prescriptive trade-offs or strict constraints.

This thesis builds upon and extends this imitation learning formulation for minimizing subdominance. Unlike the original approach, which used logistic regression to optimize the fairness-aware classifier, this thesis employs neural networks. This transition to neural networks allows for a more robust and flexible model capable of handling complex datasets and achieving better performance in terms of fairness and accuracy.

We leverage the subdominance concept not to identify a target trade-off, as previous work in inverse optimal control does to estimate a cost function, but to directly optimize our classifier. The neural network architecture provides a significant advantage in modeling non-linear relationships and capturing intricate patterns within the data, which logistic regression might miss. This results in a more effective and fair decision-making process.

We perform experiments on the Adult datasets, evaluating accuracy as a measure of performance and three conflicting fairness definitions: Demographic Parity [1], Equalized Odds [2], and Predictive Rate Parity [6]. Our approach demonstrates high levels of superhuman performance that improve significantly with increased reference decision noise and outperforms other methods that focus on more limited fairness-performance objectives.

By incorporating neural networks, this thesis enhances the capability of the fairness-aware classifier, allowing it to handle more complex decision-making tasks. This represents a significant step forward in the development of data-driven, fair, and effective machine learning models, moving beyond the limitations of previous methods that relied on simpler models like logistic regression.

CHAPTER 2

RELATED WORK

(This chapter is based on a paper published as “Superhuman Fairness” (Memarrast et al., 2023b) in the International Conference on Machine Learning 40 (ICML 2023))

2.1 Group Fairness Measures

Group fairness measures are mainly determined using confusion matrix statistics. These statistics rely on ground truth labels $y_i \in \{0, 1\}$ and classifier predictions $\hat{y}_i \in \{0, 1\}$ generated from inputs $x_i \in \mathbb{R}^M$ for examples belonging to different protected groups (e.g., $a_i \in \{0, 1\}$).

In this thesis, we concentrate on three widely recognized fairness properties:

- **Demographic Parity (DP)** requires that the rate of positive outcomes is the same across different protected groups. This can be formally defined as:

$$P(\hat{Y} = 1 \mid A = 1) = P(\hat{Y} = 1 \mid A = 0) \quad (2.1)$$

where \hat{Y} represents the predicted outcome, and A represents the group membership [1].

- **Equalized Odds (EqOdds)** requires that both the true positive rates and false positive rates are equal across groups. This is expressed as:

$$P(\hat{Y} = 1 \mid Y = y, A = 1) = P(\hat{Y} = 1 \mid Y = y, A = 0), \quad \forall y \in \{0, 1\} \quad (2.2)$$

where Y is the actual outcome [2].

- **Predictive Rate Parity (PRP)** ensures that the positive predictive value (the probability that a positive prediction is correct) and the negative predictive value are equal across groups.

It can be formulated as:

$$P(Y = 1 \mid A = 1, \hat{Y} = \hat{y}) = P(Y = 1 \mid A = 0, \hat{Y} = \hat{y}), \quad \forall \hat{y} \in \{0, 1\} \quad (2.3)$$

[6].

Violations of these fairness properties can be quantified using specific difference measures:

- **Demographic Parity Difference (D.DP)**

$$D.DP(\hat{y}, a) = \left| \frac{\sum_{i=1}^N I[\hat{y}_i = 1, a_i = 1]}{\sum_{i=1}^N I[a_i = 1]} - \frac{\sum_{i=1}^N I[\hat{y}_i = 1, a_i = 0]}{\sum_{i=1}^N I[a_i = 0]} \right| \quad (2.4)$$

- **Equalized Odds Difference (D.EqOdds)**

$$D.EqOdds(\hat{y}, y, a) = \max_{y \in \{0,1\}} \left| \frac{\sum_{i=1}^N I[\hat{y}_i = 1, y_i = y, a_i = 1]}{\sum_{i=1}^N I[a_i = 1, y_i = y]} - \frac{\sum_{i=1}^N I[\hat{y}_i = 1, y_i = y, a_i = 0]}{\sum_{i=1}^N I[a_i = 0, y_i = y]} \right| \quad (2.5)$$

- **Predictive Rate Parity Difference (D.PRPP)**

$$D.PRPP(\hat{y}, y, a) = \max_{y \in \{0,1\}} \left| \frac{\sum_{i=1}^N I[y_i = 1, \hat{y}_i = y, a_i = 1]}{\sum_{i=1}^N I[a_i = 1, \hat{y}_i = y]} - \frac{\sum_{i=1}^N I[y_i = 1, \hat{y}_i = y, a_i = 0]}{\sum_{i=1}^N I[a_i = 0, \hat{y}_i = y]} \right| \quad (2.6)$$

These measures help in assessing and quantifying the extent to which a machine learning model’s decisions are fair across different groups.

2.2 Performance-Fairness Trade-offs

Numerous algorithms for fair classification have emerged recently, focusing on one or two fairness measures [2, 7–9]. Often, predictive performance and fairness are at odds, meaning improving one can degrade the other [10]. While achieving multiple fairness objectives is appealing, it often results in significant performance loss or infeasibility [3].

To address this, many methods aim to optimize parameters θ of a classifier P_θ to balance performance and fairness [2, 10–14]. Hsu et al. [15] proposed an optimization framework to address three conflicting fairness measures—demographic parity, equalized odds, and predictive rate parity:

$$\min_{\theta} \mathbb{E}_{\hat{y} \sim P_\theta} [\text{loss}(\hat{y}, y) + \alpha_{\text{DP}} \cdot D_{\text{DP}}(\hat{y}, a) + \alpha_{\text{Odds}} \cdot D_{\text{EqOdds}}(\hat{y}, y, a) + \alpha_{\text{PRP}} \cdot D_{\text{PRP}}(\hat{y}, y, a)]$$

2.3 Preference Elicitation & Imitation Learning

Preference Elicitation

Preference elicitation, as described by [16], is a natural method for identifying desirable performance-fairness trade-offs. This approach typically involves querying users for their pairwise preferences on a series of option pairs to determine their utilities for various option characteristics.

This method has been adapted for fairness measure elicitation [17], enabling efficient learning of both linear and non-linear measures from limited and noisy preference feedback.

In contexts where multiple stakeholders make decisions [18] rather than a single individual, preference elicitation may not be very informative. While preferences from each stakeholder can be elicited, determining how these preferences should be prioritized to achieve joint outcomes remains unclear without strong additional assumptions about the decision-making process, such as outcomes determined by majority vote [19].

Imitation Learning

Imitation learning [20] is a type of supervised machine learning designed to create a general-use policy $\hat{\pi}$ based on demonstrated sequences of states and actions, $\tilde{\xi} = (\tilde{s}_1, \tilde{a}_1, \tilde{s}_2, \dots, \tilde{s}_T)$.

Inverse reinforcement learning methods [21,22] seek to explain these demonstrated trajectories as the result of (near-) optimal policies under an estimated cost or reward function. Feature matching [21] is essential in these methods, ensuring that if the expected feature counts match, the estimated policy $\hat{\pi}$ will have an expected cost equal to the average of the demonstrated trajectories:

$$\begin{aligned} \mathbb{E}_{\xi \sim \hat{\pi}} [f_k(\xi)] &= \frac{1}{N} \sum_{i=1}^N f_k(\tilde{\xi}_i), \forall k \\ \implies \mathbb{E}_{\xi \sim \hat{\pi}} [\text{cost}_{\tilde{w}}(\xi)] &= \frac{1}{N} \sum_{i=1}^N \text{cost}_{\tilde{w}}(\tilde{\xi}_i), \end{aligned} \tag{2.7}$$

where $f_k(\xi) = \sum_{s_t \in \xi} f_k(s_t)$.

[23] seeks to surpass the provided demonstrations given that the signs of the unknown cost function are known, specifically when $\tilde{w}_k \geq 0$, by making the inequality,

$$\mathbb{E}_{\xi \sim \pi} [f_k(\xi)] \leq \frac{1}{N} \sum_{i=1}^N f_k(\tilde{\xi}_i), \forall k, \quad (2.8)$$

strict for at least one feature. Subdominance minimization [24] aims to create trajectories that exceed each demonstration by a certain margin:

$$f_k(\xi) + m_k \leq f_k(\tilde{\xi}_i), \forall i, k, \quad (2.9)$$

Assuming the cost weight signs are known, this method seeks to outperform demonstrations. However, as this is often impractical, the approach instead focuses on minimizing subdominance, which quantifies the α -weighted violation of this inequality:

$$\text{subdom}_\alpha(\xi, \tilde{\xi}) \triangleq \sum_k \left[\alpha_k \left(f_k(\xi) - f_k(\tilde{\xi}) \right) + 1 \right]_+, \quad (2.10)$$

where $[f(x)]_+ \triangleq \max(f(x), 0)$ is the hinge function, and the per-feature margin is reparameterized as α_k^{-1} . Previous work [24] has applied subdominance minimization along with inverse optimal control:

$$\min_w \min_\alpha \sum_{i=1}^N \sum_{k=1}^K \text{subdom}_\alpha(\xi^*(w), \tilde{\xi}_i), \quad (2.11)$$

$$\text{where: } \xi^*(w) = \arg \min_{\xi} \sum_k w_k f_k(\xi) \quad (2.12)$$

to learn the cost function parameters \mathbf{w} for the optimal trajectory $\xi^*(\mathbf{w})$ that minimizes subdominance.

2.3.1 Subdominance Minimization for Improved Fairness-Aware Classification

In the paper "Superhuman Fairness" [25], the authors approach fair classification from an imitation learning perspective. The main objective is to create a fairness-aware classifier that consistently outperforms human-provided reference decisions in both performance and fairness on unseen data, thereby ensuring guarantees for all stakeholders.

2.3.1.1 Superhumanness and Subdominance

Consider the reference decisions $\tilde{\mathbf{y}} = \{\tilde{y}_j\}_{j=1}^M$ drawn from an unknown human decision-making process or baseline method \tilde{P} , applied to a set of M items, $\mathbf{X}_{M \times L} = \{\mathbf{x}_j\}_{j=1}^M$, where L is the number of attributes in each item \mathbf{x}_j . Group membership attributes a_m indicate the group to which item m belongs.

The predictive performance and fairness of decisions $\hat{\mathbf{y}}$ for each item are assessed using ground truth \mathbf{y} and group membership \mathbf{a} with a set of predictive loss and unfairness measures $\{f_k(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a})\}$. Higher values of these measures are less favorable. Ideally, these measures should cover all stakeholder preference functions.

Definition 3.1: A fairness-aware classifier is considered γ -superhuman if its decisions $\hat{\mathbf{y}}$ satisfy:

$$P(f(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a}) \preceq f(\tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a})) \geq \gamma.$$

Maximizing γ directly is challenging due to the discontinuity of Pareto dominance (\preceq). The subdominance serves as a convex upper bound for non-dominance in each measure $\{f_k\}$ and on $1 - \gamma$ in aggregate:

$$\text{subdom}_{\alpha_k}^k(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a}) \triangleq [\alpha_k(f_k(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a}) - f_k(\tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a})) + 1]_+,$$

$$\text{subdom}_\alpha(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a}) \triangleq \sum_k \text{subdom}_{\alpha_k}^k(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a}).$$

Given N vectors of reference decisions as demonstrations $\tilde{\mathbf{Y}} = \{\tilde{\mathbf{y}}_i\}_{i=1}^N$, the subdominance for decision vector $\hat{\mathbf{y}}$ with respect to the set of demonstrations is:

$$\text{subdom}_\alpha(\hat{\mathbf{y}}, \tilde{\mathbf{Y}}, \mathbf{y}, \mathbf{a}) = \frac{1}{N} \sum_{\tilde{\mathbf{y}} \in \tilde{\mathbf{Y}}} \text{subdom}_\alpha(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a}).$$

2.3.1.2 Performance-Fairness Subdominance Minimization

We consider probabilistic predictors $P_\theta : \mathbf{X}^M \rightarrow \Delta \mathbf{Y}^M$ that generate structured predictions over a set of items while also being capable of making conditionally independent decisions for each item.

Definition 3.2: The minimally subdominant fairness-aware classifier \hat{P}_θ has model parameters θ chosen by:

$$\arg \min_{\theta} \min_{\alpha \succeq 0} \mathbb{E}_{\hat{\mathbf{y}} | \mathbf{X} \sim P_\theta} \left[\text{subdom}_\alpha(\hat{\mathbf{y}}, \tilde{\mathbf{Y}}, \mathbf{y}, \mathbf{a}) \right] + \lambda \|\alpha\|_1.$$

Hinge loss slopes $\alpha \triangleq \{\alpha_k\}_{k=1}^K$ are also learned from the data during training. For the subdominance of the k -th measure, α_k indicates the degree of sensitivity to underperformance in that measure. Higher α_k values reduce underperformance on that measure, minimizing overall subdominance more effectively.

The optimization of θ and α differs from single-level support vector machine optimization, which is a convex optimization problem. Instead, subdominance is a quasi-convex function, implying no local optima in terms of realized predictive performance/fairness measures.

Theorem 3.3: The α_k -minimized subdominance,

$$\sum_k \Gamma_k(\hat{\mathbf{y}}, \tilde{\mathbf{Y}}, \mathbf{y}, \mathbf{a}) \triangleq \min_{\alpha_k \geq 0} \left(\text{subdom}_{\alpha_k}^k(\hat{\mathbf{y}}, \tilde{\mathbf{Y}}, \mathbf{y}, \mathbf{a}) + \lambda_k \alpha_k \right),$$

is a quasiconvex function in terms of the set of measures $\{f_k(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a})\}$.

The gradient of the expected subdominance under \hat{P}_θ with respect to the set of reference decisions $\{\tilde{\mathbf{y}}_i\}_{i=1}^N$ is:

$$\nabla_\theta \mathbb{E}_{\hat{\mathbf{y}}|\mathbf{X} \sim \hat{P}_\theta} \left[\sum_k \Gamma_k(\hat{\mathbf{y}}, \tilde{\mathbf{Y}}, \mathbf{y}, \mathbf{a}) \right] = \mathbb{E}_{\hat{\mathbf{y}}|\mathbf{X} \sim \hat{P}_\theta} \left[\left(\sum_k \Gamma_k(\hat{\mathbf{y}}, \tilde{\mathbf{Y}}, \mathbf{y}, \mathbf{a}) \right) \nabla_\theta \log \hat{P}_\theta(\hat{\mathbf{y}}|\mathbf{X}) \right].$$

Using gradient descent, the model weights θ are updated iteratively based on a set of sampled predictions $\hat{\mathbf{y}} \in \hat{\mathbf{Y}}$ from the model \hat{P}_θ :

$$\theta \leftarrow \theta + \eta \left(\sum_{\hat{\mathbf{y}} \in \hat{\mathbf{Y}}} \left(\sum_k \Gamma_k(\hat{\mathbf{y}}, \tilde{\mathbf{Y}}, \mathbf{y}, \mathbf{a}) \right) \nabla_\theta \log \hat{P}_\theta(\hat{\mathbf{y}}|\mathbf{X}) \right).$$

The algorithm for training the model involves initializing θ , sampling model predictions, sorting reference decisions by measure values, computing α , and iterating until convergence.

Algorithm 1: Subdominance policy gradient optimization

Draw N set of reference decisions $\{\tilde{\mathbf{y}}_i\}_{i=1}^N$ from a human decision-maker or baseline method $\hat{\mathbb{P}}$. Initialize: $\theta \leftarrow \theta_0$;
while θ *not converged* **do**
 Sample model predictions $\{\hat{\mathbf{y}}_i\}_{i=1}^N$ from $\hat{\mathbb{P}}_{\theta}(\cdot|\mathbf{X}_i)$ for the matching items used in reference decisions $\{\tilde{\mathbf{y}}_i\}_{i=1}^N$;
 for $k \in \{1, \dots, K\}$ **do**
 Sort reference decisions $\{\tilde{\mathbf{y}}_i\}_{i=1}^N$ in ascending order by k^{th} measure value $f_k(\tilde{\mathbf{y}}_i)$: $\{\tilde{\mathbf{y}}^{(j)}\}_{j=1}^N$;
 Compute $\alpha_k^{(j)} = \frac{1}{f_k(\tilde{\mathbf{y}}^{(j)}) - f_k(\tilde{\mathbf{y}}^{(j-1)})}$;
 $\alpha_k = \underset{\alpha_k^{(m)}}{\text{argmin}} m$ such that
 $f_k(\hat{\mathbf{y}}) + \lambda \leq \frac{1}{m} \sum_{j=1}^m f_k(\tilde{\mathbf{y}}^{(j)})$;
 Compute $\Gamma_k(\hat{\mathbf{y}}_i, \tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a})$;
 $\theta \leftarrow \theta + \frac{\eta}{N} \sum_i \left(\sum_k \Gamma_k(\hat{\mathbf{y}}_i, \tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a}) \right) \nabla_{\theta} \log \hat{\mathbb{P}}_{\theta}(\hat{\mathbf{y}}_i|\mathbf{X}_i)$;

Relationship to Thesis Contributions

Building on this imitation learning formulation for fairness, this thesis extends this approach to incorporate more complex models, such as neural networks, to enhance performance. Logistic regression models used previously within this approach [25] are limited to linear decision boundaries. Multi-layer neural networks allow much more complicated non-linear decision boundaries to be learned. These have proven to be highly beneficial in many domains, ranging

from computer vision to natural language processing. This thesis extends the imitation learning approach for fairness in the use of neural networks. This allows for improved predictive accuracy and fairness performance, further pushing the boundaries of what can be achieved in fairness-aware machine learning. In addition to the neural network, I also attempt to include additional information for the neural network, including positive and negative predictions of each sensitive group, to improve fairness metrics.

in the International Conference on Machine Learning 40 (ICML 2023)

CHAPTER 3

APPROACH

(A portion of this chapter (3.2 Mathematical Formulation) is based on a paper published as “Superhuman Fairness” (Memarrast et al., 2023b))

3.1 Background of Neural Networks

Neural network is a model of machine learning that have gained significant popularity due to their ability to model complex, non-linear relationships [26]. They consist of interconnected nodes, or neurons, arranged in layers. The fundamental building block of a neural network is the perceptron, which computes a weighted sum of its inputs and passes the result through a non-linear activation function [27].

3.1.1 Architecture of Neural Networks

A typical neural network architecture includes an input layer, several hidden layers, and an output layer. Each of these layers is made up of numerous neurons, which are completely interconnected with the neurons in the following layer. The connections between neurons are characterized by weights, which are adjusted during training to minimize a predefined loss function [28].

3.1.1.1 Feedforward Neural Networks

The most basic type of neural network is the feedforward neural network (FNN), where the data flows directly from the input layer through the hidden layers to the output layer [27].

The network does not have cycles or loops, making it straightforward and suitable for many classification and regression tasks [28].

3.1.1.2 Activation Functions

Activation functions bring non-linearity to neural networks, allowing them to capture and represent more intricate patterns. Popular activation functions such as the sigmoid function, hyperbolic tangent (\tanh), and Rectified Linear Unit (ReLU) [29] each possess unique features that influence the training process and the network's overall performance.

3.1.2 Training Neural Networks

Training a neural network is the process of adjusting the weights of the parameters based on the error between the predicted output and the actual target. This process is typically performed using a method called backpropagation [30], which calculates the gradient of the loss function for each weight by using the chain rule. Optimizers such as Stochastic Gradient Descent (SGD), Adam, and RMSprop [31] are commonly used to update the weights iteratively, reducing the loss function over time.

3.1.3 Applications of Neural Networks

Neural networks have been widely applied in different domains, including speech and image [32], natural language processing [33], and autonomous systems [34]. Their ability to automatically learn features from data without manual featurizing engineering has revolutionized many fields, making them a cornerstone of modern artificial intelligence [26].

In the context of fairness-aware classification, neural networks provide a flexible and powerful framework for modeling complex relationships between features and the target variable. By

incorporating fairness constraints into the training process, we aim to develop classifiers that not only perform well but also adhere to ethical standards of fairness [2].

3.2 Subdominance Minimization with Neural Networks

Our approach extends the concept of subdominance minimization for fairness-aware classification, as proposed in [25]. While the original method utilizes logistic regression, our work leverages neural networks to improve the flexibility and performance of the fairness-aware classifier. This chapter details the methodology used in our implementation.

We consider probabilistic predictors, that make structured predictions over a set of items. The fairness-aware classifier \hat{P}_θ is parameterized by a neural network with parameters θ . The objective is to minimize the subdominance measure, which serves as a convex upper bound for non-dominance in each performance and fairness measure.

3.2.1 Mathematical Formulation

The subdominance measure for the k -th feature is defined as:

$$\text{subdom}_k^{\alpha_k}(\hat{y}, \tilde{y}, y, a) = [\alpha_k (f_k(\hat{y}, y, a) - f_k(\tilde{y}, y, a)) + 1]_+,$$

and the aggregate subdominance measure is:

$$\text{subdom}_\alpha(\hat{y}, \tilde{y}, y, a) = \sum_k \text{subdom}_k^{\alpha_k}(\hat{y}, \tilde{y}, y, a).$$

Given N vectors of reference decisions, $\tilde{Y} = \{\tilde{y}_i\}_{i=1}^N$, the subdominance for decision vector \hat{y} with respect to the set of demonstrations is:

$$\text{subdom}_\alpha(\hat{y}, \tilde{Y}, y, a) = \frac{1}{N} \sum_{\tilde{y} \in \tilde{Y}} \text{subdom}_\alpha(\hat{y}, \tilde{y}, y, a).$$

The optimization objective is to minimize the expected subdominance under the model P_θ :

$$\arg \min_{\theta} \min_{\alpha \succeq 0} \mathbb{E}_{\hat{y}|X \sim P_\theta} \left[\text{subdom}_\alpha(\hat{y}, \tilde{Y}, y, a) \right] + \lambda \|\alpha\|_1.$$

3.2.2 Training Procedure

The training procedure involves iterative optimization of the neural network parameters θ and the hinge loss slopes $\alpha = \{\alpha_k\}_{k=1}^K$. The algorithm follows these steps:

1. Initialize the model parameters θ and hinge loss slopes α .
2. For each iteration:
 - (a) Extract the label vector Y and the feature matrix X from the training data.
 - (b) Convert X to a tensor for input into the neural network.
 - (c) Sample superhuman decisions and update the sample matrix.
 - (d) Index the samples based on the demonstration list and calculate the sample loss for each feature.
 - (e) Shuffle the demonstration indices and iterate through them:
 - i. For each demonstration, compute the predictions \hat{y} using the neural network.

- ii. Calculate the subdominance measure for each feature and update the loss.
 - iii. Backpropagate the loss and update the model parameters using gradient descent.
3. After convergence, compute the optimal α values.

CHAPTER 4

EXPERIMENTS AND RESULTS

4.1 Neural Network Architecture

In our experiments, we designed a neural network using the PyTorch library to assess the performance and fairness of our classifier. The model consists of three fully connected (FC) layers, each followed by ReLU activation functions. Each fully connected layers is followed by a ReLU activation function. The final layer of the network is used to compressed down to two outputs, which correspond to the classes to be predicted. The output from this layer is directed into a softmax function where it returns a vector of probability scores of the classes.

The figure below visualizes the structure of our neural network, illustrating the sequence of layers and activations used to achieve classification.

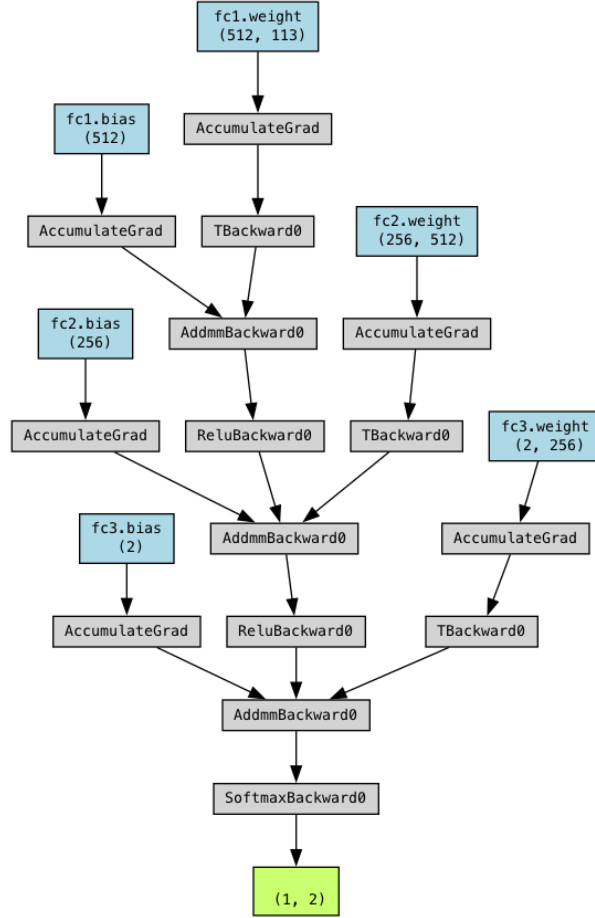


Figure 2: Visualization of the neural network structure

4.2 Training and Testing Dataset Construction

In my thesis, I focus exclusively on the UCI Adult dataset [35] to investigate decision-making processes. This dataset comprises 45,222 entries and is utilized to predict whether a household's

income exceeds \$50K/year based on census data, with gender serving as the criterion for group membership.

The dataset is initially partitioned into two disjoint subsets: a training set (*train-all*) and a test set (*test-all*), both of equal size. The test set (*test-all*) is completely excluded from the training process and is used exclusively for evaluation purposes. For each demonstration, which involves generating a vector of reference decisions, the *train-all* set is further divided randomly into two equal-sized, disjoint subsets: a training set (*train-demo*) and a test set (*test-demo*).

Using the data, we equip existing fairness-aware methods with labeled *train-demo* data and unlabeled *test-demo* data to produce decisions on the *test-demo* data, referred to as demonstrations \tilde{y} . Specifically, we apply the post-processing approach in [36], which aims to minimize both prediction error and demographic disparities using the demographic parity criterion as the fairness constraint. Demonstrations for the Adult dataset are generated using this method.

We repeat the partitioning of the *train-all* dataset $N = 50$ times to create randomized partitions of *train-demo* and *test-demo*, subsequently generating a set of demonstrations $\{\tilde{y}_i\}_{i=1}^{50}$.

4.3 Results

We run the experiments three times with randomization to ensure the performance of the model. We also experiment with neural networks of different sizes. Below is a summary of the results.

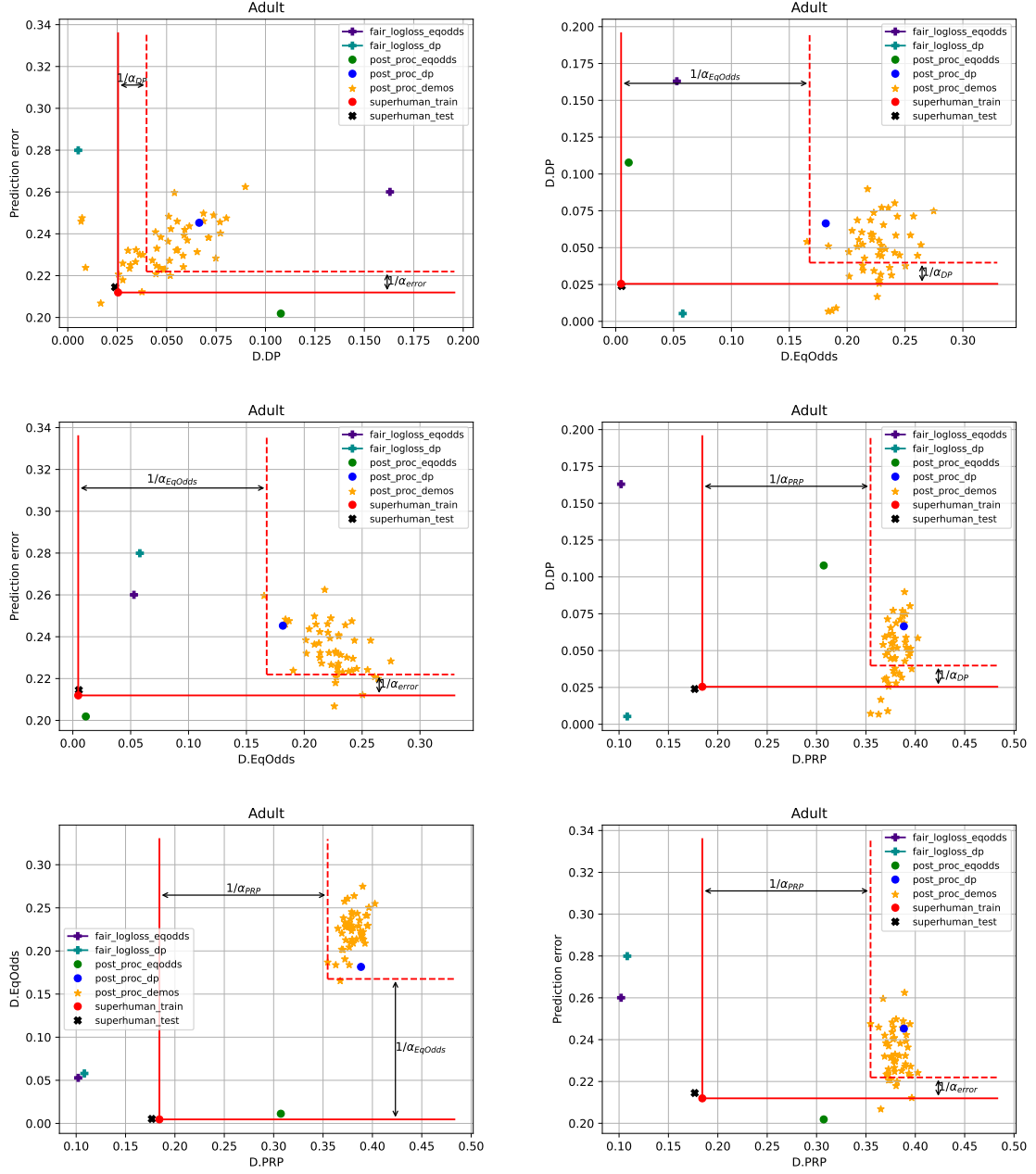


Figure 3: Experimental results on the Adult dataset.

TABLE I: Experimental results on noise-free datasets

Method \ Dataset	Adult			
	Prediction error	DP diff	EqOdds diff	PRP diff
original_logistic_regression	0.220370	0.018258	0.015092	0.178658
large_nn_w_counts	0.214597	0.023829	0.005396	0.176662
large_nn_wo_counts	0.213982	0.021831	0.011610	0.178390
small_nn_wo_counts	0.205124	0.031476	0.009779	0.171306
small_nn_w_counts	0.203609	0.033652	0.014163	0.169798
eval_pp_dp	0.245076	0.065161	0.178313	0.388746
eval_pp_eq_odds	0.202258	0.107372	0.012040	0.307997
eval_fairll_dp	0.281380	0.005055	0.064160	0.107310
eval_fairll_eqodds	0.255038	0.150623	0.042668	0.109555
eval_fairll_eqopp	0.223851	0.180576	0.156833	0.094255
eval_MFOpt	0.195696	0.063152	0.077549	0.209199

In the comparison of various methods on the **Adult** dataset, the approach of **large_nn_w_counts** exhibits competitive performance across different fairness metrics, compared to alternative methods evaluated in the study. Specifically, **large_nn_w_counts** achieves a prediction error of 0.214597, which is slightly above the best-performing **eval_MFOpt** method at 0.195696, indicating a marginally lower accuracy but still maintaining a high standard of predictive performance.

In terms of fairness metrics, the **large_nn_w_counts** method demonstrates a notable performance in the 'EqOdds diff' metric with a value of 0.005396, which is the lowest among all methods and highlights capability to ensure equal odds between different demographic groups.

However, the **large_nn_w_counts** method shows slightly higher values in 'DP diff' at 0.023829 compared to the lowest **eval_fairll_dp** at 0.005055, indicating a minor compromise in demo-

graphic parity. Similarly, in the 'PRP diff' metric, while `large_nn_w_counts` scores 0.176662, it is slightly outperformed by `eval_fairll_eqopp` which achieves the lowest value of 0.094255.

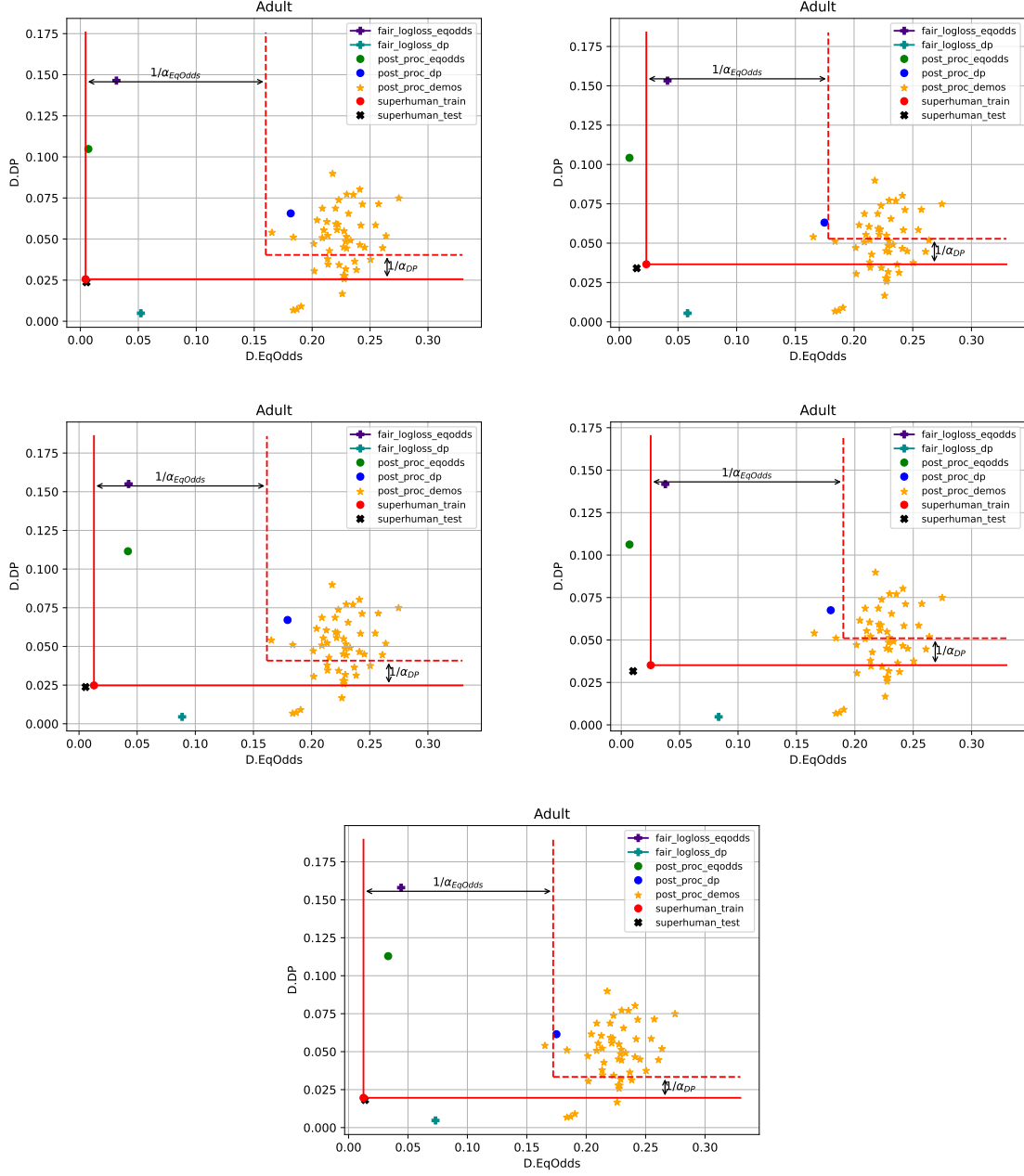
4.3.1 Comparision with Original Method

A direct comparison between `large_nn_w_counts` and `original_logistic_regression` highlights several key differences in performance metrics. The `large_nn_w_counts` method achieves a lower prediction error of 0.214597, compared to 0.220370 for the original method. For fairness metrics, `large_nn_w_counts` records a 'DP diff' of 0.023829, which is slightly higher than the 0.018258 observed for `original_logistic_regression`. However, `large_nn_w_counts` significantly outperforms `original_logistic_regression` in the 'EqOdds diff' metric with a notable lower score of 0.005396 versus 0.015092. This demonstrates a substantial enhancement in maintaining equal odds between demographic groups. In the 'PRP diff', both methods perform similarly, with `large_nn_w_counts` slightly outperforming at 0.176662 compared to 0.178658 for `original_logistic_regression`. Overall, the metrics indicate that `large_nn_w_counts` not only improves upon the accuracy of the `original_logistic_regression` but also offers significant advancements in fairness, particularly in equalizing odds across demographic groups.

Below are some additional graphs for experiments:

TABLE II: Mean and Standard Deviation of Evaluation Methods

Method	ZeroOne (m, std)	DP Diff (m, std)	EO Diff (m, std)	PRP Diff (m, std)
small_nn_w_counts	0.203609, 0.000272	0.033652, 0.000371	0.014163, 0.000666	0.169798, 0.000239
small_nn_wo_counts	0.205124, 0.000047	0.031476, 0.000141	0.009779, 0.000350	0.171306, 0.000072
large_nn_w_counts	0.214597, 0.000103	0.023829, 0.000122	0.005396, 0.000261	0.176662, 0.000055
large_nn_wo_counts	0.213982, 0.000024	0.021831, 0.000036	0.011610, 0.000000	0.178390, 0.000010
eval_pp_dp	0.245076, 0.000681	0.065161, 0.002171	0.178313, 0.001786	0.388746, 0.001255
eval_pp_eq_odds	0.202258, 0.000925	0.107372, 0.001736	0.012040, 0.007682	0.307997, 0.004493
eval_fairll_dp	0.281380, 0.003625	0.005055, 0.000281	0.064160, 0.016658	0.107310, 0.004703
eval_fairll_eqodds	0.255038, 0.004884	0.150623, 0.008970	0.042668, 0.017071	0.109555, 0.005708
eval_fairll_eqopp	0.223851, 0.000481	0.180576, 0.000995	0.156833, 0.003135	0.094255, 0.000316

Figure 4: Experimental results on the **Adult** dataset. EqOdds vs DP

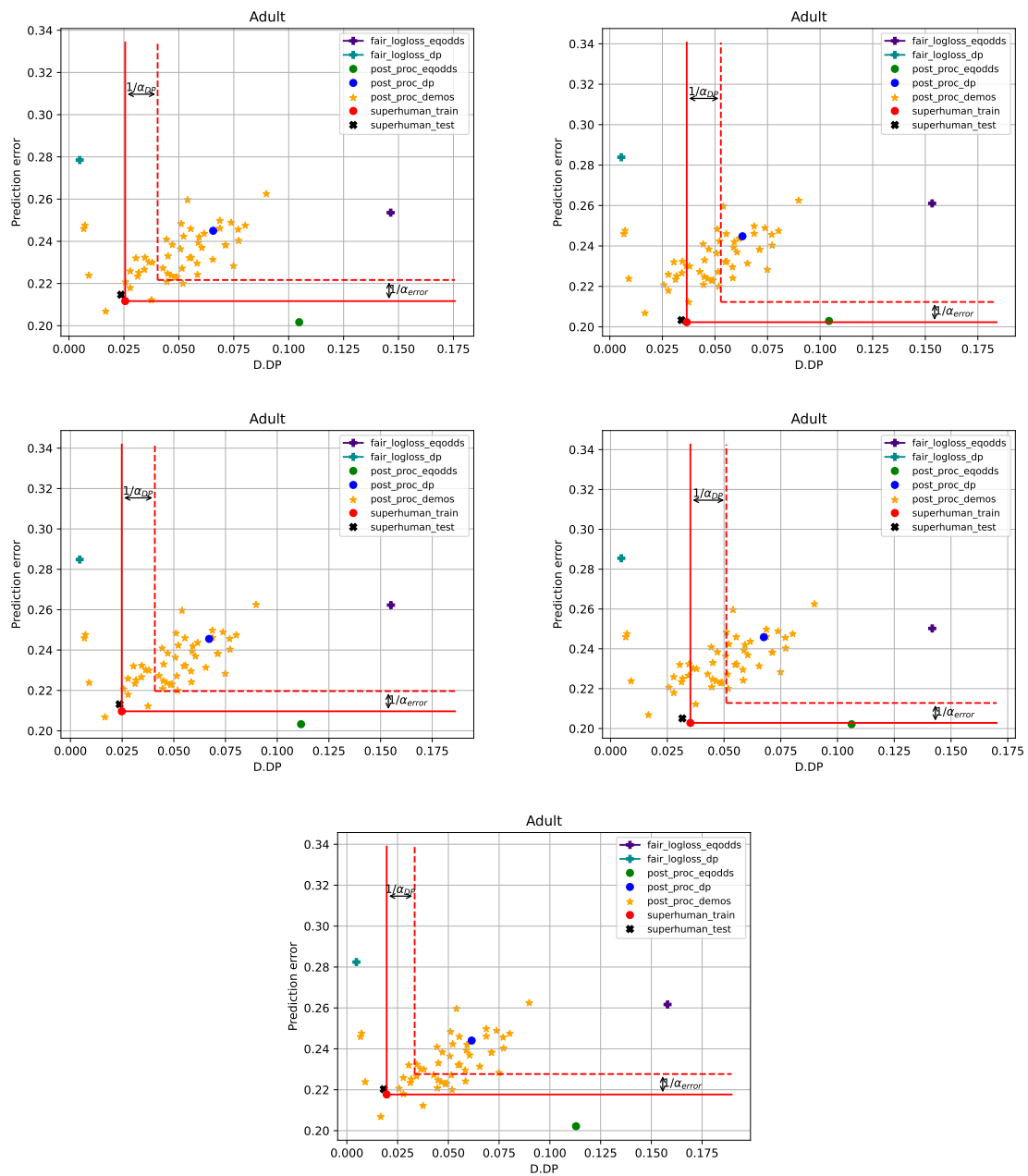


Figure 5: Experimental results on the **Adult** dataset. DP vs Error

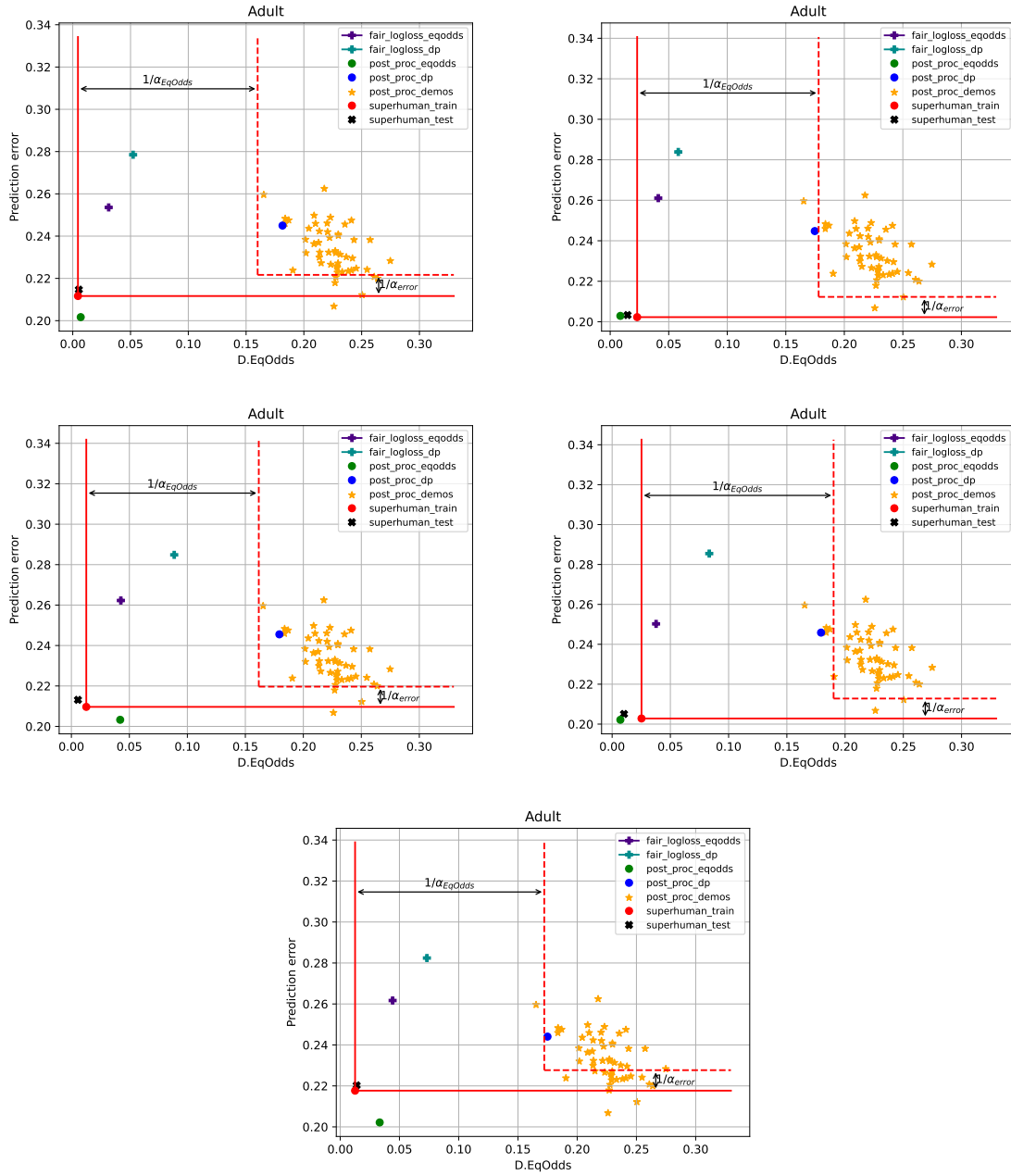


Figure 6: Experimental results on the **Adult** dataset. DP vs Error

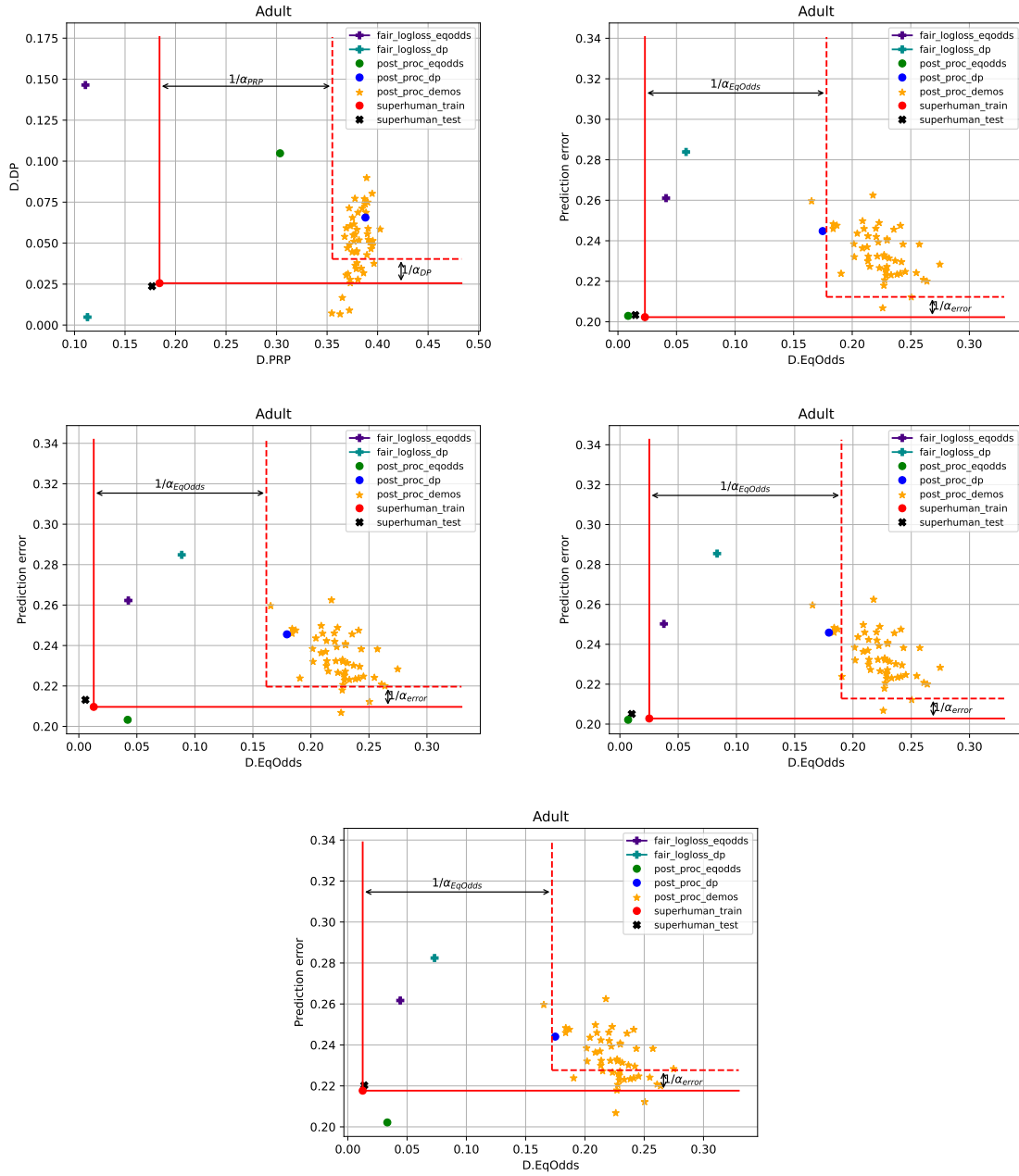
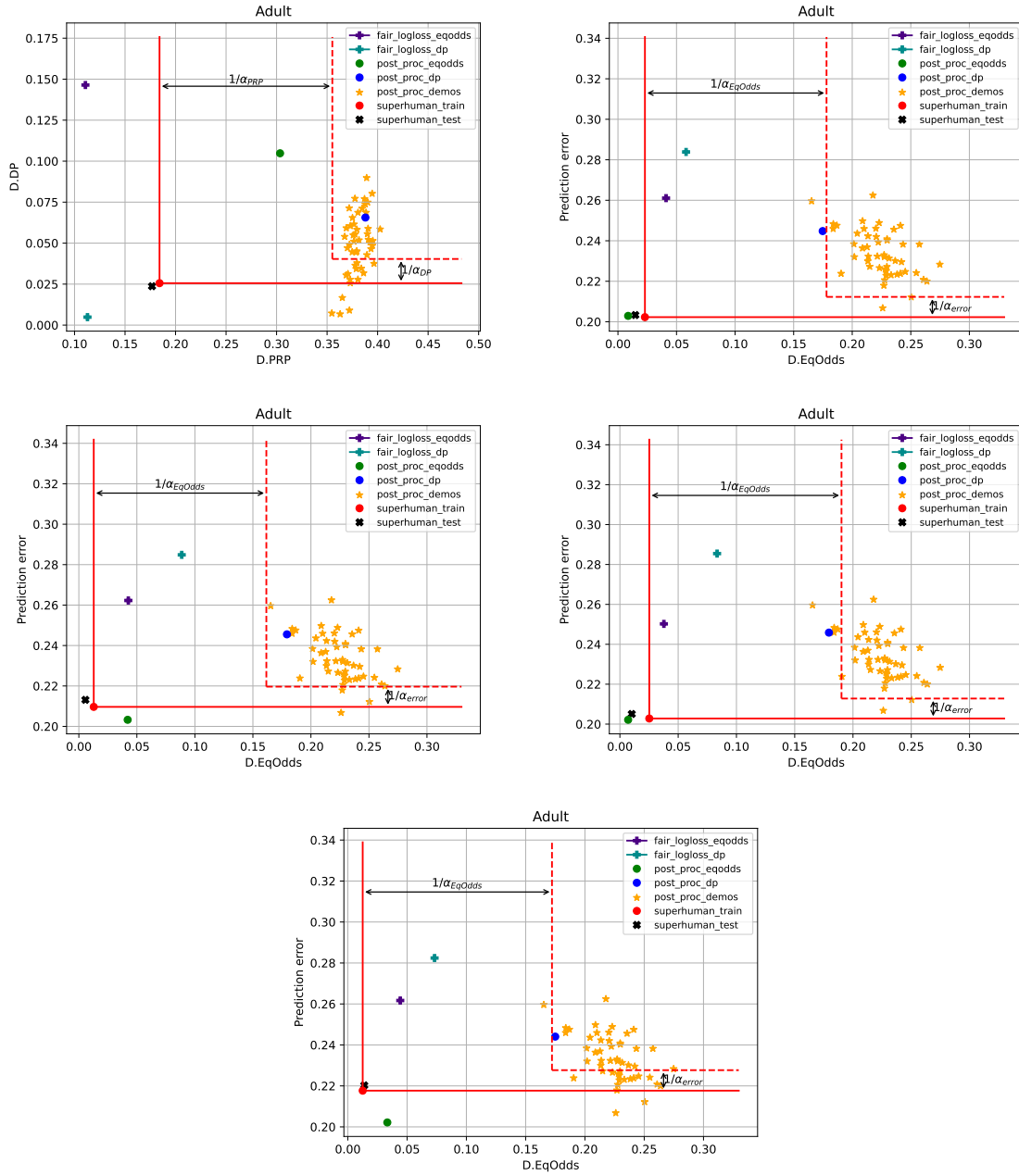
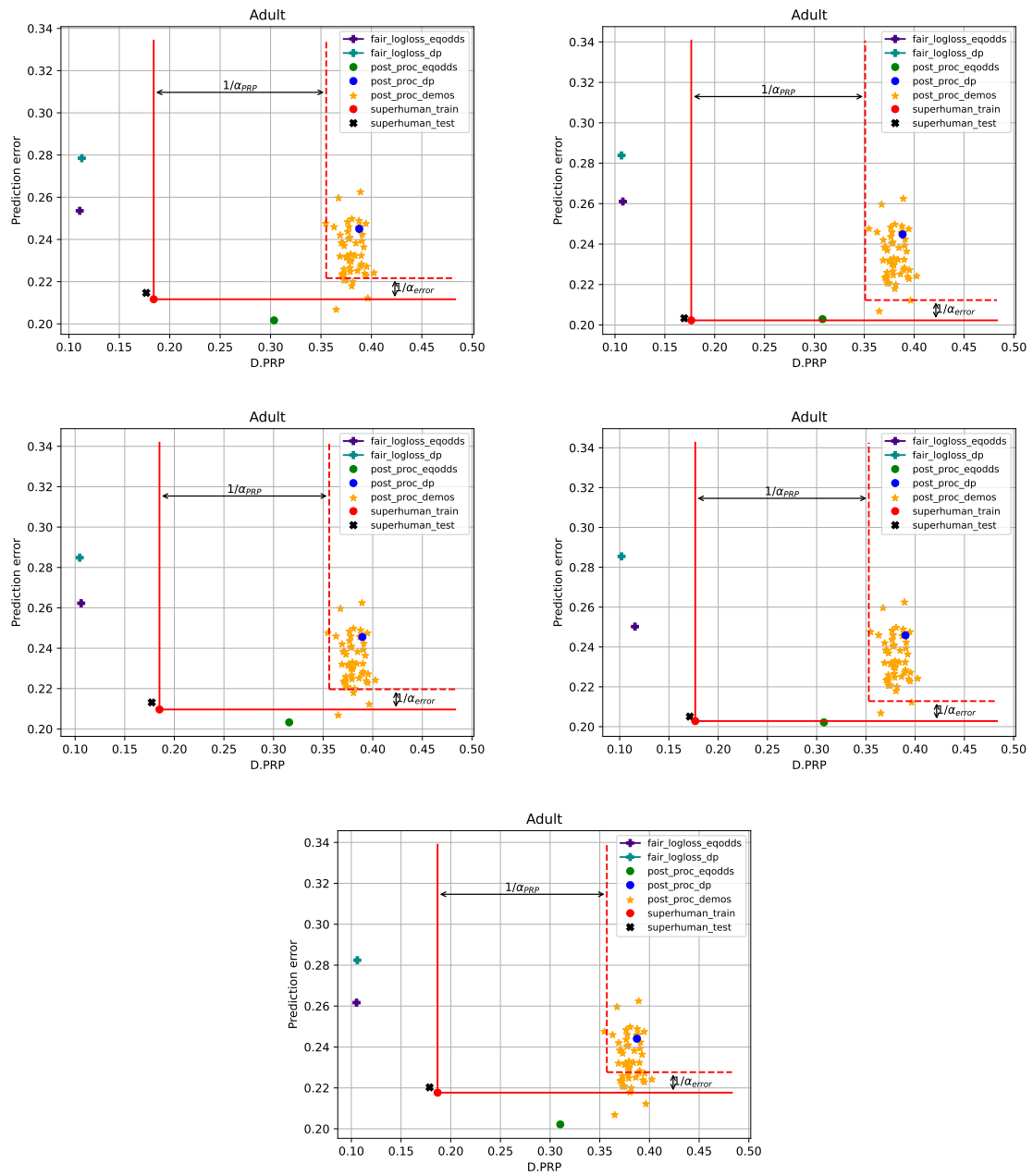


Figure 7: Experimental results on the Adult dataset. DP vs Error

Figure 8: Experimental results on the **Adult** dataset. EqOdds vs Error

Figure 9: Experimental results on the **Adult** dataset. PRP vs Error

CHAPTER 5

CONCLUSION

This project has demonstrated that the neural network model, outperforms traditional logistic regression in key performance areas. The neural network not only achieved lower prediction errors but also showed significant improvements in fairness metrics, especially in 'EqOdds diff'. This evidence strongly supports the adoption of neural networks to more effectively manage the dual demands of accuracy and fairness in predictive models.

Future research should aim to further enhance neural network architectures, with a specific focus on improving fairness metrics without sacrificing accuracy. Experimentation with different optimizers and exploring various network structures could provide insights into optimizing model performance and fairness.

APPENDIX

COPYRIGHT PERMISSIONS

The Proceedings of Machine Learning Research is a series that publishes machine learning research papers presented at conferences and workshops. Each volume is separately titled and associated with a particular workshop or conference and will be published online on the PMLR website. Authors will retain copyright and individual volume editors are free to make additional hardcopy publishing arrangements (see for example the Challenges in Machine Learning series which includes free PDFs and low cost hard copies), but JMLR will not produce hardcopies of these volumes.

Publication Agreement

This is a publication agreement¹ (“this agreement”) regarding a written manuscript currently entitled

Superhuman Fairness

(“the article”) to be published in PMLR (“the proceedings”). The parties to this Agreement are:

Omid Memarrast and PMLR

(name of corresponding author who signs on behalf of any other authors, collectively “you”) and PMLR, (“the publisher”).


1. By signing this form, you warrant that you are signing on behalf of all authors of the article, and that you have the authority to act as their agent for the purpose of entering into this agreement.
2. You hereby grant a Creative Commons copyright license in the article to the general public, in particular a Creative Commons Attribution 4.0 International License, which is incorporated herein by reference and is further specified at <http://creativecommons.org/licenses/by/4.0/legalcode> (human readable summary at <http://creativecommons.org/licenses/by/4.0>).
3. You agree to require that a citation to the original publication of the article in the proceedings as well as a hyperlink to the PMLR web site linking to the original paper be included in any attribution statement satisfying the attribution requirement of the Creative Commons license of paragraph 2.
4. You retain ownership of all rights under copyright in all versions of the article, and all rights not expressly granted in this agreement.
5. To the extent that any edits made by the publisher to make the article suitable for publication in the proceedings amount to copyrightable works of authorship, the publisher hereby assigns all right, title, and interest in such edits to you. The publisher agrees to verify with you any such edits that are substantive. You agree that the license of paragraph 2 covers such edits.

¹The language of this publication agreement is based on Stuart Shieber’s model open-access journal publication agreement, version 1.2, available at <http://bit.ly/1m9UsNt>.

6. You further warrant that:

1. The article is original, has not been formally published in any other peerreviewed journal or in a book or edited collection, and is not under consideration for any such publication.
 2. You are the sole author(s) of the article, and that you have a complete and unencumbered right to make the grants you make.
 3. The article does not libel anyone, invade anyone's copyright or otherwise violate any statutory or common law right of anyone, and that you have made all reasonable efforts to ensure the accuracy of any factual information contained in the article. You agree to indemnify the publisher against any claim or action alleging facts which, if true, constitute a breach of any of the foregoing warranties or other provisions of this agreement, as well as against any related damages, losses, liabilities, and expenses incurred by the publisher.
7. This is the entire agreement between you and the publisher, and it may be modified only in writing. It will be governed by the laws of the Commonwealth of Massachusetts. It will bind and benefit our respective assigns and successors in interest, including your heirs. It will terminate if the publisher does not publish, in any medium, the article within one year of the date of your signature.

I HAVE READ AND AGREE FULLY WITH THE TERMS OF THIS AGREEMENT.

- Corresponding Author:
 - Signed: 
 - Date: 05/30/2023

CITED LITERATURE

1. Calders, T., Kamiran, F., and Pechenizkiy, M.: Building classifiers with independency constraints. In 2009 IEEE International Conference on Data Mining Workshops, pages 13–18. IEEE, 2009.
2. Hardt, M., Price, E., and Srebro, N.: Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems, pages 3315–3323, 2016.
3. Kleinberg, J., Mullainathan, S., and Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. In Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS), 2017.
4. Liu, Y. and Vicente, R.: Trade-offs in algorithmic fairness: Theory and practice. arXiv preprint arXiv:2202.05535, 2022.
5. Ziebart, B. D. et al.: Subdominance minimization for fairness-aware imitation learning. In Advances in Neural Information Processing Systems, 2022.
6. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2):153–163, 2017.
7. Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P.: Fairness constraints: A flexible approach for fair classification. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 259–268, 2015.
8. Goel, S., Olah, M. J., and Vohra, R.: Non-discriminatory machine learning through convex fairness criteria. In Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency, pages 129–138, 2018.
9. Aghaei, S. A., Azizi, M. J., and Vayanos, P.: Learning optimal fair policies. arXiv preprint arXiv:1905.10666, 2019.
10. Menon, A. K. and Williamson, R. C.: The cost of fairness in binary classification. In Proceedings of the Conference on Fairness, Accountability, and Transparency, pages 107–118, 2018.

11. Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 35–50. Springer, 2012.
12. Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K.: Classification with fairness constraints: A meta-algorithm with provable guarantees. In Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, pages 319–328, 2019.
13. Martinez, V., Wu, S., Wu, X., and Ren, S.: Minimax pareto fairness: A multi-objective perspective. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 5417–5424, 2020.
14. Rezaei, P., Roig-Solvas, B., Tarzanagh, D. A., Michielssen, E., Prabhu, P., and Marzbanrad, E.: Fairness in supervised learning: An information theory perspective. Entropy, 22(1):65, 2020.
15. Hsu, T., Xu, H., and Zhou, H.: A multi-objective approach to fair classification. arXiv preprint arXiv:2203.11543, 2022.
16. Chen, L. and Pu, P.: A survey of preference elicitation methods. ACM Transactions on Interactive Intelligent Systems (TiiS), 5(4):1–19, 2015.
17. Hiranandani, G., Guo, W., and Vaughan, J. W.: Fair active learning. Advances in Neural Information Processing Systems, 33:12251–12262, 2020.
18. Donaldson, T. and Preston, L. E.: The stakeholder theory of the corporation: Concepts, evidence, and implications. Academy of Management Briarcliff Manor, NY 10510, 1995.
19. Dowling, G. R. and Moran, P.: A framework for stakeholder engagement and sustainable development in mnes. Journal of International Business Studies, 47(6):730–760, 2016.
20. Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., and Peters, J.: An algorithmic perspective on imitation learning. Foundations and Trends® in Robotics, 7(1-2):1–179, 2018.
21. Abbeel, P. and Ng, A. Y.: Apprenticeship learning via inverse reinforcement learning. In Proceedings of the twenty-first international conference on Machine learning, page 1, 2004.

22. Ziebart, B. D., Maas, A. L., Bagnell, J. A., and Dey, A. K.: Maximum entropy inverse reinforcement learning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 22, pages 1433–1438, 2008.
23. Syed, U. and Schapire, R. E.: A game-theoretic approach to apprenticeship learning. Advances in neural information processing systems, 20, 2007.
24. Ziebart, B. D., Ratliff, N. D., and Ratliff, N. D.: Towards robust suboptimal control via subdominance minimization. arXiv preprint arXiv:2201.03464, 2022.
25. Memarrast, O., Vu, L., and Ziebart, B. D.: Superhuman fairness. In Proceedings of the 40th International Conference on Machine Learning, eds. A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, volume 202 of Proceedings of Machine Learning Research, pages 24420–24435. PMLR, 23–29 Jul 2023.
26. LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning. Nature, 521(7553):436–444, 2015.
27. Rosenblatt, F.: The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6):386–408, 1958.
28. Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning. MIT Press, 2016. <http://www.deeplearningbook.org>.
29. Dubey, S. R., Singh, S. K., and Chaudhuri, B. B.: A comprehensive survey and performance analysis of activation functions in deep learning. CoRR, abs/2109.14545, 2021.
30. Rumelhart, D. E., Hinton, G. E., and Williams, R. J.: Learning representations by back-propagating errors. nature, 323(6088):533–536, 1986.
31. Ruder, S.: An overview of gradient descent optimization algorithms. CoRR, abs/1609.04747, 2016.
32. Krizhevsky, A., Sutskever, I., and Hinton, G. E.: Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, volume 25, pages 1097–1105, 2012.
33. Young, T., Hazarika, D., Poria, S., and Cambria, E.: Recent trends in deep learning based natural language processing. IEEE Computational Intelligence Magazine, 13(3):55–75, 2018.

- 34. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., et al.: End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316, 2016.
- 35. Dheeru, D. and Karra Taniskidou, E.: Uci machine learning repository. 2017.
- 36. Hardt, M., Price, E., and Srebro, N.: Equality of opportunity in supervised learning. Proceedings of the 30th International Conference on Neural Information Processing Systems, pages 3315–3323, 2016.

VITA

NAME	Trong Linh Vu
EDUCATION	B.S, Computer Science, University of Illinois Chicago, Chicago, Illinois, 2023
EXPERIENCE	Graduate Assistant , University of Illinois Chicago, IL, 2023-2024 Research Assistant , University of Illinois Chicago, IL, 2022-2023 Teaching Assistant , University of Illinois Chicago, IL, 2022-2023 Research Intern , Vin Big Data, Hanoi, Vietnam, 2020-2021 Data Science Intern , Becker Dickinson Israel, Haifa, Israel, 2019
PUBLICATIONS	Omid Memarrast, Linh Vu , Brian Ziebart. "Superhuman Fairness" In Proceedings of the International Conference on Machine Learning, ICML 2023.
WORKSHOPS	Omid Memarrast, Linh Vu , and Brian Ziebart. "Superhuman Fairness" In ICLR Workshop on Pitfalls of limited data and computation for Trustworthy ML, 2023.