Learning Fair Representations with Kolmogorov-Arnold Networks

Amisha Priyadarshini[1], **Sergio Gago-Masague**[1]

[1]Department of Computer Science, University of California, Irvine, USA
apriyad1@uci.edu, sgagomas@uci.edu

Abstract

Despite recent advances in fairness-aware machine learning, predictive models often exhibit discriminatory behavior towards marginalized groups. Such unfairness might arise from biased training data, model design, or representational disparities across groups, posing significant challenges in high-stakes decision-making domains such as college admissions. While existing fair learning models aim to mitigate bias, achieving an optimal trade-off between fairness and accuracy remains a challenge. Moreover, the reliance on black-box models hinders interpretability, limiting their applicability in socially sensitive domains.

To circumvent these issues, we propose integrating Kolmogorov-Arnold Networks (KANs) within a fair adversarial learning framework. Leveraging the adversarial robustness and interpretability of KANs, our approach facilitates stable adversarial learning. We derive theoretical insights into the spline-based KAN architecture that ensure stability during adversarial optimization. Additionally, an adaptive fairness penalty update mechanism is proposed to strike a balance between fairness and accuracy. We back these findings with empirical evidence on two real-world admissions datasets, demonstrating the proposed framework's efficiency in achieving fairness across sensitive attributes while preserving predictive performance.

Introduction

In recent years, the widespread adoption of Machine Learning (ML) models in high-stakes decision-making domains, such as college admissions, healthcare, and hiring, has underscored the need for ethically aligned, fairness-aware Artificial Intelligence (AI) systems. While modern deep learning (DL) models offer high predictive capacity, they remain vulnerable to amplifying historical bias in real-world datasets (Bickel, Hammel, and O'Connell 1975). In this context, undergraduate college admissions have undergone notable transformations, emphasizing the need for fairness and equal opportunity for all applicants. Recent shifts in admission policies, such as the University of California's Non-Discriminatory Policy (UC 2024), strive towards non-discriminatory review process with elimination of standardized testing. Although these approaches represent definite strides towards an equitable system, the socioeconomically marginalized groups continue to face systemic barriers in the admissions process (Chetty, Deming, and Friedman 2023). The disparity in resources and support available to these groups often result in unequal academic performance. This underscores the critical need for robust algorithmic approaches that ensure a balanced predictive performance while maintaining fairness across protected attributes to prevent disparate outcomes.

Various fair learning frameworks have been proposed in the past to address these concerns in ML models (Priyadarshini and Gago-Masague 2024), (Petrović et al. 2022). These include pre-processing, in-processing, and post-processing techniques aimed at mitigating bias (Mehrabi et al. 2021). Given the sensitive nature of real-world datasets, particularly in the domain of college admissions where the broad spectrum of input features contributes towards a holistic evaluation (Coleman, Arthur L. 2018), choosing a fairness mechanism is a challenge. The pre-processing approaches modify the input distributions, and hence are unsuitable for the case. Similarly, post-processing techniques have a limited direct influence over the model's internal representations, which curbs their effectiveness and raises fairness concerns. This further emphasizes the need for in-processing methods, such as adversarial debiasing (Zhang, Lemoine, and Mitchell 2018), that enforce fairness constraints directly within the training process, while preserving the integrity of the data. In this paper, we examine adversarial debiasing as a means to mitigate socioeconomic bias in our college admissions decision-making process.

To meet the goals of adversarial robustness and model interpretability, we adopt Kolmogorov-Arnold Networks (KANs) (Liu et al. 2024) - a novel architecture grounded in Kolmogorov-Arnold (KA) representation theorem (Schmidt-Hieber 2021). Unlike traditional Multi-Layer Perceptron (MLP) models, which employ scalar weights and fixed activation functions, KANs leverage a composition of learnable univariate spline functions. This allows for a flexible function approximation with inherent model interpretability and smoothness. The edge-based spline parameterization offers fine-grained control over feature transformations, making them well-suited for structured tabular domains. In this work, we propose an adversarially trained KAN framework with an adaptive λ update policy, aimed at balancing fairness metrics and predictive scores. We validate this framework across two different sets of real-world freshman applicants data to the University of California Irvine (UCI). Theoretical analysis and numerical experiments further demonstrate the efficiency of our proposed framework in achieving robust performance under fairness constraints while outperforming state-of-the-art (SOTA) baseline models.

Problem Formulation

In the context of college admissions, several studies, (Bickel, Hammel, and O'Connell 1975), (Bhattacharya, Kanaya, and Stevens 2017), depict the need for fairness-aware decision-making. Given the profound societal impact, the admissions process must uphold equitable treatment across socio-demographic groups. Historical bias embedded in the data often results in unfair outcomes, disproportionately impacting the marginalized communities (Gándara et al. 2024). To systematically analyze such disparities, we formalize the admission decision task into a binary classification problem.

In our formulation, the model is trained using m examples $(x_i, z_i, y_i)_{i=1}^m$, where each of these is composed of a feature vector $x_i \in \mathbb{R}^n$, containing n predictors, a binary sensitive attribute z_i, and a binary label y_i. These examples are sampled from a training distribution, say, $\Gamma = (X, Z, Y) \sim s$. Let $\mathcal{X}$, and $\mathcal{Y}$ be the feature space and label space respectively. We aim to develop a framework that leverages KANs within an adversarial debiasing setup using an adaptive penalty update mechanism. The goal is to mitigate socioeconomic bias while maintaining predictive performance by effectively balancing the trade-off between fairness and accuracy. We try to approximate the predictor model, f, with parameters θ_f in a zero-sum game setup with an adversary, g, with parameters θ_g. One of the key factors for algorithmic fairness is group fairness. For our case, we consider two types of fairness definitions: *Demographic Parity and $p\%$-Rule*. In this section, we present the fairness notions and works that are necessary to ground our proposition.

Demographic Parity

Demographic Parity (DP), also known as statistical parity, is satisfied when the model's predicted positive outcomes are equally distributed across different sensitive groups (Dwork et al. 2012). Formally for a binary sensitive attribute $z \in \{0, 1\}$, DP requires the following conditions to hold:

$$\mathbb{E}\left[\mathbb{I}(\hat{f}_{\theta_f}(x) = 1) \mid z = 0\right] = \mathbb{E}\left[\mathbb{I}(\hat{f}_{\theta_f}(x) = 1) \mid z = 1\right] \tag{1}$$

Here, $\hat{f}_{\theta_f}(x) = \mathbb{I}(f_{\theta_f}(x) > 0.5)$ denotes the predicted label after applying a decision threshold to the model output, and $\mathbb{I}(\cdot)$ is the indicator function. Based on this criterion, the fairness-aware learning objective can be formulated as:

$$\begin{aligned} &\arg\min_{\theta_f} \; \mathbb{E}_{(x,z,y)\sim s} \, \mathcal{L}_{\mathcal{Y}}(f_{\theta_f}(x), y) \\ \text{s.t.} \quad & |\mathbb{E}_s[\hat{f}_{\theta_f}(x)|z=1] - \mathbb{E}_s[\hat{f}_{\theta_f}(x)|z=0]| < \epsilon \end{aligned} \tag{2}$$

where ϵ is a small tolerance parameter that controls the allowable disparity between groups.

$p\%$-Rule

The $p\%$-Rule is a widely used group fairness metric that quantifies the disparity in the rates of positive outcomes across different demographic groups (Hardt, Price, and Srebro 2016). For a binary sensitive attribute $z \in \{0, 1\}$, the predicted positive outcome rates for both the groups can be defined as, $r_0 = \mathbb{E}\left[\mathbb{I}(\hat{f}_{\theta_f}(x) = 1) \mid z = 0\right]$, $r_1 = \mathbb{E}\left[\mathbb{I}(\hat{f}_{\theta_f}(x) = 1) \mid z = 1\right]$. Then, the $p\%$-Rule requires:

$$\frac{\min(r_0, r_1)}{\max(r_0, r_1)} \geq \frac{p}{100} \tag{3}$$

A higher value of the $p\%$-Rule (closer to 1 or 100%) indicates improved fairness, as it implies similar treatment across groups.

Limitations of Existing Approaches

In standard supervised learning, a neural network (NN) model, f, is typically trained by minimizing the expected prediction loss, $\mathcal{L}_{\mathcal{Y}}$, without incorporating any fairness constraints. When applied to real-world datasets that reflect historical bias, such models can inadvertently encode and even amplify disparities in predicted outcomes across sensitive groups. This violates principles of group fairness, particularly when the sensitive attribute, z, is correlated with the outcome, y.

Adversarial debiasing framework extends this formulation by introducing an adversary g_{θ_g} that attempts to infer the sensitive attribute z from the predictor's output $f_{\theta_f}(x)$. The predictor f_{θ_f} is simultaneously trained to minimize classification loss while maximizing the adversary's prediction error, resulting in the following optimization objective:

$$\begin{aligned} &\min_{\theta_f} \; \mathbb{E}_{(x,z,y)\sim s} \, \mathcal{L}_{\mathcal{Y}}(f_{\theta_f}(x), y) \\ \text{s.t.} \quad &\min_{\theta_g} \; \mathbb{E}_{(x,z,y)\sim s} \, \mathcal{L}_{\mathcal{Z}}(g_{\theta_g}(f_{\theta_f}(x)), z) > \epsilon' \end{aligned} \tag{4}$$

where, ϵ' denotes the fairness threshold, and $\mathcal{L}_{\mathcal{Z}}$ represents the adversary loss. To obtain a better balance between the predictions of the predictor and adversary, (Zhang, Lemoine, and Mitchell 2018) proposes a more relaxed formulation incorporating a fairness coefficient, λ:

$$\begin{aligned} &\min_{\theta_f} \max_{\theta_g} \mathbb{E}_{(x,z,y)\sim s} \, \mathcal{L}_{\mathcal{Y}}(f_{\theta_f}(x), y) \\ &-\lambda \cdot \mathbb{E}_{(x,z,y)\sim s} \, \mathcal{L}_{\mathcal{Z}}(g_{\theta_g}(f_{\theta_f}(x)), z) \end{aligned} \tag{5}$$

where, $\lambda \in \mathbb{R}^+$ controls the degree of fairness for striking a balance between the tradeoff.

While adversarial debiasing has proven effective in reducing group-level disparities, we discuss two persistent challenges that remain. *Firstly*, fairness constraints (such as DP) lead to provable loss in accuracy, especially when base rates differ across sensitive groups. According to (Zhao and Gordon 2022), any predictor that satisfies exact DP incurs an average classification error that is lower bounded by half the base rate gap between the groups. This formalizes the inherent fairness–accuracy tradeoff that can manifest as instability during model training. Moreover, enforcing strict criteria

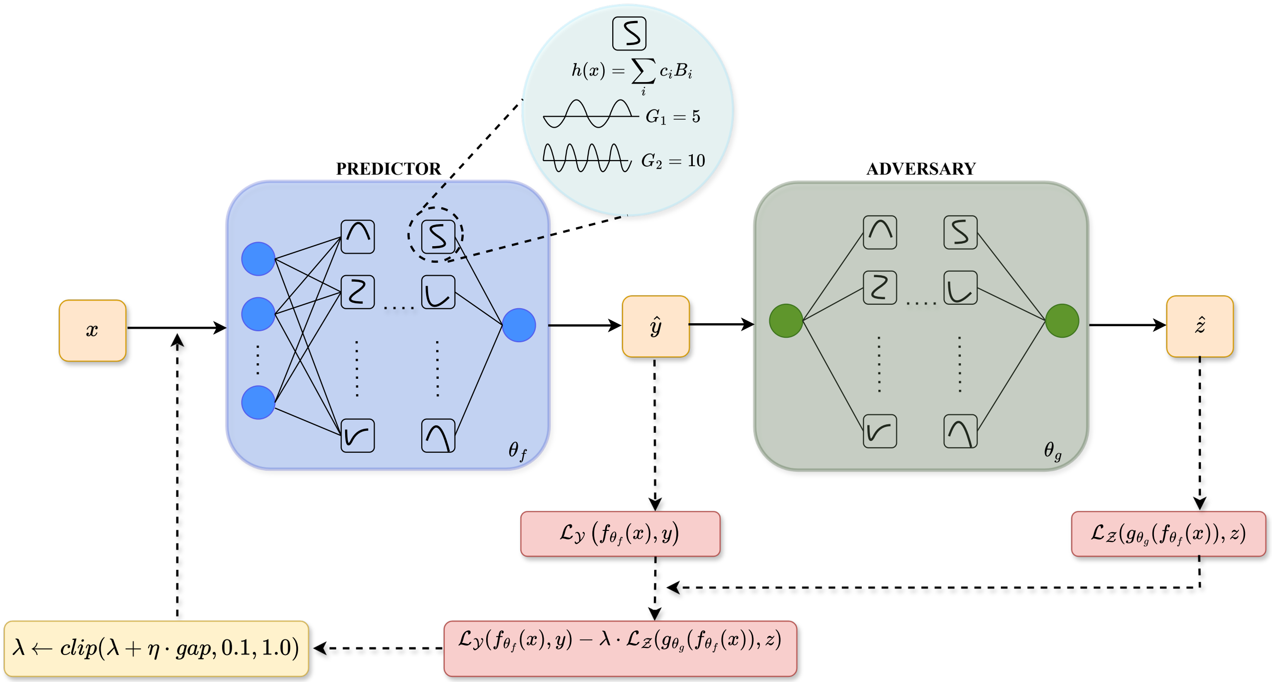


Figure 1: Schematic overview of the proposed adversarial debiasing framework using KANs. The classifier (KAN), f, learns predictive representations from input features, x, while an adversary (also a KAN), g, attempts to infer the sensitive attribute, z, from the classifier's output, y, using a min-max objective. Fairness parameter, λ, is updated adaptively after every training epoch in an attempt to balance accuracy and fairness scores.

like DP can make models **less robust** to data perturbations or distribution shifts. The requirement to equalize output distributions across groups may lead to unreliable models, particularly in high-stake settings. *Secondly*, while effective at mitigating group-level bias, the adversarial formulation suffers from a critical limitation: the learned neural predictor f_{θ_f} remains a black-box function that lacks interpretability. Standard NNs do not provide explicit functional representations of feature-wise contributions, making it challenging to analyze decision logic, particularly in sensitive real-world settings such as college admissions.

In contrast, KAN's spline-based architecture and explicit functional decomposition offer a structured, interpretable, and robust alternative to traditional NN. Moreover, KAN retains the flexibility necessary for learning fair decision boundaries. In this study, we present a theoretical analysis and experimental backing demonstrating the suitability of KAN representations for adversarial learning setup.

Related Work

Several works, (Chetty, Deming, and Friedman 2023), (Barocas, Hardt, and Narayanan 2023), (Zimdars 2010), (Woo et al. 2023), showcase the critical need for fair learning in the college admissions setup. Although various works like (Doleck et al. 2020), (Waters and Miikkulainen 2014) demonstrate ML implementation in this context, adhering to fairness criteria or prioritizing the socioeconomically marginalized groups in the decision-making is still overlooked. This shortcoming is furthered by the reliance on DL models that are hard to interpret, leading to trust issues (Rudin 2019). In the context of college admissions, interpretability plays a significant role, given the direct impact of these decisions on applicants and their subject to public scrutiny.

Although fair-learning approaches like (Zhang, Lemoine, and Mitchell 2018), (Madras et al. 2018), (Lahoti et al. 2020), (Priyadarshini and Gago-Masague 2025), implement adversarial learning to incorporate fairness constraints, they are defined in the context of DL setups. Even several mathematically grounded works such as, (Jia, Zhang, and Vosoughi 2024) introducing Lipschitz-based fairness criterion, and (Chzhen et al. 2020) that leverage optimal mass transport for regression tasks, provide rigorous formulations but are mostly tailored to specific problem settings, limiting their applicability. Secondly, works like (Zhao and Gordon 2022), (Calders, Kamiran, and Pechenizkiy 2009), (Kamiran and Calders 2009) further affirm the inherent fairness-accuracy tradeoffs, cited as one of the major problems in our case study. These underlying challenges motivate our proposed approach. While various works like (Kiamari, Kiamari, and Krishnamachari 2024), (Bodner et al. 2024), (Vaca-Rubio et al. 2024), (Alter, Lapid, and Sipper 2024) have explored KANs on various domains, to the best of our knowledge, our proposed work is the first to elaborate on

KAN-based adversarial debiasing framework.

Robust Adversarial Learning with KANs

The Adversarial Debiasing (Zhang, Lemoine, and Mitchell 2018) framework adopts a zero-sum game approach, as shown in Eq. 5. However, due to the inherent fairness-accuracy tradeoff, as discussed in (Zhao and Gordon 2022), improvements in fairness often comes at the cost of predictive performance. In particular, (Hardt, Price, and Srebro 2016) shows that enforcing DP can eliminate an otherwise perfect predictor, underscoring the importance of theoretically quantifying this tradeoff in classification tasks. Although several adversarial debiasing techniques based on deep NN models (Priyadarshini and Gago-Masague 2024) have been proposed they often suffer from lack of interpretability. In contrast, we propose utilizing KAN as the learning model in an adversarial debiasing framework. Having demonstrated adversarial robustness (Alter, Lapid, and Sipper 2024), and structural interpretability, KAN models have a definite edge over MLPs. We further incorporate an adaptive λ-update policy in the KAN-based adversarial framework with various optimizer settings. This helps enhance the model's robustness to fairness-accuracy trade-off. Despite the high complexity of the KAN representations, we further affirm their efficacy at generalization without overfitting to fairness constraints. An overview of the proposed adversarial learning framework is presented in Figure 1.

Preliminaries

The KANs (Liu et al. 2024) are inspired by the KA Representation theorem (Kolmogorov 1957), which states that any multivariate continuous function can be written as a finite superposition of continuous univariate functions. A distinctive feature of these networks is their use of learnable univariate spline functions in place of the scalar weights and fixed activation functions of traditional MLPs. These characteristics improve the model's capacity for accurate and flexible approximations. A traditional KAN function approximator, $f : [0, 1]^n \to \mathbb{R}$, can be expressed as:

$$f(x_1, x_2, ..., x_n) = \sum_{q=1}^{2n+1} \Phi_q(\sum_{r=1}^{n} \phi_{q,r}(x_r)) \tag{6}$$

where the inner and outer univariate functions map as, $\phi_{q,r} : [0, 1] \to \mathbb{R}$ and $\Phi_q : \mathbb{R} \to \mathbb{R}$ respectively. While the early critiques questioned the practical utility of the KA theorem due to the non-smoothness of inner functions (Girosi and Poggio 1989a), (Liu et al. 2024) demonstrate its viability in ML by leveraging smoother, sparser compositional structures, extending the network to a deep KAN architecture (Lin, Tegmark, and Rolnick 2017).

Given, all the functions to be learned are univariate, represented in Eq. 6, each 1D function, ϕ_i, is parametrized as a B-spline curve, say $h(x)$. It can be defined as $h(x) = \sum_i c_i B_i(x)$, where c_is are the learnable coefficients of the local B-spline basis functions. Each spline function is a piecewise-polynomial (typically cubic) function, and is $(k + 1)$ (or, twice differentiable for cubic splines) times continuously differentiable (Girosi and Poggio 1989b). In the following section, we discuss the structural properties of KANs that make them well-suited for fairness-aware adversarial learning with theoretical justifications.

Assumption 1. Given $\mathcal{X} \subset \mathbb{R}^n$ denotes the input feature space, and $Z \in \{0, 1\}$ represents the binary sensitive attribute, we impose the following conditions:

1. The input space $\mathcal{X}$ is compact; that is, it is bounded and closed.
2. Each univariate spline function, $h(x)$, used in the KAN architecture is piecewise polynomial (usually cubic), and is twice continuously differentiable on its domain.

Theoretical Analysis

Lemma 1. *Each univariate spline function is Lipschitz continuous on a bounded range, is differentiable, and defined over a compact interval. Hence, f is Lipschitz continuous on a bounded domain.*

Proof. Given $h_i : \mathbb{R} \to \mathbb{R}$ denotes the univariate spline functions used in KAN architecture. Based on our Assumption 1, each h_i is a cubic spline continuous function that is C^2 smooth (i.e., twice differentiable), and defined over a compact interval ($\subset \mathbb{R}$). Hence, by the Mean Value Theorem, there exists a constant $L_i > 0$ such that:

$$|h_i(x) - h_i(x')| \leq L_i.|x - x'|$$

for all $x, x' \in \mathbb{R}$. As the overall KAN function, f, is a finite summation and composition of these univariate spline functions, it follows that f itself is L-Lipschitz continuous on a bounded input domain (Rudin 1976). □

Lemma 2. *Every univariate spline function, in KAN is β_i-smooth, and the finite sum of spline functions preserves the smoothness. Hence, f is β-smooth.*

Proof. Given, h_i is twice differentiable implies $|h_i''(x)| \leq \beta_i$ for all $x \in \mathcal{X}$, for some $\beta_i \in [0, \infty)$. This proves that h_i is β_i-smooth. Provided the overall KAN function, f, is constructed by combining these univariate spline functions through finite summations and compositions, we note two important properties: *firstly*, the finite sum of spline functions preserves smoothness, and *secondly*, the composition of smooth functions preserves smoothness (Rudin 1976). Hence, it implies that f is β-smooth. A full justification of the said properties has been provided in the Appendix. □

As shown in Lemma 2, the KAN function, f, inherits the smoothness of its constituent univariate functions, and remains β-smooth on the bounded domain, $\mathcal{X}$. This ensures stable convergence during optimization owing to bounded variations in gradients (Schmidt-Hieber 2021). Furthermore, Lemma 1 proves Lipschitz continuity of KAN models, hence ensuring model robustness to small input perturbations (Alter, Lapid, and Sipper 2024). These structural properties play a crucial role in stabilizing the adversarial training dynamics and ensure effective debiasing. Further mathematical evidence suggesting the efficiency of KAN structure in an adversarial debiasing setup has been provided in the Appendix.

Algorithm 1: Adversarial training with adaptive λ policy

Input: Initialize KAN-based classifier f and adversary g with grid $\mathcal{G}$ and order k. Training data (x, z, y), classifier learning rate η_{clf}, adversary learning rate η_{adv}, threshold τ, fairness learning rate η.

Parameter: Initialize fairness penalty weight λ

```
for each grid level G_i ∈ G do
    if i = 0 then
        Initialize f and g on grid G_0
    else
        Initialize f and g from trained models on G_{i-1}
    end if
    Freeze classifier and compute outputs f(x)
    Train adversary:
        Update g to minimize L_Z(g(f(x)), z)
    Train classifier with adversarial debiasing:
    for each epoch t = 1, ..., T do
        Update f to minimize:
            L_Y(f(x), y) − λ · L_Z(g(f(x)), z)
        Evaluate fairness metric: compute p%-Rule
        Compute gap: δ ← (τ − p%-Rule)/τ
        Update penalty: λ ← clip(λ + η · δ, 0.1, 1.0)
    end for
end for
```

Adaptive Penalty Mechanism One of the key challenges in an adversarial debiasing setup lies in balancing fairness and predictive performance, which is governed by the fairness coefficient λ. It scales the adversarial loss relative to the predictive loss. However, in a non-adaptive setup, determining λ is non-trivial. The process requires extensive manual tuning that is computationally expensive (Sattigeri et al. 2019), especially given the high network complexity of KANs. To address these limitations, we introduce an adaptive λ update mechanism that dynamically adjusts the fairness penalty during training. At each training epoch, we compute the fairness gap and update the penalty weights in proportion to this gap. The mechanism follows a linear rule with a learning rate, η, and is clipped within a safe range to ensure stability in trade-off as shown in Algorithm 1. To further affirm our claims, we empirically demonstrate the effectiveness of our approach in the following section.

Numerical Experiments

Simulation Setup

Various works, (Marcinkowski et al. 2020), (Alvero et al. 2020), have documented socioeconomic factors to significantly influence college admission outcomes. To empirically assess the efficacy of our proposed method in mitigating such bias, we conduct experiments on two distinct college admissions dataset collected from freshman applicants to the Dept. of Computer Science, UCI. The first dataset, denoted as $D_{(1)}$, comprises of 4,442 application records from the Fall 2021 admission cycle, and the second dataset, denoted as $D_{(2)}$, includes 26,243 application records spanning admission cycles from Fall 2018 to Fall 2024. Each of the datasets contain approximately 140 features, encompassing demographic, academic records, high school information, and essay question responses (Priyadarshini, Martinez-Neda, and Gago-Masague 2023). The broad spectrum of features facilitates *holistic evaluation*, a comprehensive approach to college admissions that ensures applicants are assessed based on a wide range of criteria. For the study, after a systematic analysis of input features, we consider two binary sensitive attributes: low-income status and first-generation flag, which directly affect the socioeconomic demographics. We justify this choice by empirically demonstrating the presence of bias in model predictions using Kernel Density Estimation (KDE) plots, which are included in the Appendix. Furthermore, we use the final read score as the target variable for the binary classification task. To quantitatively evaluate model fairness, we utilize two fairness metrics, $p\%$-Rule, and DP. In addition, to assess the predictive performance of our model, we employ Accuracy and AUROC scores. Next, we incorporate three different types of optimizers, namely Adam (Kingma and Ba 2014), Optimistic Adam (OAdam) (Daskalakis et al. 2017), and ADOPT (Taniguchi et al. 2024), for each training configuration. Despite KAN's structural robustness and theoretical guarantees in adversarial setting, we empirically explore different optimization strategies to enhance model convergence and stability during the debiasing process. Further discussions have been included in the Appendix.

Baseline Models

To evaluate the effectiveness of our proposed framework, we compare our approach against three baseline models:

- Baseline $B_{(1)}$ is a standard adversarial debiasing framework using a fully connected feedforward neural network (FFNN) as the classifier and adversary models (Priyadarshini and Gago-Masague 2024),
- Baseline $B_{(2)}$ is a SOTA implementation of exponential gradient-based debiasing model (Agarwal et al. 2018),
- Baseline $B_{(3)}$ is a SOTA method based on the ROAD (Robust Optimization for Adversarial Debiasing) framework (Grari et al. 2023).

Simulation Results and Discussion

In this section, we provide a comprehensive comparison of our proposed KAN-based adversarial debiasing framework with an adaptive λ update mechanism across the two datasets, $D_{(1)}$ and $D_{(2)}$, using three distinct optimizer settings. In the experimental setup, as showcased in Table 1, we use a fixed B-spline order of $k = 3$. This spline order choice is consistent with (Liu et al. 2024), that highlight it as striking an effective balance between expressiveness, smoothness, and training stability. Table 1a reports the key performance metrics, including classification accuracy, AUROC, $p\%$-Rule, and DP across the two binary sensitive attributes, and Table 1b lists the necessary notations.

On training the KAN models on $D_{(1)}$ dataset, we observe the Adam-based model to offer stronger predictive power

(a)

Model	Optimizer	Acc	AUROC	$p\%$-Rule$_{(1)}$	$p\%$-Rule$_{(2)}$	DP$_{(1)}$	DP$_{(2)}$
$K_{(1)}$	Adam	74.92	80.27	85.79	89.65	0.028	0.035
	OAdam	74.47	76.86	90.39	92.99	0.072	0.084
	ADOPT	**74.58**	**81.45**	**89.21**	**90.92**	**0.047**	**0.055**
$K_{(2)}$	Adam	82.36	85.60	84.11	81.22	0.037	0.043
	OAdam	81.69	86.26	99.25	99.55	0.026	0.03
	ADOPT	**82.51**	**86.68**	**99.25**	**99.70**	**0.032**	**0.037**
$B_{(1)}$	Adam	72.55	72.53	98.32	99.61	0.009	0.011
	OAdam	74.13	74.13	98.52	95.04	0.004	0.018
	ADOPT	70.19	70.20	97.15	93.50	0.019	0.022
$B'_{(1)}$	Adam	74.87	71.77	93.97	97.79	0.01	0.004
	OAdam	78.61	63.05	96.56	97.06	0.006	0.015
	ADOPT	77.75	63.45	90.64	99.0	0.006	0.004
$B_{(2)}$	Adam	72.44	76.12	95.22	97.59	0.024	0.012
$B'_{(2)}$	Adam	79.12	64.47	98.00	88.84	0.003	0.018
$B_{(3)}$	Adam	74.92	80.95	83.10	83.91	0.085	0.08
$B'_{(3)}$	Adam	78.36	81.99	97.72	93.29	0.004	0.014

(b)

$K_{(1)}$	KAN (trained on $D_{(1)}$)
$K_{(2)}$	KAN (trained on $D_{(2)}$)
$B_{(1)}$	trained on $D_{(1)}$
$B'_{(1)}$	trained on $D_{(2)}$
$B_{(2)}$	trained on $D_{(1)}$
$B'_{(2)}$	trained on $D_{(2)}$
$p\%$-Rule$_{(1)}$	$p\%$-Rule (Low Inc.)
$p\%$-Rule$_{(2)}$	$p\%$-Rule (First Gen.)
DP$_{(1)}$	DP Gap (Low Inc.)
DP$_{(2)}$	DP Gap (First Gen.)

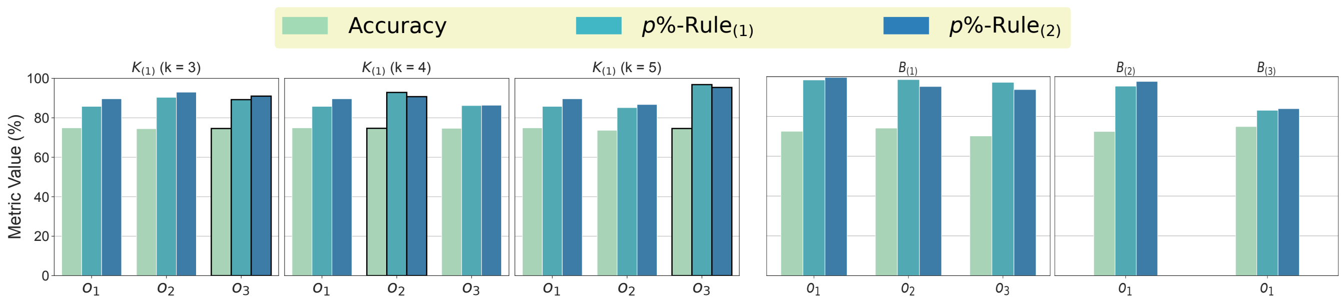
Table 1: (a) Performance and fairness comparison across proposed frameworks, and baseline models, trained on two distinct datasets, $D_{(1)}$ and $D_{(2)}$, under fairness constraints with three different optimizers, Adam, OAdam, and ADOPT. (b) Notation reference for model identifiers and fairness metric used in (a).

(0.45% *incr. in Acc.,* 3.41% *incr. in AUROC*) compared to OAdam, which enhances fairness outcomes significantly. However, the model trained using ADOPT exhibits a balanced performance across fairness and accuracy. On the other hand, training the KAN models on the larger $D_{(2)}$ dataset yields notable improvements in predictive performance and fairness. Although all setups trained on $D_{(2)}$ show significant performance improvement, the ADOPT-based model stands out by achieving balanced optimization between accuracy (0.15%*, and* 0.82% *incr. in Acc. compared to Adam setup*) and fairness metrics ($\{15.14\%, 18.48\%\}$ *incr. in* $p\%$*-Rule compared to Adam setup*). Meanwhile, the $B_{(1)}$ and $B'_{(1)}$ models under-perform relative to the KAN models, especially on the $D_{(2)}$ dataset. This suggests KAN-based adversarial learning to be more effective in maintaining fairness-accuracy trade-offs while capturing complex patterns, even in a feature-rich setting. SOTA baseline models, $B_{(2)}$ and $B_{(3)}$, display varying behavior. $B_{(2)}$ achieves high fairness on $D_{(1)}$ but suffers a drop on $D_{(2)}$, whereas $B_{(3)}$ starts with lower fairness on $D_{(1)}$ but improves substantially on $D_{(2)}$ with proper fine-tuning.

The empirical evidence demonstrates that the KAN-based adversarial debiasing framework, regardless of the choice of optimizer, consistently outperforms the baseline models across both the datasets. This can be largely attributed to the adaptive policy usage within the training module. Its ability to dynamically adjust and bound the fairness constraint, λ, achieves to secure a balanced performance. Amongst all the configurations, the model trained on ADOPT demonstrates comparative superior balanced performance. This observation suggests ADOPT to be well suited to the spline-based architecture of KANs, owing to their adaptive convergence behavior (Taniguchi et al. 2024). We highlight the frameworks that demonstrate a balanced fairness-accuracy trade-off in Table 1 for clarity.

Ablation Study

To investigate how KAN spline knot complexity, k, affects the adversarial learning process, we perform an ablation study focusing on three spline orders $k = \{3, 4, 5\}$. We chose the $D_{(1)}$ dataset for the experiments due to its controlled small-scale setting. It enables focused assessment of the model's sensitivity to architectural changes under constrained data conditions. In addition to the KAN-based adversarial learning models, $K_{(1)}$, we also train the baseline model, $B_{(1)}$, across the three optimizer settings, and the SOTA baseline models, $B_{(2)}$, and $B_{(3)}$ across the Adam optimizer. As illustrated in Figure 2a, increasing the spline order k leads to subtle variations in accuracy and fairness metrics. Although the $K_{(1)}$ models exhibit relatively stable accuracy across different k setups, the fairness metrics, measured via the $p\%$-Rule, tend to improve with higher spline orders. This suggests that more expressive spline functions, in an adversarial setting, can better model fair representations. Although higher k (e.g., $k = 4, 5$) may introduce convergence instability, as observed in Figure 3, the model trained with ADOPT exhibits comparatively stable training dynamics than those trained on Adam or OAdam. Next, when comparing the $K_{(1)}$ models with the baseline MLP models, $B_{(1)}$, we observe that the former consistently outperforms the latter in achieving a balance between accuracy and fairness. As observed in Figure 2b, $B_{(1)}$ models



(a) (b)

Figure 2: Ablation study to compare predictive performance and fairness scores across proposed frameworks with different spline knot complexities, and baseline models under varying optimization strategies, evaluated on dataset $D_{(1)}$. (a) Illustrates metrics of the proposed KAN-based adversarial frameworks using three different optimizer techniques (o_1: Adam, o_2: OAdam, o_3: ADOPT). Similarly, (b) depicts the Baseline model performance across varying optimizers.

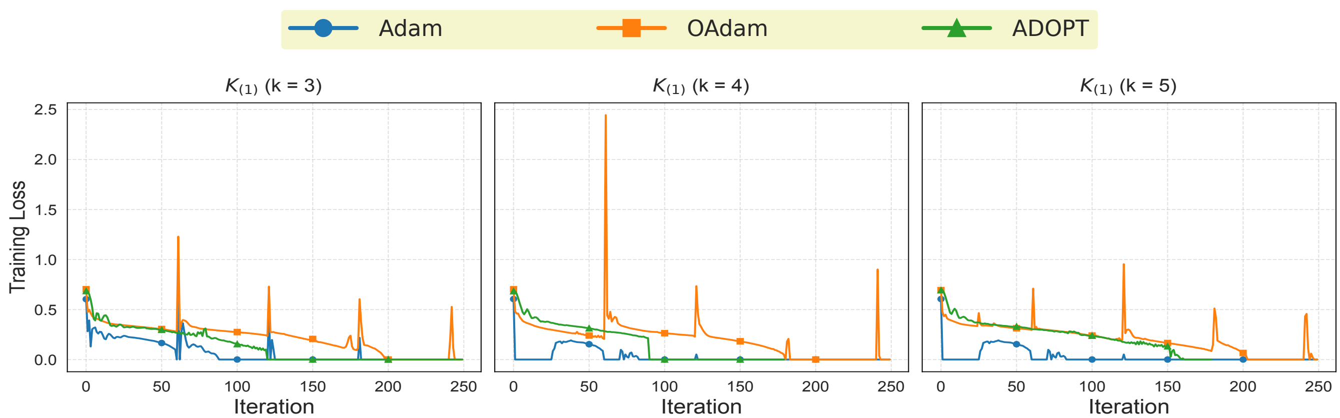


Figure 3: Training Loss convergence of the KAN-based adversarial learning framework across three different spline knot orders ($k \in \{3, 4, 5\}$), under three different optimization techniques. Each plot illustrates the convergence behavior of the model, highlighting the impact of optimizer choice and KAN model's spline complexity on training stability.

have a higher fairness score but comparatively lower accuracy across the three optimizer settings. This further empirically substantiates the efficiency of the adaptive λ update policy. The SOTA baseline models, $B_{(2)}$ and $B_{(3)}$, show strong performance in fairness but incur slight drops in accuracy as seen in Figure 2b. As seen in Figure 3, the $K_{(1)}$ model training loss curve exhibits pronounced spikes. This could be attributed to the smoothness-dimension trade-off (Samadi, Müller, and Schuppert 2024) in KANs. While this can induce oscillatory behavior during training, it is inherent to preserving the universal approximation. A formal discussion of this property is provided in the Appendix. Despite the fluctuations in KAN-based setups, our findings help us empirically establish the efficiency of the proposed adversarial learning framework.

Conclusion and Future Work

In this study, we propose a KAN-based adversarial learning framework with adaptive λ update policy to mitigate socioeconomic bias in a college admission setting. We discuss the various structural properties of KAN, providing proper mathematical justification behind the adversarial robustness of the networks. Through extensive experiments on two distinct real-world admissions datasets, we demonstrate the efficiency of our proposed framework by consistently outperforming SOTA baselines while achieving a balance between the inherent accuracy-fairness tradeoff. While our proposed approach effectively mitigates output-level bias, it opens up several promising enhancements. The incorporation of explicit feature-level bias detection through a reweighting mechanism could improve the overall fairness of the model while also helping to adhere to any data privacy constraints. Additionally, extending the current framework by incorporating alternative fairness definitions beyond DP could provide new insights. We believe this work offers a promising step towards aligning algorithmic performance with social responsibility, encouraging future explorations into KAN-based fairness-aware systems that could uphold both ethical and performance standards in critical decision-making.

Acknowledgments

We gratefully acknowledge the University of California, Irvine, for providing access to the admissions datasets used in this study. We also thank Prof. Iftekhar Ahmed for providing access to GPU resources used for large-scale experiments.

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. In *International conference on machine learning*, 60–69. PMLR.

Alter, T.; Lapid, R.; and Sipper, M. 2024. On the robustness of kolmogorov-arnold networks: An adversarial perspective. *arXiv preprint arXiv:2408.13809*.

Alvero, A.; Arthurs, N.; Antonio, A. L.; Domingue, B. W.; Gebre-Medhin, B.; Giebel, S.; and Stevens, M. L. 2020. AI and holistic review: informing human reading in college admissions. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 200–206.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT press.

Bhattacharya, D.; Kanaya, S.; and Stevens, M. 2017. Are university admissions academically fair? *Review of Economics and Statistics*, 99(3): 449–464.

Bickel, P. J.; Hammel, E. A.; and O'Connell, J. W. 1975. Sex Bias in Graduate Admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science*, 187(4175): 398–404.

Bodner, A. D.; Tepsich, A. S.; Spolski, J. N.; and Pourteau, S. 2024. Convolutional kolmogorov-arnold networks. *arXiv preprint arXiv:2406.13155*.

Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*, 13–18. IEEE.

Chetty, R.; Deming, D. J.; and Friedman, J. N. 2023. Diversifying society's leaders? The determinants and causal effects of admission to highly selective private colleges. Technical report, National Bureau of Economic Research.

Chzhen, E.; Denis, C.; Hebiri, M.; Oneto, L.; and Pontil, M. 2020. Fair regression with wasserstein barycenters. *Advances in Neural Information Processing Systems*, 33: 7321–7331.

Coleman, Arthur L. 2018. Understanding Holistic Review in Higher Education Admissions. https://educationcounsel.com/our_work/publications/roadmap-framework/understanding-holistic-review-in-higher-education-admissions-guiding-principles-and-model-illustrations-2.

Daskalakis, C.; Ilyas, A.; Syrgkanis, V.; and Zeng, H. 2017. Training gans with optimism. *arXiv preprint arXiv:1711.00141*.

Doleck, T.; Lemay, D. J.; Basnet, R. B.; and Bazelais, P. 2020. Predictive analytics in education: a comparison of deep learning frameworks. *Education and Information Technologies*, 25(3): 1951–1963.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Gándara, D.; Anahideh, H.; Ison, M. P.; and Picchiarini, L. 2024. Inside the black box: Detecting and mitigating algorithmic bias across racialized groups in college student-success prediction. *aera Open*, 10: 23328584241258741.

Girosi, F.; and Poggio, T. 1989a. Representation properties of networks: Kolmogorov's theorem is irrelevant. *Neural Computation*, 1(4): 465–469.

Girosi, F.; and Poggio, T. 1989b. Representation properties of networks: Kolmogorov's theorem is irrelevant. *Neural Computation*, 1(4): 465–469.

Grari, V.; Laugel, T.; Hashimoto, T.; Lamprier, S.; and Detyniecki, M. 2023. On the fairness road: Robust optimization for adversarial debiasing. *arXiv preprint arXiv:2310.18413*.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Jia, Y.; Zhang, C.; and Vosoughi, S. 2024. Aligning relational learning with lipschitz fairness. In *The Twelfth International Conference on Learning Representations*.

Kamiran, F.; and Calders, T. 2009. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*, 1–6. IEEE.

Kiamari, M.; Kiamari, M.; and Krishnamachari, B. 2024. Gkan: Graph kolmogorov-arnold networks. *arXiv preprint arXiv:2406.06470*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kolmogorov, A. N. 1957. On the representations of continuous functions of many variables by superposition of continuous functions of one variable and addition. In *Dokl. Akad. Nauk USSR*, volume 114, 953–956.

Lahoti, P.; Beutel, A.; Chen, J.; Lee, K.; Prost, F.; Thain, N.; Wang, X.; and Chi, E. 2020. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33: 728–740.

Lin, H. W.; Tegmark, M.; and Rolnick, D. 2017. Why does deep and cheap learning work so well? *Journal of Statistical Physics*, 168(6): 1223–1247.

Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T. Y.; and Tegmark, M. 2024. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, 3384–3393. PMLR.

Marcinkowski, F.; Kieslich, K.; Starke, C.; and Lünich, M. 2020. Implications of AI (un-) fairness in higher education admissions: the effects of perceived AI (un-) fairness on exit, voice and organizational reputation. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 122–130.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Petrović, A.; Nikolić, M.; Radovanović, S.; Delibašić, B.; and Jovanović, M. 2022. FAIR: Fair adversarial instance reweighting. *Neurocomputing*, 476: 14–37.

Priyadarshini, A.; and Gago-Masague, S. 2024. Fair Evaluator: An Adversarial Debiasing-based Deep Learning Framework in Student Admissions. In *2024 IEEE 6th International Conference on Cognitive Machine Intelligence (CogMI)*, 152–161. IEEE.

Priyadarshini, A.; and Gago-Masague, S. 2025. FairFusion: Distributionally Robust Fair-Multimodal Learning for College Admissions. In *Women in Machine Learning Workshop@ NeurIPS 2025*.

Priyadarshini, A.; Martinez-Neda, B.; and Gago-Masague, S. 2023. Admission prediction in undergraduate applications: an interpretable deep learning approach. In *2023 Fifth International Conference on Transdisciplinary AI (TransAI)*, 135–140. IEEE.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5): 206–215.

Rudin, W. 1976. *Principles of Mathematical Analysis*. McGraw-Hill, 3rd edition.

Samadi, M. E.; Müller, Y.; and Schuppert, A. 2024. Smooth Kolmogorov Arnold networks enabling structural knowledge representation. *arXiv preprint arXiv:2405.11318*.

Sattigeri, P.; Hoffman, S. C.; Chenthamarakshan, V.; and Varshney, K. R. 2019. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5): 3–1.

Schmidt-Hieber, J. 2021. The Kolmogorov–Arnold representation theorem revisited. *Neural networks*, 137: 119–126.

Taniguchi, S.; Harada, K.; Minegishi, G.; Oshima, Y.; Jeong, S. C.; Nagahara, G.; Iiyama, T.; Suzuki, M.; Iwasawa, Y.; and Matsuo, Y. 2024. ADOPT: Modified Adam Can Converge with Any β_2 with the Optimal Rate. *Advances in Neural Information Processing Systems*, 37: 72438–72474.

UC. 2024. Non-Discrimination Policy. https://grad.uci.edu/non-discrimination-policy/.

Vaca-Rubio, C. J.; Blanco, L.; Pereira, R.; and Caus, M. 2024. Kolmogorov-arnold networks (kans) for time series analysis. *arXiv preprint arXiv:2405.08790*.

Waters, A.; and Miikkulainen, R. 2014. Grade: Machine learning support for graduate admissions. *Ai Magazine*, 35(1): 64–64.

Woo, S. E.; LeBreton, J. M.; Keith, M. G.; and Tay, L. 2023. Bias, fairness, and validity in graduate-school admissions: A psychometric perspective. *Perspectives on Psychological Science*, 18(1): 3–31.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

Zhao, H.; and Gordon, G. J. 2022. Inherent tradeoffs in learning fair representations. *Journal of Machine Learning Research*, 23(57): 1–26.

Zimdars, A. 2010. Fairness and undergraduate admission: a qualitative exploration of admissions choices at the University of Oxford. *Oxford Review of Education*, 36(3): 307–323.