ECAI 2025
I. Lynce et al. (Eds.)

doi:10.3233/FAIA251214

Promoting Intersectional Fairness Through Knowledge Distillation

Md Fahim Sikder[a,*], Resmi Ramachandranpillai[b], Daniel de Leng[a] and Fredrik Heintz[a]

[a]Department of Computer and Information Science (IDA), Linköping University, Sweden
[b]Institute for Experiential AI, Northeastern University, USA

Abstract.

As Artificial Intelligence-driven decision-making systems become increasingly popular, ensuring fairness in their outcomes has emerged as a critical and urgent challenge. AI models, often trained on open-source datasets embedded with human and systemic biases, risk producing decisions that disadvantage certain demographics. This challenge intensifies when multiple sensitive attributes interact, leading to intersectional bias, a compounded and uniquely complex form of unfairness. Over the years, various methods have been proposed to address bias at the data and model levels. However, mitigating intersectional bias in decision-making remains an under-explored challenge. Motivated by this gap, we propose a novel framework that leverages knowledge distillation to promote intersectional fairness. Our approach proceeds in two stages: first, a teacher model is trained solely to maximize predictive accuracy, followed by a student model that inherits the teacher's representational knowledge while incorporating intersectional fairness constraints. The student model integrates tailored loss functions that enforce parity in false positive rates and demographic distributions across intersectional groups, alongside an adversarial objective that minimizes protected attribute information within the learned representation. Empirical evaluation across multiple benchmark datasets demonstrates that we achieve a 52% increase in accuracy for multi-class classification and a 61% reduction in average false positive rate across intersectional groups and outperforms state-of-the-art models. This distillation-based methodology provides a more stable optimization opportunity than direct fairness approaches, resulting in substantially fairer representations, particularly for multiple sensitive attributes and underrepresented demographic intersections.

1 Introduction

With the increasing availability of extensive datasets, Artificial Intelligence (AI) systems have seen rapid development in automated decision-making processes across diverse domains such as healthcare, finance, recruitment, and criminal justice [18]. While AI systems hold promising efficiency and reducing subjectivity, they frequently exhibit biases against specific demographics groups [21, 20]. Such biases generally arise from underlying discriminatory patterns embedded within training datasets, reflecting either human or historical inequities encoded during data collection or processing [4, 18]. A notable example of bias in automated decision making is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) risk assessment tool, which notoriously produced higher false-positive rates for African-American defendants, unfairly categorizing them as high-risk individuals [1]. Additionally, the Dutch Tax and Customs Administration in 2021 improperly utilized residents citizenship data to assess fraud risks in childcare benefit claims, leading to systemic biases in their selection processes and significant distress for the affected individuals [10, 31].

Over the years, numerous techniques aimed at mitigating bias have been developed to enhance the model fairness while maintaining data utility. These methodologies generally fall into three distinct categories based on the stage at which fairness interventions are applied [4]. Pre-processing technique involve adjusting the data prior to its use in model training, thus ensuring the data itself adheres to fairness standards. In-processing methods entail modifications to the model architecture of the learning process itself, guiding the model's outputs to comply with fairness constraints. Examples of in-processing techniques include approaches that learn fair representations directly from biased dataset [33, 32, 16], as well as generative models specifically designed to produce fair synthetic samples [23, 17, 29, 24, 18, 25]. Post-processing approaches adjust the model's outcome after the training phase, ensuring that the results are equitable across different demographic groups [4].

While these approaches address bias along individual protected attributes, such as gender or race, considered in isolation, discrimination often manifests at the intersections of multiple protected characteristics, a phenomenon known as intersectional bias. Buolamwini et al. [3] show that evaluations of commercial facial analysis systems found error rates as high as 34.7% for darker-skinned females, while lighter-skinned males had error rates as low as 0.8%. Crucially, analyzing gender and skin type separately hides these differences, only through intersectional evaluation across these demographic combinations did the true extent of bias become clear, showing how discrimination can be much worse for people who belong to multiple groups at the same time.

Ensuring fairness across multiple protected attributes (intersectional fairness) is technically more complex than working with a single attribute. First, the number of intersectional groups increases rapidly as more attributes are considered, which often leads to data sparsity and unreliable statistics in many real-world datasets. Second, enforcing fairness across these overlapping groups results in complicate optimization problems that are challenging to solve effectively. Some recent works tried to address the intersectional bias issue, e.g. Jang et al. [12] tackle fairness through representation learning using

* Corresponding Author
Email: md.fahim.sikder@liu.se, fahimsikder01@gmail.com

a non-zero-sum adversarial approach that enforces non-separability across sensitive attributes and separability for class labels. RELIANT [7] and FairGKD [34] use graph neural networks to mitigate biases. However, these graph-based approaches are not directly applicable to tabular data settings where intersectional fairness research is predominantly conducted, creating a gap for intersectional fairness methods in non-graph domains.

In this work, we introduce a novel framework for learning fair representations that mitigate intersectional bias by bridging the trade-off between predictive performance and fairness. Our approach leverages knowledge distillation from a high-capacity teacher model into a student model trained with explicit fairness objectives. Rather than jointly optimizing for accuracy and fairness, a task that becomes increasingly difficult in the intersectional settings, our approach decouples these objectives across two models. A teacher model focuses exclusively on learning meaningful predictive representations, unconstrained by fairness requirements. A student model then distills knowledge from the teacher while incorporating explicit constraints to ensure fairness across intersectional groups. This separation of concerns allows each model to specialize, leading to more effective optimization and ultimately better fairness-accuracy trade-offs.

Our key insight is that knowledge distillation provides an ideal mechanism for transferring predictive power while allowing for selective adaptation to satisfy fairness constraints. Specifically, by inheriting useful representations from the teacher, the student can focus its learning capacity on addressing fairness considerations rather than rediscovering predictive patterns from scratch. This approach is particularly valuable in the intersectional setting, where the complex interplay between multiple protected attributes creates challenges for direct optimization approaches. While previous works have explored knowledge distillation for model compression and efficiency, our approach represents the systematic investigation of distillation specifically for intersectional fairness. We model the interactions among multiple protected attributes through tailored loss functions designed to mitigate bias patterns emerging from their intersectional structure.

Our framework targets two central fairness challenges in multiclass classification: disparities in false positive rates (FPR) and violations of demographic parity across intersectional groups. These concerns are especially critical in domains such as criminal justice and financial services, where misclassifications can reinforce structural disadvantage. To mitigate these issues, we integrate fairness-driven loss terms into the training objective that penalize deviations in group-level statistics. By quantifying and minimizing gaps in metrics like FPR and positive prediction rates, the framework promotes equitable outcomes across diverse sub-populations. Furthermore, its modular design allows practitioners to tailor fairness objectives to specific deployment contexts without altering the core architecture.

Empirically, we evaluate our framework on several benchmark datasets, including MIMIC-III, MIMIC-IV, COMPAS, Adult-Income, and German Credit. Our evaluation demonstrate that knowledge distillation consistently outperforms direct fairness optimization across multiple metrics. We achieve 52% increase in accuracy for multiclass-classification, 61% reduction in average false positive rate across intersectional groups compared to state-of-the-art method. Beyond empirical performance, our framework offers practical advantages for deployment. The student model is more compact than the teacher, reducing computational requirements for inference.

The contributions of this paper are as follows:

1. We propose a novel two-stage framework that leverages knowledge distillation to transfer predictive performance from a teacher model to a fair student model, ensuring high utility while enabling fairness interventions.
2. We introduce a modular fairness constraint architecture for the student model, incorporating multiple fairness objectives, including false positive rate (FPR) parity, demographic parity, and conditional independence (targeted specifically at intersectional subgroups).
3. We design an intersectional fairness loss that simultaneously penalizes disparities in group-level FPRs and demographic rates while minimizing their absolute values, promoting both fairness and calibration across intersectional identities.
4. We empirically demonstrate that our approach achieves a superior trade-off between predictive accuracy and fairness compared to prior works, including adversarial representation learning (FSNS) [12], and diffusion-based Fair model [24] on real-world tabular datasets.

2 Related Works

2.1 Fairness in Machine Learning

In recent years, researchers have introduced multiple approaches for deriving fair representations from datasets exhibiting biases. For instance, Learning Fair Representations (LFR) formulates representation learning as an optimization challenge, demonstrating fairness in subsequent predictive tasks [33]. FairDisco pursues fairness by explicitly minimizing the variance correlation between sensitive and non-sensitive attributes, thereby creating bias-free representations [18]. Additionally, adversarial techniques have emerged, incorporating an adversary to identify and challenge potentially biased outcomes, thus enforcing fairness constraints [8, 19, 32].

Along these approaches, a number of fair generative modeling techniques have been introduced to synthesize data that preserves utility while reducing bias. For instance, TabFairGAN [23] incorporates fairness constraints into the training of generative adversarial networks to produce equitable tabular data. Similarly, FLDGMs [24] leverage adversarial and diffusion-based mechanisms to generate fair representation in the latent space. While these models show promising results on fairness and predictive capability but they only work with single sensitive attributes and sometimes only on binary classification tasks [23, 18, 24].

2.2 Knowledge Distillation in Fairness Research

Knowledge distillation has emerged as a powerful approach for addressing bias. Recent works such as RELIANT [7] and FairGKD [34] use distillation techniques to transfer knowledge from complex teacher models to student architectures while adding fairness constraints. However, these approaches primarily operate on graph-structured data, restricting their applicability to domains where relationships between entities can be naturally represented as graphs, and neither specifically targets intersectional fairness across multiple protected attributes. Our framework addresses intersectional fairness in tabular data through knowledge distillation. We introduce intersectional fairness constraints and use a teacher-student framework where the teacher provides accurate representations while the student jointly optimizes for fairness and utility.

3 Preliminaries

This section provides the foundational concepts and formal definitions necessary to understand our approach to intersectional fairness

through knowledge distillation.

3.1 Problem Formulation

We address the challenge of learning fair representations for multi-class classification tasks and promote intersectional fairness through knowledge distillation. Given a feature space $\mathcal{X} \subseteq \mathbb{R}^d$, a label space $\mathcal{Y} = \{0, 1, .., N-1\}$, and a joint protected attribute space $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \times \cdots \times \mathcal{S}_m$, our primary objective is to learn fair and robust representations that mitigate bias across intersectional groups.

Specifically, we aim to develop a framework where knowledge is distilled from a high-capacity teacher model $T = (f_T, g_T)$ to a student model $S_t = (f_S, g_S)$, where $f_T : \mathcal{X} \to \mathcal{Z}$ and $f_S : \mathcal{X} \to \mathcal{Z}$ are encoders mapping inputs to a latent representation space $\mathcal{Z} \subseteq \mathbb{R}^k$, and $g_T, g_S : \mathcal{Z} \to [0, N-1]$ are classifiers making predictions based on these representations.

The key insight of our approach is that representation learning for intersectional fairness can be more effectively accomplished through distillation, the teacher model learns useful predictive representations without fairness constraints, while the student model inherits this knowledge but adapts it to satisfy fairness requirements across intersectional groups, defined by specific combinations of protected attributes $g \in \mathcal{S}$.

3.2 Fairness Definition

Artificial Intelligence-based models can retain and amplify biases present in the training data particularly against marginalized demographic groups. We seek to mitigate these biases through various metrics and constraints and in this study we focus on two fairness metrics particularly relevant to decision making systems.

Definition 1 (False Positive Rate Parity)**.** *A classifier $h : X \to \{0, 1, .., N-1\}$ with N classes, satisfies false positive rate parity with respect to sensitive attribute $\mathcal{S}$ if, for each class c:*

$$\Pr(h(X) = c \mid Y \neq c, S = s) = \Pr(h(X) = c \mid Y \neq c, S = s') \quad \forall s, s' \in \mathcal{S} \tag{1}$$

where X denotes feature, Y denotes label and $Pr(.)$ is the probability.

Definition 2 (Demographic Parity Ratio, Weerts et al. [30])**.** *A classifier $h : X \to \{0, 1, .., N-1\}$ with N classes, the Demographic Parity Ratio (DPR) for class c across sensitive attribute value is defined as:*

$$DPR = \frac{\Pr(h(X) = c \mid S = s)}{\Pr(h(X) = c \mid S = s')} \quad \forall s, s' \in \mathcal{S},\ \forall c \in \mathcal{Y} \tag{2}$$

3.3 Intersectional Fairness

Intersectionality refers to the unique situation when aspects of social and political identity (e.g. gender, race) combine to create different modes of discrimination or privilege [5]. In our study, this translates to ensuring equal outcome across combination of sensitive attributes.

Definition 3 (Intersectional Group, Jang et al. [12])**.** *Given m protected attributes $S_1, S_2, ..., S_m$, an intersectional group $g \in \mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \times \cdots \times \mathcal{S}_m$ represents a specific combination of sensitive attribute values.*

3.4 Knowledge Distillation

Knowledge distillation is a form of transfer learning where a compact student model is trained to replicate the behavior of a larger, pre-trained teacher model [11]. One common approach, known as response-based distillation, involves aligning the student's output logits with those of the teacher, effectively encouraging the student to imitate the teacher's soft predictions [11, 15, 6]. This is typically achieved by combining the standard cross-entropy loss with a Kullback Leibler (KL) divergence term applied to the softmax outputs of both models.

$$\mathcal{L}_{\text{distillation}} = \mathcal{L}_{\text{CE}}(y, \hat{y}) + \mathcal{L}_{\text{KL}}(p, q) \tag{3}$$

Here, $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss and $\mathcal{L}_{\text{KL}}$ is the KL-divergence loss.

4 Methodology

In this section, we present the detail of our framework. First we give an overview of the framework, followed by the model architecture (both Teacher and Student model), and then we present our fairness loss functions.

4.1 Framework Overview

We propose a knowledge distillation framework for intersectional fairness that operates in two phases. First, a high-capacity teacher model is trained solely for predictive performance without fairness constraints. Then, a student model learns from the teacher while incorporating explicit fairness constraints for intersectional groups. This approach separates the objectives of learning predictive representations and ensuring fairness, leading to more effective optimization and robust performance.

The key insight of our framework is that directly optimizing for both accuracy and fairness across intersectional groups leads to complex optimization challenges. By distilling knowledge from a teacher model trained for accuracy, the student model inherits powerful representations while adapting them to satisfy fairness constraints. This approach is particularly effective for addressing false positive rate (FPR) disparities and demographic parity disparities across intersectional groups. We further incorporates an adversarial component that operates along the distillation process to further enhance fairness. This adversarial mechanism attempts to predict sensitive attributes from the learned representations, while gradient reversal layer encourages the student encoder to produce representations that actively resist such prediction [22]. This adversarial training dynamic creates an implicit regularization effect that promotes the removal of sensitive information related to projected attributes from the latent space.

4.2 Model Architecture

4.2.1 Teacher Model

The teacher model consists of an encoder f_T and a classifier g_T. The encoder is implemented as a three-layer multi-layer perceptron (MLP) with ReLU activation and batch normalization after each linear transformation. Formally, we define:

$$f_T(\mathbf{x}) = f_3(f_2(f_1(\mathbf{x})))$$

where:

$$f_1(\mathbf{x}) = \text{ReLU}(\text{BN}(W_1^T \mathbf{x} + b_1^T))$$

$$f_2(\mathbf{h}_1) = \text{ReLU}(\text{BN}(W_2^T \mathbf{h}_1 + b_2^T))$$

$$f_3(\mathbf{h}_2) = W_3^T \mathbf{h}_2 + b_3^T$$

Here, W_i and b_i are learnable parameters, BN denotes batch normalization, and ReLU is the rectified linear activation. The classifier g_T is similarly structured where V_i and c_i as learnable parameters:

$$g_T(\mathbf{z}) = V_3^T(\text{ReLU}(\text{BN}(V_2^T(\text{ReLU}(\text{BN}(V_1^T\mathbf{z}+c_1^T)))+c_2^T)))+c_3^T$$

4.2.2 Student Model

The student model employs a more compact architecture with an encoder f_S and classifier g_S:

$$f_S(\mathbf{x}) = W_2^S(\text{ReLU}(\text{BN}(W_1^S\mathbf{x} + b_1^S))) + b_2^S$$

$$g_S(\mathbf{z}) = \sigma(V_2^S(\text{ReLU}(\text{BN}(V_1^S\mathbf{z} + c_1^S))) + c_2^S)$$

The reduced architecture contributes to model efficiency while still maintaining sufficient capacity to learn fair representations.

4.2.3 Adversarial Component

We incorporate an adversarial component to further enhance fairness. This component, $d_\psi : \mathcal{Z} \to \mathbb{R}^{|\mathcal{S}|}$, attempts to predict protected attributes from the latent representations:

$$d_\psi(\mathbf{z}) = U_3(\text{ReLU}(U_2(\text{ReLU}(U_1\mathbf{z} + e_1)) + e_2)) + e_3$$

Through gradient reversal [22], the encoder is trained to produce representations that prevent the adversary from accurately predicting protected attributes.

4.3 Fairness-Aware Distillation

Our training process involves first training the teacher model to learn data representation to improve predictive performance. Then we employ multiple specialized loss functions designed to balance predictive performance and intersectional fairness during the student model training (distillation process).

4.3.1 Teacher Training

The teacher model is trained to minimize standard cross-entropy loss involving teacher encoder f_T and classifier g_T parameterized by θ_T and ϕ_T respectively:

$$\mathcal{L}_T(\theta_T, \phi_T) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C} y_{i,c} \log\left([g_T(f_T(x_i))]_c\right) \tag{4}$$

This ensures that the teacher learns representations optimized purely for predictive performance.

4.3.2 Fairness Loss Functions

In this section, we present several specialized loss functions for fairness:

FPR Fairness Loss This penalizes disparities in class-specific false positive rates across intersectional groups. For a given class $c \in \{1, ..., C\}$ and group $g \in \mathcal{S}$, the per-group false positive rate is defined as:

$$\text{FPR}_{g,c} = \frac{\sum_{i:s_i=g,\ y_i\neq c} \mathbb{1}\left[\arg\max g_S(f_S(x_i)) = c\right]}{\sum_{i:s_i=g,\ y_i\neq c} 1} \tag{5}$$

Here, $\mathbb{1}$ is an indicator function and $\text{FPR}_{g,c}$ measures the proportion of non-class-c examples in group g that are incorrectly predicted as class c. To encourage the FPR parity across groups and classes, we define the following FPR fairness loss:

$$\mathcal{L}_{\text{FPR}}(\theta_S, \phi_S) = \sum_{c=1}^{C}\Bigg[\underbrace{\text{Var}_{g\in\mathcal{S}}(\text{FPR}_{g,c})}_{\text{variance}} + \underbrace{\max_{g,g'}|\text{FPR}_{g,c} - \text{FPR}_{g',c}|}_{\text{max gap}} + \underbrace{\sum_{g\in\mathcal{S}} w_{g,c}\cdot\text{FPR}_{g,c}^2}_{\text{magnitude penalty}}\Bigg] \tag{6}$$

The three components of the loss function are designed to (1) minimize the variance of FPR across groups, (2) reduce the maximum pairwise difference in FPR, and (3) drive the absolute magnitude of FPR toward zero for all groups. The weights $w_{\mathbf{s}}$ account for potential imbalances in group sizes.

Demographic Parity Loss This loss penalizes disparities in class-specific positive prediction rates (PPR) across intersectional groups. For a group $g \in \mathcal{S}$ and class $c \in \{1, ..., C\}$ the positive prediction rate is defined as:

$$\text{PPR}_{g,c} = \frac{\sum_{i:\,\mathbf{s}_i=g} \mathbb{1}\left[\arg\max g_S(f_S(x_i)) = c\right]}{\sum_{i:\,\mathbf{s}_i=g} 1} \tag{7}$$

This measures the proportion of group g's samples that are predicted as class c. To promote demographic parity across all classes and groups, we define the following loss:

$$\mathcal{L}_{\text{DP}}(\theta_S, \phi_S) = \sum_{c=1}^{C}\Bigg[\underbrace{\text{Var}_{g\in\mathcal{S}}(\text{PPR}_{g,c})}_{\text{variance}} + \underbrace{\max_{g,g'\in\mathcal{S}}|\text{PPR}_{g,c} - \text{PPR}_{g',c}|}_{\text{max gap}} + \underbrace{\sum_{g,g'\in\mathcal{S}} w_{g,g'}\cdot(\text{PPR}_{g,c} - \text{PPR}_{g',c})^2}_{\text{pairwise penalty}}\Bigg] \tag{8}$$

The three components of the loss functions are similar like the FPR fairness loss where we minimize the variance of PPR, reduce the maximum pairwise difference in PPR and finally applies a weighted penalty for all group-pair differences, ensuring a smoother and more distributed alignment of PPRs across groups.

Conditional Independence Loss This loss promotes independence between protected attributes and representations, conditioned

on the true label. This is motivated by the fairness criterion of Equality of Odds which requires that prediction (or representations) be independent of sensitive attributes once the true label is known [9]. Formally, the loss is defined as:

$$\mathcal{L}_{\text{CI}}(\theta_S) = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \sum_{j=1}^{m} \|\text{Corr}(f_S(\mathbf{X}^y), S_j^y)\|_1 \tag{9}$$

where $\mathbf{X}^y$ and S_j^y represent features and protected attributes for samples with label y, and Corr$(,)$ computes the correlation matrix.

Adversarial Loss The Adversarial loss consists of two components. The first trains the adversary to predict protected attributes:

$$\mathcal{L}_{\text{adv}}^{d}(\psi) = \frac{1}{n} \sum_{i=1}^{n} \|d_\psi(f_S(\mathbf{x}_i)) - \mathbf{s}_i\|_2^2 \tag{10}$$

The second performs gradient reversal to encourage the encoder to remove protected attribute information [22]:

$$\mathcal{L}_{\text{adv}}^{f}(\theta_S) = -\lambda_{\text{adv}} \cdot \frac{1}{n} \sum_{i=1}^{n} \|d_\psi(f_S(\mathbf{x}_i)) - \mathbf{s}_i\|_2^2 \tag{11}$$

where λ_{adv} controls the strength of the adversarial component. The gradient reversal is performed on the student encoder.

4.3.3 Distillation Loss Functions

We employ two distillation losses to transfer knowledge from the teacher to the student:

Representation Distillation Loss The representation distillation loss encourages the student's encoder to learn representations to align with the teacher's. It operates on L2 normalized representation to emphasize angular similarity (cosine-like), independent of magnitude [11]:

$$\mathcal{L}_{\text{repr}}(\theta_S, \theta_T) = \frac{1}{n} \sum_{i=1}^{n} \left\| \frac{f_S(\mathbf{x}_i)}{\|f_S(\mathbf{x}_i)\|_2} - \frac{f_T(\mathbf{x}_i)}{\|f_T(\mathbf{x}_i)\|_2} \right\|_2^2 \tag{12}$$

Prediction Distillation Loss The prediction distillation loss transfers the teacher model's class distribution to the student by minimizing the KL divergence between their softened predictions [11]:

$$\mathcal{L}_{\text{pred}}(\theta_S, \phi_S, \theta_T, \phi_T) = \frac{1}{n} \sum_{i=1}^{n} \tau^2 \cdot \text{KL}\left(\frac{g_S(f_S(\mathbf{x}_i))}{\tau} \middle\| \frac{g_T(f_T(\mathbf{x}_i))}{\tau} \right) \tag{13}$$

Here, $\tau > 1$, is a temperature parameter that smooths the output logits to better capture class similarities. The loss encourages the student to mimic the teacher's probabilistic output distribution, improving generalization and training stability.

4.3.4 Combined Training Objective

The student model is trained to minimize a weighted combination of all loss functions:

$$\begin{aligned}\mathcal{L}_S(\theta_S, \phi_S, \psi) = {} & \alpha_1 \mathcal{L}_{\text{cls}} + \alpha_2 \mathcal{L}_{\text{repr}} + \alpha_3 \mathcal{L}_{\text{pred}} \\ & + \alpha_4 \mathcal{L}_{\text{FPR}} + \alpha_5 \mathcal{L}_{\text{DP}} + \alpha_6 \mathcal{L}_{\text{CI}} + \alpha_7 \mathcal{L}_{\text{adv}}^{f}\end{aligned} \tag{14}$$

Here, $\mathcal{L}_{\text{cls}}$ is the defined like Eq. 4, but applied to the student encoder and classifier instead of teacher's. And, α_1 through α_7 are hyperparameters controlling the relative importance of each loss component. We use `Hyperopt` [2], a bayesian optimization based grid search library to find optimal hyperparameters for balancing accuracy and fairness objectives effectively. More details about the hyperparameters can be found in Table 5 of the supplementary document [27].

5 Experiments

In this section, we present the experimental setup and result of our framework that leverages an encoder to create fair representation from an biased teacher model through distillation technique. We evaluated our framework using five distinct datasets, measuring the performance through various evaluation metrics from the perspective of fairness and data utility including False Positive Rate (FPR), Demographic Parity Ratio (DPR). We compare the performance of our framework with several state-of-the-art models. We also perform an ablation study to show the effectiveness and necessity of the different loss function component we use in our framework. While training the benchmark, we use the same hyperparameters as in their respective original publication. For loading the dataset and training, we use a fairness benchmarking tool called FairX [26]. Evaluation is done using held-out testing dataset.

5.1 Datasets

We evaluate on five benchmark datasets spanning diverse domains: Adult-Income, COMPAS [1], German Credit, and healthcare datasets MIMIC-III [13] and MIMIC-IV [14]. The MIMIC datasets are configured for multi-class classification to demonstrate effectiveness beyond binary prediction. Additional dataset details are in the supplementary material [27].

5.2 Baselines

To assess the quality of our approach, we conducted extensive comparisons against prominent fairness-aware frameworks across various domains. Our evaluation benchmarks include adversarial representation learning (FSNS) [12], which employs non-zero-sum adversarial techniques to enforce attribute non-separability. Generative approaches like TabFairGAN [23], which incorporates fairness constraints into GANs for tabular data, representation learning methods such as FairDisco [18], which minimizes variance correlation between sensitive and non-sensitive attributes, and diffusion-based approaches like FLDGMs [24], which leverage latent generative mechanisms for fairness-aware data synthesis.

While knowledge distillation for fairness would seem a logical point of comparison of our method, however current approaches in this domain primarily target graph-structured data [7, 34]. For instance, model such as RELIANT [7] and FairGKD [34] uses distillation principles to enhance fairness, but their architectural designs specifically target graph-structured data, which is fundamentally different data modality than our framework's current focus. Consequently, we have selected baseline comparators that aligns with our data domain while still representing diverse fairness enhancement paradigms.

5.3 Evaluation Metrics

We evaluate the proposed framework along several primary criteria:

- **Fairness:** To assess fairness, we measure the *False Positive Rate (FPR)* across intersectional groups and compute the *Demographic Parity Ratio (DPR)* for a downstream classification task (see Definition 2).
- **Predictive Utility:** To ensure that improvements in fairness do not come at the expense of model performance, we evaluate predictive utility using standard metrics: *Accuracy* on the downstream task.

Table 1. Multi-class Classification on MIMIC-III and MIMIC-IV dataset, Here 'G' represents Gender and 'E' represents Ethnicity (bold indicates best result, ↑ indicates higher as better, ↓ indicates lower as better)

Dataset	Evaluation	FLDGMs	FSNS	TabFair GAN	Ours
MIMIC-III	Accuracy(↑)	0.69	0.44	0.59	**0.92**
	DPR (G)(↑)	0.94	0.95	0.97	**0.98**
	DPR (E)(↑)	0.20	0.03	0.58	**0.70**
MIMIC-IV	Accuracy(↑)	0.70	0.53	0.56	**0.97**
	DPR (G)(↑)	0.94	0.95	**0.97**	**0.97**
	DPR (E)(↑)	0.41	0.06	0.00	**0.71**

5.4 Results

In this section, we present the empirical evaluation of our proposed framework. We report results on both fairness and predictive utility metrics, as described in Section 5.3. Specifically, we compare our method against several baseline approaches to assess its effectiveness in mitigating intersectional bias while preserving predictive performance. Detailed results are provided in Table 3, 4, and supplementary Table 2 [27] for binary classification with intersectional fairness evaluation. In addition, Table 1 along with Supplementary Table 3 and 4 [27] presents the evaluation of multi-class classification in the intersectional fairness settings.

Our findings shows that the student models distilled with intersectional fairness constraints consistently achieve lower False Positive Rate and improved Demographic Parity Ratio across sensitive groups, while maintaining competitive accuracy relative to the baselines. On COMPAS, we achieve 61% reduction in FPR disparities and 5% improvement in Demographic Parity Ratios compared to FSNS (Table 4), with minimal accuracy trade-off (5% decrease). Notably, these fairness gains are achieved while preserving competitive classification accuracy relative to state-of-the-art alternatives suggests our distillation framework successfully addresses the trade-off between algorithmic fairness and predictive performance.

Table 2. Ablation Study, COMPAS Dataset, Here 'G' represents Gender and 'R' represents Race (bold indicates best result, ↑ indicates higher as better, ↓ indicates lower as better)

Loss	FPR GAP (↓)	DPR(G) (↑)	DPR(R) (↑)	Accuracy (↑)
Teacher	0.17	0.77	0.83	**0.63**
No Adversarial	0.05	0.00	0.00	0.62
No Adversarial + No Fairness	0.10	0.47	0.52	0.62
Full Model	**0.04**	**0.93**	**0.97**	0.61

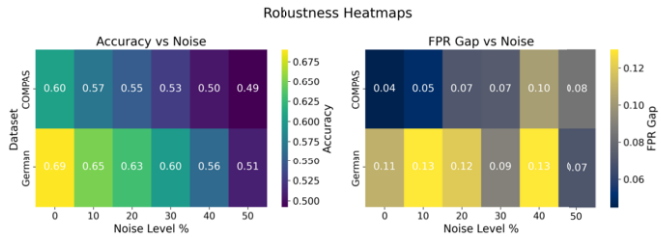

Figure 1. Robustness of our Model, Accuracy vs Noise level (0-50%) and FPR GAP vs Noise level

5.4.1 Multi-class Classification

While most fairness methods focus on binary classification, real-world decision-making frequently involves multi-class outcomes. We evaluate our framework's multi-class capability using MIMIC-III and MIMIC-IV datasets for ICU length of stay prediction (four classes). The non-binary "Ethnicity" attribute creates a challenging intersectional setting.

Our model achieves strong performance on both fairness and predictive metrics (Table 1). We achieve 52% accuracy increase in MIMIC-III compared to FSNS and 96% improvement in Demographic Parity Ratio for "Ethnicity", demonstrating our framework's effectiveness beyond binary classification. Additional FPR results are in supplementary Table 3 and 4 [27].

5.4.2 Model Robustness

We evaluate robustness under label noise (0-50%) on COMPAS and German Credit datasets (Figure 1). Our framework demonstrates notable resilience: COMPAS accuracy degrades moderately from 0.60 to 0.49 (18% relative decrease) while FPR gap remains stable (0.045 to 0.084, peak 0.10 at 40% noise). German Credit shows similar patterns. This stability indicates fairness guarantees are not dependent on perfect labeling, demonstrating applicability in real-world settings with imperfect data.

5.4.3 Ablation Study

We conduct an ablation study on COMPAS to assess component contributions and presented the result in Table 2. The teacher model achieves highest accuracy of 0.63 but violates fairness with FPR gap of 0.17 and DPR of 0.77 (gender) and 0.83 (race). Removing adversarial components reduces FPR gap but eliminates demographic parity (DPR = 0.00), confirming adversarial training's necessity for balancing multiple fairness criteria. Our full model optimally balances metrics with lowest FPR gap of 0.04 and near-perfect DPR of 0.93 (gender) and 0.97 (race), while maintaining competitive accuracy of 0.61, demonstrating that all components are essential for effective intersectional fairness.

5.4.4 Computational Efficiency

Our approach distills fairness-aware representations into compact student models with significantly fewer parameters. For COMPAS, the student network uses 30,016 parameters versus 76,736 in the teacher (2.56× reduction) while maintaining competitive accuracy and fairness performance. This compression reduces memory usage,

Table 3. Evaluation on Adult-Income Dataset (bold indicates best result, ↑ indicates higher as better, ↓ indicates lower as better

Race	Gender	FLDGMs	FairDisco	TabFairGAN	FSNS	Baseline	Ours
Subgroup FPRs (↓)							
Non-white	Female	0.21	0.14	0.20	0.02	0.10	**0.00**
Non-white	Male	0.19	0.11	0.07	0.05	0.12	**0.01**
White	Female	0.20	0.13	0.27	0.03	0.07	**0.01**
White	Male	0.18	0.12	0.07	0.11	0.15	**0.01**
Overall Metrics							
Accuracy (↑)		0.65	0.68	0.70	**0.81**	0.60	**0.81**
DPR Race (↑)		0.90	0.91	0.26	0.89	0.51	**0.92**
DPR Gender (↑)		0.86	**0.87**	0.69	0.80	0.45	0.84

Table 4. Evaluation on COMPAS Dataset (bold indicates best result, ↑ indicates higher as better, ↓ indicates lower as better)

Gender	Race	FLDGMs	FairDisco	TabFairGAN	FSNS	Baseline	Ours
Subgroup FPRs (↓)							
Male	Non-Caucasian	0.82	0.91	0.10	0.22	0.36	**0.04**
Male	Caucasian	0.83	0.91	0.36	0.08	0.26	**0.06**
Female	Non-Caucasian	0.69	0.93	0.06	0.05	0.23	**0.02**
Female	Caucasian	0.88	0.90	0.38	0.04	0.20	**0.03**
Overall Metrics							
Accuracy (↑)		0.45	0.47	0.56	**0.64**	0.63	0.61
DPR Gender (↑)		0.85	0.83	0.90	0.89	0.76	**0.93**
DPR Race (↑)		0.90	0.91	0.51	0.93	0.62	**0.97**

accelerates inference, and lowers training costs, making our framework suitable for resource-constrained environments where both fairness and efficiency are critical. Additional computational details are in supplementary Table 1 [27].

6 Discussion

Our experimental results demonstrate that knowledge distillation provides an effective framework for addressing intersectional fairness challenges while maintaining competitive predictive performance. By decoupling accuracy and fairness objectives across teacher and student models, we achieve more stable optimization than direct fairness approaches, particularly in complex intersectional settings.

Several key insights emerge from our findings. First, the student model's ability to inherit useful representations from the teacher while selectively adapting them to satisfy fairness constraints provides a more balanced approach to the fairness-accuracy trade-off than single-stage optimization methods. This is particularly evident in the robustness analysis, where our framework maintains fairness guarantees even under significant label noise, a critical consideration for real-world deployments where perfect labeling cannot be assumed.

Limitations & Future Works While our approach shows promising results, we currently requires explicit identification of sensitive attributes during training, like most fairness-aware methods in the literature [12, 18, 24]. This design choice enables precise quantification and mitigation of disparities across intersectional groups. In our future work, we would like to explore privacy-preserving adaptations of our framework that operate within limited or absent demographic information.

EU Regulatory Alignment Our framework offers particular relevance in the European context, where recent regulatory developments have placed increasing emphasis on algorithmic fairness and non-discrimination. The EU AI Act, which classifies systems used for evaluating creditworthiness, recruitment, education access, and criminal justice as "high-risk", imposes strict requirements for fairness, transparency, and human oversight [28]. Our approach directly addresses these requirements through its explicit fairness guarantees.

7 Conclusion

In this work, we introduce a novel knowledge distillation framework that effectively addresses the complex challenge of intersectional fairness in machine learning systems. By decoupling the optimization of predictive performance and fairness objectives across teacher and student models, our approach provides a more stable optimization pathway than direct fairness methods. The student model inherits valuable predictive knowledge while incorporating specialized fairness constraints, enabling superior performance across multiple fairness metrics without significant accuracy degradation. Our comprehensive empirical evaluation demonstrates that this approach yields significant improvements in intersectional fairness across diverse domains, reducing FPR disparities and improving demographic parity ratios while maintaining competitive accuracy. The framework's resilience to label noise further underscores its potential for real-world deployment. Moreover, the student model's parameter efficiency offers practical computational advantages for resource-constrained environments. This work opens several promising research directions, including the development of privacy-preserving variants. By bridging the gap between fairness and utility through knowledge distillation, our framework contributes to the ongoing effort to develop AI-based systems that make equitable decisions across all demographic intersections.

Acknowledgements

This work was funded by the Knut and Alice Wallenberg Foundation, Sweden, the ELLIIT Excellence Center at Linköping-Lund for

Information Technology, Sweden (portions of this work were carried out using the AIOps/Stellar). The computations were enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

References

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016. [Online; accessed 12-Aug-2024].

[2] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, and D. D. Cox. Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, 8(1):014008, 2015.

[3] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[4] S. Caton and C. Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38, 2024.

[5] K. W. Crenshaw. Mapping the margins: Intersectionality, identity politics, and violence against women of color. In *The public nature of private violence*, pages 93–118. Routledge, 2013.

[6] X. Dai, Z. Jiang, Z. Wu, Y. Bao, Z. Wang, S. Liu, and E. Zhou. General instance distillation for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7842–7851, 2021.

[7] Y. Dong, B. Zhang, Y. Yuan, N. Zou, Q. Wang, and J. Li. Reliant: Fair knowledge distillation for graph neural networks. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, pages 154–162. SIAM, 2023.

[8] X. Gao, J. Zhai, S. Ma, C. Shen, Y. Chen, and Q. Wang. Fairneuron: improving deep neural network fairness with adversary games on selective neurons. In *Proceedings of the 44th International Conference on Software Engineering*, pages 921–933, 2022.

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

[10] M. Heikkilä. Dutch scandal serves as a warning for europe over risks of using algorithms. https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms, 2021. [Online; accessed 12-Aug-2024].

[11] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *stat*, 1050:9, 2015.

[12] T. Jang, H. Gao, P. Shi, and X. Wang. Achieving fairness through separability: a unified framework for fair representation learning. In *International Conference on Artificial Intelligence and Statistics*, pages 28–36. PMLR, 2024.

[13] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1): 1–9, 2016.

[14] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10 (1):1, 2023.

[15] S. Jung, D. Lee, T. Park, and T. Moon. Fair feature distillation for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12115–12124, 2021.

[16] M. Knobbout. ALFR++: A Novel Algorithm for Learning Adversarial Fair Representations. In *ECAI 2023*, pages 1280–1287. IOS Press, 2023.

[17] J. Li, Y. Ren, and K. Deng. Fairgan: Gans-based fairness-aware learning for recommendations with implicit feedback. In *Proceedings of the ACM web conference 2022*, pages 297–307, 2022.

[18] J. Liu, Z. Li, Y. Yao, F. Xu, X. Ma, M. Xu, and H. Tong. Fair representation learning: An alternative to mutual information. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1088–1097, 2022.

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

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

[21] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, et al. Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356, 2020.

[22] E. Raff and J. Sylvester. Gradient reversal against discrimination: A fair neural network learning approach. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 189–198. IEEE, 2018.

[23] A. Rajabi and O. O. Garibay. Tabfairgan: Fair tabular data generation with generative adversarial networks. *Machine Learning and Knowledge Extraction*, 4(2):488–501, 2022.

[24] R. Ramachandranpillai, M. F. Sikder, and F. Heintz. Fair Latent Deep Generative Models (FLDGMs) for Syntax-Agnostic and Fair Synthetic Data Generation. In *ECAI 2023*, pages 1938–1945. IOS Press, 2023.

[25] R. Ramachandranpillai, M. F. Sikder, D. Bergström, and F. Heintz. Bt-GAN: Generating Fair Synthetic Healthdata via Bias-transforming Generative Adversarial Networks. *Journal of Artificial Intelligence Research (JAIR)*, 79:1313–1341, 2024.

[26] M. F. Sikder, R. Ramachandranpillai, D. de Leng, and F. Heintz. Fairx: A comprehensive benchmarking tool for model analysis using fairness, utility, and explainability. In *Proceedings of the 2nd Workshop on Fairness and Bias in AI, co-located with ECAI 2024*. CEUR-WS.org/Vol-3808, 2024.

[27] M. F. Sikder, R. Ramachandranpillai, D. de Leng, and F. Heintz. Promoting intersectional fairness through knowledge distillation - supplementary materials. *10.5281/zenodo.15866767*, 2025.

[28] N. A. Smuha. Regulation 2024/1689 of the eur. parl. & council of june 13, 2024 (eu artificial intelligence act). *International Legal Materials*, pages 1–148.

[29] B. Van Breugel, T. Kyono, J. Berrevoets, and M. Van der Schaar. Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems*, 34:22221–22233, 2021.

[30] H. Weerts, M. Dudík, R. Edgar, A. Jalali, R. Lutz, and M. Madaio. Fairlearn: Assessing and improving fairness of ai systems. *Journal of Machine Learning Research*, 24(257):1–8, 2023.

[31] H. J. Weerts, R. Theunissen, and M. C. Willemsen. Look and you will find it: Fairness-aware data collection through active learning. In *IAL@ PKDD/ECML*, pages 74–88, 2023.

[32] H. Xu, X. Liu, Y. Li, A. Jain, and J. Tang. To be robust or to be fair: Towards fairness in adversarial training. In *International conference on machine learning*, pages 11492–11501. PMLR, 2021.

[33] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[34] Y. Zhu, J. Li, L. Chen, and Z. Zheng. The devil is in the data: Learning fair graph neural networks via partial knowledge distillation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1012–1021, 2024.