Fairness-Aware Checkpoint Screening for Neural Models via Multi-Task Learning and Monte Carlo Dropout

Anonymous authors
Paper under double-blind review

Abstract

Machine learning models deployed in high-stakes domains often exhibit trade-offs between predictive performance and group fairness, and identifying models that navigate this trade-off remains challenging in practice. We present a neural in-processing framework that combines multi-task learning and Monte Carlo (MC) dropout to support uncertainty-aware checkpoint selection for fairness-aware prediction. Our approach jointly predicts a primary target and a protected attribute using a shared representation, then evaluates saved training checkpoints using predictive performance and a group-fairness objective based on disparate impact ratio. We use MC dropout to characterize checkpoint-level predictive variability and perform Pareto-based screening over fairness–performance trade-offs on a validation set, enabling selection of candidate checkpoints that better balance these competing objectives. We evaluate the approach on three datasets: ADULT, MIMIC-III, and SNAPSHOT, and compare against standard fairness baselines including reweighing, adversarial reweighted learning, and FairRF where applicable. Across these settings, the proposed selection strategy often identifies checkpoints with improved demographic-parity trade-offs relative to baseline models, while maintaining competitive predictive performance. We further provide qualitative saliency-map analyses to illustrate how feature emphasis may shift across selected checkpoints. Our results suggest that uncertainty-aware checkpoint screening can serve as a practical mechanism for navigating fairness–performance trade-offs in neural prediction pipelines. We discuss limitations, including dependence on neural architectures with MC dropout and the current focus on a demographic-parity-style fairness criterion.

1 Introduction

Bias in machine learning (ML), as initially framed by Mitchell (Mitchell, 1980), can support generalization but may also lead to unfair outcomes. When models are used in sensitive applications such as healthcare, hiring, or criminal justice, bias can disproportionately harm specific groups based on attributes like race, gender, or age (Zanna et al., 2022).

Bias may stem from data through unrepresentative sampling or labeling errors, or from algorithmic underestimation, where models under-predict rare outcomes in minority groups (Kamishima et al., 2012; Cunningham & Delany, 2021). Removing sensitive features alone is often ineffective, as models can infer them from correlated attributes (Paulus & Kent, 2020). These challenges underscore the importance of methods that address fairness while making model selection trade-offs more transparent.

A key difficulty in bias mitigation is balancing fairness with predictive performance (Buijsman, 2023; Pessach & Shmueli, 2022). Recent work has employed Pareto optimality from multi-objective optimization to model this trade-off without reducing multiple objectives to a single scalar value, as in traditional linear scalarization methods, which can miss significant trade-offs due to the assumption of linear relationships (Wei & Niethammer, 2022; Sener & Koltun, 2018; Liu & Vicente, 2022; Wang et al., 2021; Liang et al., 2022; Little et al., 2022; Kamani et al., 2021). Pareto-based approaches more effectively explore fairness-accuracy trade-offs by identifying optimal solutions across competing objectives.

As ML plays a more significant role in decision-making across industries, there is growing interest in fairness-aware methods that are both practically deployable and transparent about the trade-offs they impose (Rabonato & Berton, 2025; Özbulak et al., 2025). In many applications, the challenge is not only to reduce bias, but also to identify candidate models that navigate the tension between predictive performance and group fairness in a controlled and inspectable way. This motivates approaches that treat fairness-aware model selection itself as part of the learning pipeline.

In this work, we present a neural framework for fairness-aware checkpoint selection that combines multi-task learning (MTL), Monte Carlo (MC) dropout, and Pareto-based screening. We train an MTL model to jointly predict a primary target and a protected attribute using a shared representation, then save checkpoints across training (Qin et al., 2025). Rather than assuming that a single final epoch provides the best fairness-performance trade-off, we evaluate saved checkpoints on a validation set using both predictive performance and a group-fairness objective based on disparate impact ratio (DIR). MC dropout is used to obtain stochastic predictions and summarize checkpoint-level predictive variability during this screening process (Ahmed et al., 2023; Gal & Ghahramani, 2016; Schmal & Mäder, 2026). We then apply Pareto-based checkpoint screening over the resulting validation-set fairness and performance metrics to identify candidate checkpoints with favorable trade-offs, followed by a pre-specified selection rule for final model choice (Almasi et al., 2026; Nagpal et al., 2025; Wei & Niethammer, 2022).

This framing intentionally treats fairness-aware model selection as a practical multi-objective screening problem rather than as a claim that uncertainty alone guarantees fairness improvements. Unlike prior work that modifies training objectives or data distributions, we formalize fairness-aware model selection itself as an intervention, using checkpoint-level stochastic evaluation and Pareto-based screening. In our experiments, the combination of MTL, checkpoint-level stochastic evaluation, and Pareto-based screening often identifies checkpoints with improved demographic-parity trade-offs relative to baseline models while maintaining competitive predictive performance. However, we view this relationship as empirical rather than theoretically guaranteed, and we return to this limitation in Section 7.3.

We evaluate the framework using three datasets spanning distinct application domains: ADULT (finance), MIMIC-III (in-hospital mortality prediction), and SNAPSHOT (student wellbeing / affect-related prediction). These settings allow us to examine whether the same checkpoint-screening strategy can recover useful fairness-performance trade-offs across different neural prediction tasks. In addition to quantitative evaluation, we include saliency-map visualizations as a qualitative post hoc diagnostic to inspect how feature emphasis may differ between baseline and selected checkpoints (Simonyan et al., 2013). These analyses are intended as descriptive model-inspection tools rather than as causal evidence of the mechanism underlying fairness changes.

Our major contributions include:

- We propose a neural framework for fairness-aware prediction that combines MTL and MC dropout to support validation-based checkpoint screening for fairness-performance trade-offs.
- We formulate checkpoint selection as a multi-objective screening problem over validation-set predictive performance and a group-fairness criterion based on disparate impact ratio (DIR), using Pareto-based selection to identify candidate checkpoints.
- We empirically evaluate the framework across three application domains including finance (ADULT), healthcare (MIMIC-III), and student wellbeing (SNAPSHOT) to assess whether checkpoint-level screening can recover favorable demographic-parity trade-offs.
- We compare the selected checkpoints against established fairness baselines, including reweighing, adversarial reweighted learning (ARL), and FairRF where applicable.
- We provide saliency-map visualizations as qualitative post hoc diagnostics to examine how feature emphasis may differ between baseline and selected checkpoints, while emphasizing that these analyses are descriptive rather than causal.

2 Related Works

2.1 Fairness Challenges in Machine Learning

Fairness, long studied in philosophy and psychology, has become a central concern in ML as predictive systems are increasingly deployed in high-stakes settings such as healthcare, finance, education, and hiring (Pleiss et al., 2017; Saxena et al., 2019; Caton & Haas, 2024). In this context, fairness is commonly understood as the reduction or absence of systematic disadvantage in decisions affecting individuals or groups defined by sensitive attributes (Mehrabi et al., 2021; Caton & Haas, 2024). A substantial body of work has therefore focused on algorithmic interventions that mitigate bias while preserving predictive utility.

Existing bias mitigation strategies are commonly organized into three broad categories: pre-processing, in-processing, and post-processing methods (Calders & Verwer, 2010; Kamiran & Calders, 2012; Zafar et al., 2017; Caton & Haas, 2024). Pre-processing methods modify the training data before model fitting, often by reweighting, resampling, or generating transformed representations intended to reduce the influence of sensitive attributes (d'Alessandro et al., 2017; Yan et al., 2020). These methods are often straightforward to implement and can be attractive because they may not require access to sensitive labels at deployment time. However, they can be difficult to adapt in domains with strong structural or clinical constraints, such as healthcare, where aggressive data transformation or synthetic data generation may distort clinically meaningful relationships (Chen et al., 2021). Reweighing-based approaches are effective in many settings, but their performance may degrade when subgroup representation is highly imbalanced or when fairness-relevant structure is not easily corrected through sample-level weighting alone (Paviglianiti & Pasero, 2020; Yang et al., 2022).

In-processing methods intervene directly in the learning algorithm, typically by modifying the loss, adding fairness constraints, or introducing adversarial or regularization-based objectives (Gorrostieta et al., 2019; Liu & Vicente, 2022). These methods can provide stronger control over fairness objectives, but they often require access to the model internals and may be tightly coupled to a particular architecture or fairness definition, which can reduce transferability and increase implementation complexity (Yan et al., 2020; Gorrostieta et al., 2019; Pagano et al., 2022; Oneto et al., 2019). Post-processing methods instead adjust model outputs after training to satisfy fairness criteria (Oneto et al., 2019; Yan et al., 2020). They are often more modular and model-agnostic, but they may offer less flexibility when the goal is to jointly navigate predictive performance and fairness during model development.

A recurring challenge across these families is that fairness interventions often induce nontrivial trade-offs with predictive performance, and these trade-offs can vary substantially across datasets, tasks, and fairness definitions (Caton & Haas, 2024). Recent work has increasingly framed this tension in terms of explicit fairness-utility or fairness-accuracy frontiers, emphasizing that model development may require selecting among multiple candidate operating points rather than optimizing a single scalar objective (Tang et al., 2023). This perspective motivates our focus on fairness-aware checkpoint screening: instead of assuming that the final trained model is necessarily the best fairness-performance compromise, we treat model selection itself as part of the fairness-aware learning pipeline.

Unlike purely pre-processing methods, our framework uses a multi-task neural architecture and is therefore best understood primarily as an in-processing approach. Its distinctive feature is that final model choice is deferred to a post-training checkpoint-screening stage based on validation-set fairness and predictive performance. We therefore view the method as a hybrid in-processing workflow with post-training model selection, rather than as a fully model-agnostic intervention.

2.2 Multi-task Learning in Fairness-Aware Prediction

MTL trains a model to solve multiple related prediction tasks simultaneously by learning a shared representation while preserving task-specific outputs (Zhang & Yang, 2018). In fairness-related settings, this shared-representation structure has been explored as a way to expose the model to information about both the primary prediction task and protected-attribute structure during training, potentially affecting how features are encoded and how subgroup disparities emerge.

Prior work has shown that MTL can improve both predictive performance and fairness in certain domains. For example, Li et al. (Li et al., 2023) reported fairness improvements in clinical prediction settings by dynamically adjusting task gradients across subgroup-defined tasks. Gao et al. (Gao et al., 2022) introduced negative multi-task learning (NMTL), in which auxiliary tasks are selected to counteract biased correlations and reduce the propagation of undesirable bias signals. Oneto et al. (Oneto et al., 2019) further demonstrated that fairness can be incorporated as an auxiliary learning objective, allowing shared representations to reduce disparity with limited loss in predictive accuracy.

More recent work has continued to revisit fairness in multi-task settings, including analyses of how shared representations and task interactions affect fairness-utility trade-offs in modern MTL pipelines. Qin et al. (Qin et al., 2025) study fairness in MTL through performance variance across tasks and propose a gradient aggregation strategy that explicitly seeks to reduce imbalance during multi-task optimization. This line of work reinforces that fairness-related behavior in MTL can arise not only from auxiliary-task design, but also from how task interactions are managed during training.

These studies motivate the use of MTL as a flexible architectural foundation for fairness-aware prediction, but they do not imply that MTL alone guarantees improved fairness. In our framework, MTL is used primarily to jointly model the primary target and the protected attribute within a shared neural representation. This shared structure creates a family of candidate checkpoints across training whose behavior can then be evaluated under fairness and performance criteria.

Our contribution extends prior MTL-based fairness work in two ways. First, rather than treating fairness solely as a property of the final trained model, we treat fairness-aware model choice as a post-training selection problem over saved checkpoints. Second, we pair the multi-task architecture with MC dropout-based stochastic evaluation and Pareto-style screening on the validation set. In the present study, this screening is instantiated using predictive performance together with a primary fairness objective based on DIR, while additional disparity metrics are reported as secondary diagnostics. This design allows us to examine whether different checkpoints within the same multi-task training trajectory exhibit meaningfully different fairness-performance trade-offs, and whether validation-based screening can recover favorable operating points across multiple application domains.

2.3 Model Uncertainty in Fairness-Aware Model Selection

Uncertainty estimation has become an important component of trustworthy ML, particularly in high-stakes settings where models must support decisions under distributional shift, sparse subgroup coverage, or ambiguous labels (Gal & Ghahramani, 2016; Hüllermeier & Waegeman, 2021). Broadly, uncertainty can be decomposed into epistemic uncertainty, which reflects uncertainty in the model parameters or structure, and aleatoric uncertainty, which reflects irreducible noise in the data-generating process (Hüllermeier & Waegeman, 2021). In neural networks, MC dropout provides a practical approximation to Bayesian inference by enabling stochastic forward passes at test time, thereby producing a distribution over predictions rather than a single deterministic output (Gal & Ghahramani, 2016).

Prior work has used uncertainty estimates for calibration, abstention, active learning, and risk-sensitive decision-making, but their role in fairness-aware learning remains comparatively underexplored. Recent work has begun to investigate this connection more directly. Heuss et al. (Heuss et al., 2023) propose an uncertainty-aware post hoc bias mitigation method in ranking, showing that predictive uncertainty can be used to navigate utility-fairness trade-offs without retraining the underlying ranking model. More recently, Rosenblatt and Witter (Rosenblatt & Witter, 2024) introduced *FairlyUncertain*, a benchmark for evaluating uncertainty in algorithmic fairness, highlighting that uncertainty estimates can improve calibration and consistency in some settings but do not automatically resolve downstream group disparities. These findings are especially relevant because they caution against assuming a simple or universal relationship between predictive uncertainty and fairness.

Our framework builds on this emerging perspective but uses uncertainty differently. We do not treat uncertainty as a standalone fairness intervention, nor do we claim that uncertainty alone guarantees fairer predictions. Instead, MC dropout is used as a stochastic evaluation mechanism during checkpoint screening: saved checkpoints from the multi-task training trajectory are evaluated on the validation set using

repeated stochastic forward passes, from which we summarize predictive variability and compute fairness and performance metrics. These validation-set measurements are then used to compare checkpoints under a multi-objective fairness-performance criterion.

This framing positions uncertainty as an auxiliary signal in a broader fairness-aware model selection workflow. In our setting, uncertainty helps characterize checkpoint behavior under stochastic prediction, but the final checkpoint is selected based on validation-set fairness and predictive performance rather than on uncertainty alone. This distinction is important: the empirical relationship between checkpoint-level stochastic behavior and fairness may be useful in practice, but it should not be interpreted as a theoretical guarantee that higher or lower uncertainty directly causes improved fairness outcomes.

3 Fairness Terminology

In this section, we introduce and define some of the concepts and terminologies generally used in algorithmic fairness research that are relevant to this paper.

- **Protected (sensitive) label**: An attribute that partitions a population into groups whose outcomes should have parity (such as race, gender, income, etc.).

- **Privileged class**: A protected label value indicating a group that is at an advantage. Generally represented by the value 1.

- **Unprivileged class**: A protected label value indicating a group that is at a disadvantage. Generally represented by the value 0.

- **Disparate impact ratio (DIR)**: This fairness metric compares the rate of positive outcomes in the unprivileged class to that in the privileged class, as shown in Equation 1. A value of 1 corresponds to parity, while values below or above 1 indicate group disparities in opposite directions. In this work, we report both the raw DIR values and their deviation from parity, $DIR = 1$ (1 - DIR) and use DIR as the primary fairness objective for checkpoint screening.

$$\text{DIR} = \frac{Pr(Y = 1|D = unprivileged)}{Pr(Y = 1|D = privileged)} \tag{1}$$

  where Y represents the target label and D the class of the protected demographic label.

- **Equalized odds**: This fairness metric enforces that the model correctly identifies the positive outcome at equal rates across groups, and miss-classify the positive outcome at equal rates across groups (creating the same proportion of True Positives and False Positives across groups). In this work, we represent this metric using 2 values; Difference in False Negative (FN) scores, and Difference in False Positive (FP) scores.

In this study, we use the DIR as the primary fairness objective for checkpoint screening and final model selection. We selected this metric because it is straightforward to interpret across all three application domains and permits a consistent trade-off analysis with predictive performance.

We also report error-based disparities, including false-negative and false-positive differences where relevant, as secondary diagnostics rather than primary optimization targets. We emphasize that the current framework is not restricted in principle to demographic parity: alternative or multiple fairness objectives (e.g., equal opportunity or equalized odds) could be substituted into the checkpoint-selection stage. We leave systematic evaluation of multi-metric or alternative-fairness selection to future work.

4 Proposed Fairness-Aware Checkpoint Screening Framework

4.1 Overview

Our framework combines MTL, MC dropout, and validation-based checkpoint screening to support fairness-aware model selection in neural prediction pipelines. Rather than assuming that the final training epoch provides the best fairness-performance trade-off, we retain checkpoints across training and evaluate them on a held-out validation set using both predictive performance and a group-fairness objective based on DIR.

The central motivation is practical rather than mechanistic. During multitask training, different checkpoints can exhibit meaningfully different trade-offs between predictive performance and group fairness, even when they arise from the same training run. We therefore treat checkpoint selection itself as part of the fairness-aware learning pipeline. In this setting, the goal is not to identify a universally optimal model, but to recover candidate checkpoints whose validation-set behavior reflects favorable operating points under competing objectives.

MC dropout is used to obtain stochastic predictions and summarize checkpoint-level predictive variability during this screening process. In our framework, this predictive variability is treated as an auxiliary diagnostic signal rather than as a direct optimization target or a guaranteed mechanism for fairness improvement. The primary model-selection mechanism is Pareto-based checkpoint screening over validation-set fairness and predictive performance, followed by a pre-specified tie-breaking rule for final checkpoint choice.

We therefore frame the method as a practical fairness-aware checkpoint-selection procedure: MTL provides a shared representation over target and protected-label tasks, MC dropout provides stochastic evaluation of saved checkpoints, and Pareto-based screening makes fairness-performance trade-offs explicit across candidate models. Figure 1 illustrates the overall workflow; the following subsections describe each component in detail.

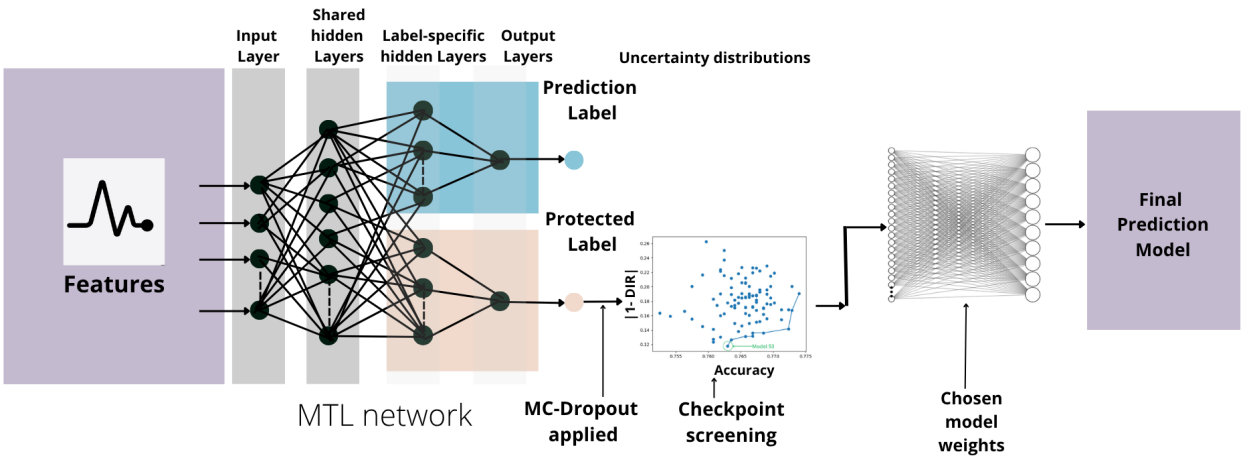


Figure 1: Proposed Method Structure
(MTL: multi-task learning, MC monte carlo dropout, DIR: disparate impact ratio)

4.2 Multi-Task Learning for Shared Representation Learning

We first train a multi-task learning (MTL) model that jointly predicts the primary target label and a selected protected attribute using a shared neural representation. Let x denote the input features, y the target label, and a the protected attribute. The model consists of a shared feature extractor (or backbone) $f_\theta(\cdot)$, which maps the input into a latent representation

$$h = f_\theta(x),$$

where h denotes the shared hidden representation learned from the input. This shared representation is then passed to two task-specific prediction heads: a target-label head $g_{\phi_y}(\cdot)$ and a protected-attribute head

$g_{\phi_a}(\cdot)$. These heads produce

$$\hat{y} = g_{\phi_y}(h), \qquad \hat{a} = g_{\phi_a}(h),$$

where $\hat{y}$ is the predicted target label and $\hat{a}$ is the predicted protected attribute.

Training is performed on the training split using a weighted multitask objective:

$$\mathcal{L}_{\text{MTL}} = \lambda_y \mathcal{L}_y(y, \hat{y}) + \lambda_a \mathcal{L}_a(a, \hat{a}),$$

where $\mathcal{L}_y$ and $\mathcal{L}_a$ are task-specific loss terms and λ_y, λ_a control their relative contributions.

We use this multitask formulation as a shared-representation mechanism rather than as a standalone fairness guarantee. The key role of the MTL stage is to produce a training trajectory of saved checkpoints whose predictive and fairness behavior can differ across epochs. These checkpoints are subsequently evaluated and screened on the validation set using the procedure described in Sections 4.3 and 4.4.

4.3 Monte Carlo Dropout for Stochastic Checkpoint Evaluation

After training the multi-task model, we evaluate saved checkpoints using MC dropout to obtain stochastic predictions under repeated forward passes. The purpose of this stage is to characterize checkpoint behavior under stochastic inference before fairness-performance screening on the validation set.

MC dropout provides a practical approximation to Bayesian model averaging in neural networks by retaining dropout at inference time and performing multiple stochastic forward passes through the same checkpoint (Gal & Ghahramani, 2016). For a given saved checkpoint and input example x, we perform T stochastic forward passes, yielding predictions

$$\{\hat{y}^{(1)}(x), \hat{y}^{(2)}(x), \ldots, \hat{y}^{(T)}(x)\}.$$

From these samples, we compute the mean predictive output

$$\bar{y}(x) = \frac{1}{T} \sum_{t=1}^{T} \hat{y}^{(t)}(x),$$

which is used as the checkpoint's aggregated prediction for downstream evaluation.

To summarize checkpoint-level predictive variability, we also compute the empirical variance across stochastic passes:

$$u(x) = \frac{1}{T} \sum_{t=1}^{T} \left(\hat{y}^{(t)}(x) - \bar{y}(x) \right)^2.$$

This quantity captures how sensitive the checkpoint's prediction is to stochastic dropout perturbations for a given input. In our framework, such variability is treated as a descriptive diagnostic of checkpoint behavior rather than as a standalone fairness objective.

For each saved checkpoint, MC dropout inference is applied on the validation split. Using the aggregated predictions, we compute target-task predictive performance and fairness metrics, including the DIR. The predictive variability summaries may help characterize how checkpoints differ under stochastic evaluation, but final checkpoint choice is not based on uncertainty alone. Instead, these validation-set measurements feed into the Pareto-based checkpoint-screening procedure described in Section 4.4, where checkpoints are compared using predictive performance together with fairness.

4.4 Pareto-based checkpoint screening

After applying MC-dropout inference on the validation split to estimate predictive uncertainty and obtain target-task predictions, we compute checkpoint-level predictive performance and fairness metrics, including the DIR. We then perform Pareto-based checkpoint screening over validation-set predictive performance and fairness, retaining checkpoints that are not dominated under the objective of maximizing predictive

performance and minimizing $|1 - DIR|$. Among Pareto-optimal checkpoints, we select the final checkpoint by prioritizing the smallest absolute deviation from parity, $|1 - DIR|$, with predictive performance used as a secondary criterion when multiple checkpoints have comparable fairness. The test split is not used during this checkpoint screening or selection; it is reserved for final evaluation after the checkpoint has been chosen.

4.5 End-to-End Training, Checkpoint Screening, and Evaluation Protocol

For clarity, we summarize the full experimental pipeline as follows:

- Train multi-task model on the training split: We train the MTL model on the training split, jointly predicting the primary target and the selected protected attribute.
- Save checkpoints across epochs: During training, we save model checkpoints at regular intervals (e.g., each epoch), producing a set of candidate checkpoints.
- Evaluate checkpoints on the validation split. For each saved checkpoint, we perform MC-dropout inference on the validation split to estimate predictive variability and compute target-task predictive performance together with the fairness metric (DIR).
- Construct the checkpoint Pareto set and select a final checkpoint: Using the validation-set metrics, we perform Pareto-based checkpoint screening across candidate checkpoints and select the final checkpoint according to our predefined tie-breaking rule (e.g., smallest $|1 - DIR|$, with performance used to break ties).
- Evaluate once on the test split: The held-out test split is used after checkpoint selection is fixed. No test-set information is used in checkpoint screening or model selection.

5 Experiments

5.1 Datasets

We evaluate the effectiveness of our proposed method using three datasets, (1) ADULT (individual income and demographic information) (Kohavi et al., 1996), (2) MIMIC-III (data from patients in critical care units) (Johnson et al., 2016), and (3) SNAPSHOT datasets (multi-modal physiological, behavioral, and survey data collected from 350 college students) (Sano et al., 2015). Table 1 describes the input and binary target variables for each of these datasets used in our experiments. More information on the details of these datasets and data processing can be found in Appendix A.

For the ADULT dataset, we regard age, sex(gender) and race as the protected labels, and for MIMIC-III, age, gender, insurance type, and marital status. For the SNAPSHOT dataset, we assign gender, race, and ethnicity as the protected labels in our experiments. We also use personality types, including openness, conscientiousness, extraversion, agreeableness, and neuroticism, as protected labels in our experiments for SNAPSHOT. Table 2 details how we binary-encode each protected label and assign them as privileged and unprivileged classes to analyze them using our chosen fairness metrics. These thresholds were chosen to support consistent binary fairness analyses across datasets and should be interpreted as practical experimental binarizations rather than uniquely normative cut points.

5.2 Baseline Models and Fairness Analyses

The first part of our experiments involves developing baseline models that make the target predictions for all datasets (income for ADULT, IHM for MIMIC-III, and 4 happiness and calmness labels for SNAPSHOT). Next, we analyze our prediction results for fairness by calculating the DIR and equalized odds (FP and FN rates) for predictions against the protected labels for all datasets. We conduct these analyses to determine whether and where the models were introducing bias and ensure that we only apply our method to aspects of the data that are actually biased. Details of the models can be found in Appendix A.

Table 1: Input and Target Variables for ADULT, MIMIC-III, and SNAPSHOT Datasets

Dataset	Input Variables	Target Variables
ADULT	Age, workclass, education, Marital-status, occupation, relationship, race, sex	Salary
MIMIC-III	Vital signs, medications, laboratory measurements, fluid balance, procedure codes, hospital length of stay, survival data, laboratory test results, diagnostic codes, patient demographics, billing information, observations and notes, etc.	In-hospital mortality (IHM)
SNAPSHOT	Physiological features, mobile phone usage, weather data	Morning happiness Evening happiness Morning Calmness Evening Calmness

Table 2: Binarization of Protected labels for ADULT, MIMIC-III, and SNAPSHOT Datasets

Dataset	**Label**	**Privileged Class (1)**	**Unprivileged Class (0)**
ADULT	Age	Less than or equal to 40 years	Greater than 40 years
	Race	White, Asian-Pac-Islander	Black, Amer-Indian-Eskimo, other
	Sex	Male	Female
MIMIC-III	Age	Greater than 60 years	Less than or equal to 60 years
	Gender	Male	Female
	Insurance	Private, other	Medicare, Medicaid
	Marital Status	Married, Life partner	Single, Widowed, Divorced, Separated
SNAP-SHOT	Gender	Male	Female
	Race	White	Non-white
	Ethnicity	Non-Hispanic/latino	Hispanic/latino
	Openness	$<= 50$	> 50
	Conscientiousness	> 50	$<= 50$
	Extraversion	$<= 50$	> 50
	Agreeableness	$<= 50$	> 50
	Neuroticism	$<= 50$	> 50

After performing fairness analysis on each baseline model, we select the protected labels by which our models showed the most bias and test our methods on all datasets. For all fairness analyses in our experiments, we utilize a fairness toolkit developed by IBM - AI Fairness 360 (Bellamy et al., 2019).

5.3 Checkpoint Evaluation and Selection Protocol

For each dataset, we partition the data into training, validation, and test splits. The multi-task neural model is trained using the training split, and model checkpoints are saved across epochs to form a candidate checkpoint set.

For each saved checkpoint, we perform multiple stochastic forward passes with MC dropout on the validation split to estimate predictive variability and obtain checkpoint-level uncertainty summaries. Using the same validation split, we compute predictive performance metrics for the target task and a group-fairness metric based on the DIR for the protected attribute under study.

We then apply the Pareto-based checkpoint screening procedure described in Section 4.4 to the validation-set fairness and performance metrics across checkpoints, producing a Pareto set of candidate checkpoints for each dataset. Final checkpoint selection is performed using validation-set metrics only, according to a pre-specified rule that prioritizes fairness while preserving competitive predictive performance. Specifically, among Pareto-optimal checkpoints, we select the checkpoint with the smallest $|1 - DIR|$, breaking ties by choosing the checkpoint with the highest validation-set predictive performance. The test split is used only once, after checkpoint selection is complete, to evaluate the selected checkpoint and compare it against baseline methods.

In our figures, we visualize checkpoint-level trade-offs to illustrate the fairness–performance landscape; however, all reported "selected" checkpoints correspond to models chosen on the validation split and shown on the test set only for final comparison.

5.4 Comparison Baselines for Fairness Intervention

We compare the proposed checkpoint-screening framework against three fairness baselines drawn from prior literature. These baselines were selected to represent different stages of intervention in the learning pipeline, including pre-processing, in-processing, and feature-level mitigation. This provides a broader empirical context for evaluating whether validation-based checkpoint screening can recover competitive fairness-performance trade-offs relative to established fairness-aware learning strategies. We prioritize widely-used and interpretable baselines spanning different intervention stages rather than exhaustively comparing to all recent deep fairness methods.

(1) Reweighting (Kamiran & Calders, 2012) (ADULT, MIMIC-III, and SNAPSHOT): a standard pre-processing method that adjusts instance weights to reduce disparities between protected and unprotected groups before model training. We include reweighting because it is one of the most widely used fairness baselines and serves as a simple, broadly applicable reference for group-fairness mitigation across domains.

(2) Adversarial Reweighted Learning (ARL) (Lahoti et al., 2020) (ADULT and MIMIC-III): an in-processing method that adaptively learns sample weights during training in order to emphasize examples associated with higher unfairness. We include ARL as a representative optimization-based fairness baseline that reshapes the effective training distribution during learning.

(3) FairRF (Zhao et al., 2022) (ADULT and MIMIC-III): a fairness-aware feature-reweighting method that adjusts the influence of features associated with protected attributes. We include FairRF as a representative feature-level mitigation baseline that operates directly on feature contributions rather than on checkpoint selection.

Taken together, these baselines allow us to compare the proposed framework against established fairness interventions that act at different points in the modeling pipeline. By comparison, our method treats fairness-aware model selection as the primary intervention: candidate checkpoints from a multitask neural training trajectory are evaluated on the validation split, screened using fairness and predictive-performance

criteria, and then compared against these baselines on the held-out test split. Additional implementation details for all baseline methods are provided in Appendix A.

For all baselines, hyperparameters and implementation choices follow prior work or standard library defaults as described in Appendix A, and all final comparisons are reported on the same held-out test split used for the selected checkpoint from our framework.

5.5 Saliency Maps

We use saliency maps as a qualitative post hoc diagnostic to inspect whether selected checkpoints appear to shift feature emphasis relative to baseline models. These visualizations are intended for descriptive analysis and should not be interpreted as causal evidence of the mechanism driving fairness changes. We utilize a saliency map technique (Simonyan et al., 2013) to visualize the importance of model weights on our input time axis and features. According to the predicted class c, for example, the decision-making process of the model can be represented as $S_c(I) = w_c^T I + b_c$, where $I, S_c(I), w, b$ are the model input, output, weights, and bias, respectively. The purpose of this method is to calculate the model weights on the input layer by gradients, for example, $w = \frac{\partial S_c}{\partial I}|_{I_0}$. The calculated w represents the model saliency associated with the input layer. In this study, we fetch the saliency maps for all the test samples and calculate the average saliency map to develop the general intuition of important features and time steps.

We calculate the feature importance for our experiments according to the formulas in this section and plot them in the figures presented in Section 6.

6 Results

6.1 ADULT Dataset

6.1.1 Fairness Analyses

Initial fairness analysis indicates substantial disparity in the ADULT dataset across all three protected attributes considered. Before model training, the DIR score is above 2 for age, approximately 0.36 for sex, and approximately 0.47 for race in both the training and test data. After training the baseline model, these disparities persist and, in some cases, become more pronounced: the DIR increases further for age, falls below 0.1 for sex, and remains near 0.5 for race. These results indicate that the baseline model exhibits meaningful group disparity for all three protected attributes and therefore provides a suitable setting for evaluating the proposed checkpoint-screening framework.

6.1.2 Methods Implementation

Because the baseline model exhibited substantial disparities across age, sex, and race, we applied the proposed checkpoint-screening framework separately for each protected attribute. For each experiment, Pareto fronts were constructed using validation-set fairness and performance metrics across saved checkpoints. The highlighted checkpoints in Figure 2 correspond to the validation-selected models and indicate where the final chosen checkpoint lies within the fairness-performance trade-off landscape. In these experiments, the selected checkpoints tended to favor improved fairness because predictive performance differences among Pareto-optimal candidates were relatively small. All comparative results reported below are measured on the held-out test set after checkpoint selection was fixed.

Figure 3 compares the final selected checkpoint from our method against reweighing, ARL, and FairRF relative to the baseline model. In this figure, the x-axis represents the percentage improvement in DIR relative to the baseline with respect to the ideal value of 1, and the y-axis shows the percentage change in accuracy relative to the baseline model.

Reweighing improves fairness for age and sex while producing comparatively small changes in accuracy. For race, however, the fairness gain is minimal. FairRF improves fairness more strongly for age and sex than the other comparison methods, but does so with a larger performance cost. For race, FairRF reduces both fairness and accuracy relative to baseline.

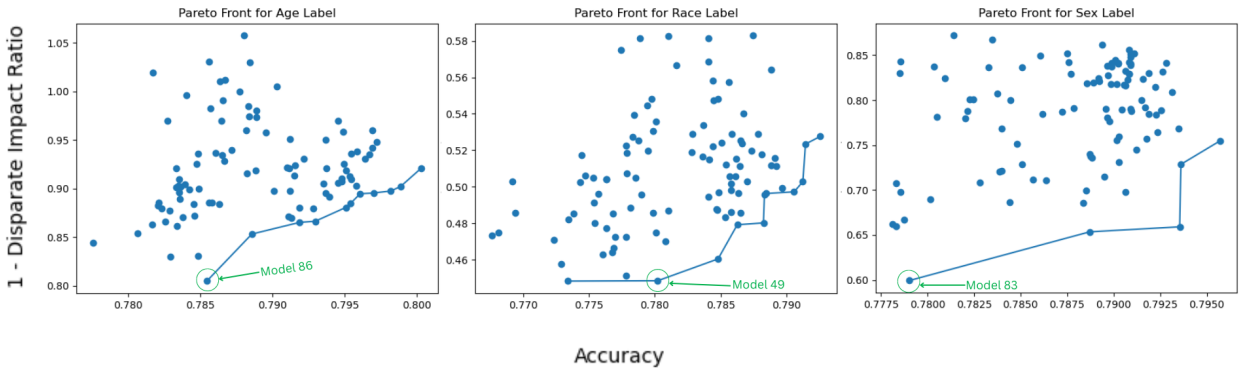


Figure 2: ADULT: Pareto Front Plot for Age, Race, and Sex Labels

The checkpoint selected by our proposed method yields larger DIR improvements than reweighing for all three protected attributes, with improvements of 128% for age, 32% for sex, and 4.2% for race. These gains come with a modest average decrease of 1.7% in accuracy across the three labels. ARL performs comparably to the proposed method for the age attribute, improving DIR by 114%, but it does not improve fairness for sex or race. Overall, these results suggest that the proposed framework does not uniformly dominate every baseline across all settings, but often identifies a favorable fairness-performance compromise when fairness improvement is prioritized without severe loss in predictive performance.

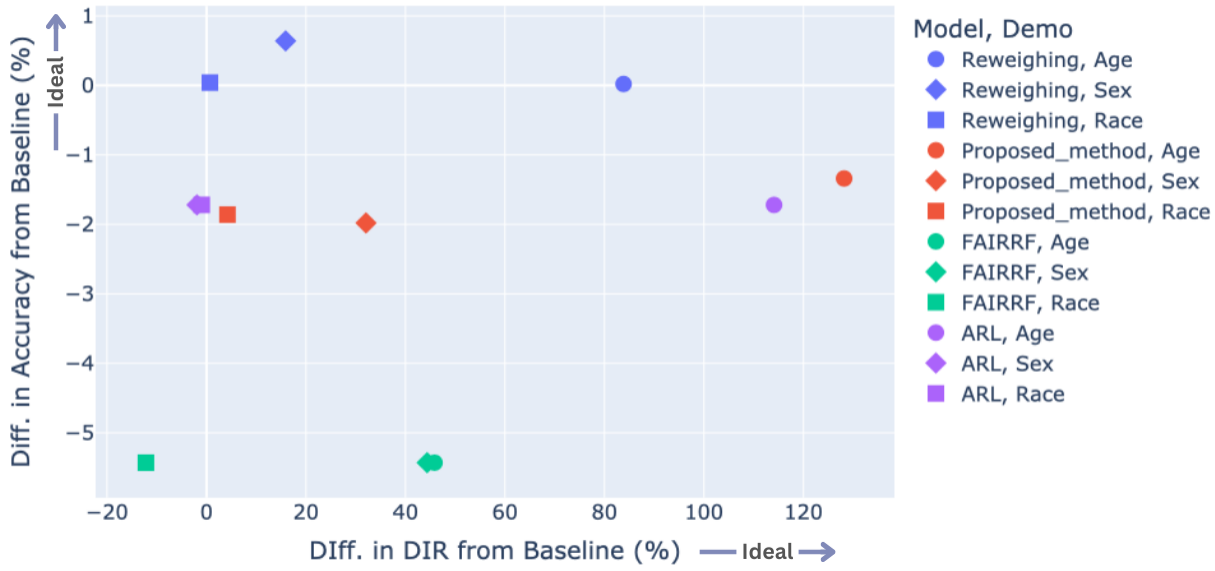
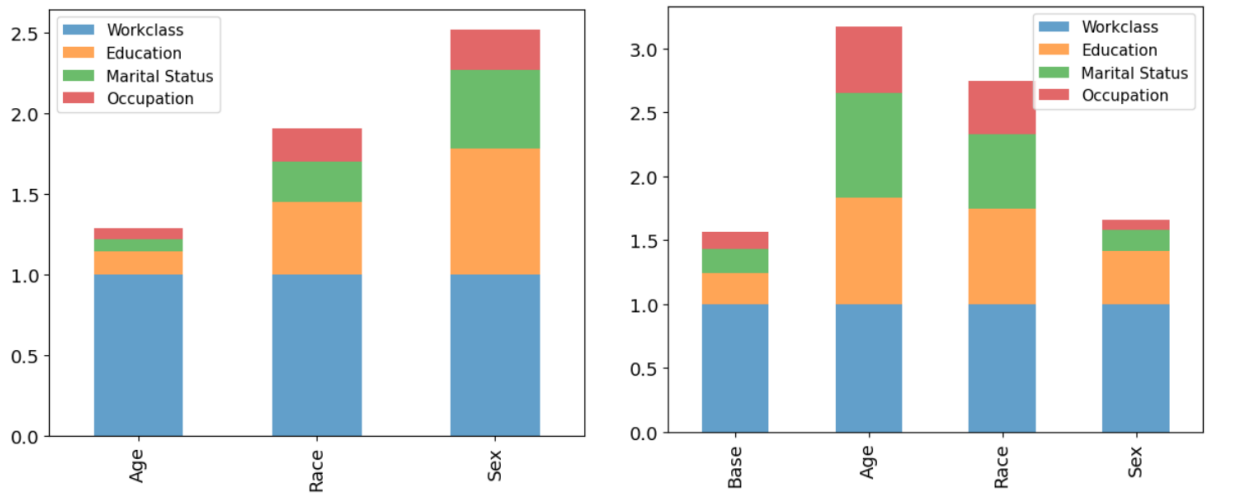


Figure 3: ADULT: DIR and Accuracy Scores of Models

6.1.3 Saliency-Based Qualitative Analysis

Figure 4 provides a qualitative comparison of feature-importance patterns between the baseline model and the validation-selected checkpoints for age, race, and sex. Because the ADULT data do not contain temporal structure, we use bar-chart summaries rather than saliency heatmaps for ease of visual comparison. These plots should be interpreted as descriptive diagnostics of feature sensitivity rather than as causal evidence of the mechanism underlying fairness changes.

Across all three experiments, *workclass* remains the most influential feature and *relationship* remains negligible. The more notable differences appear in *education*, *marital status*, and *occupation*, whose relative importance changes across the selected checkpoints depending on the protected attribute used in screening. For example, in the age-based experiment, the selected checkpoint shows a shift in feature sensitivity among these three variables relative to the baseline. Similar redistribution patterns are visible for the race- and sex-based experiments. We interpret these changes as qualitative evidence that the selected checkpoints are not trivially identical to the baseline model and may rely on somewhat different internal representations, rather than as proof of a specific debiasing mechanism.



(a) Feature Importance for Demographic Labels (b) Feature Importance of Models for Target Label

Figure 4: ADULT: Feature Importance Comparison Between Protected and Target Labels

6.2 MIMIC-III Dataset

6.2.1 Fairness Analyses

The baseline model for in-hospital mortality (IHM) prediction on MIMIC-III achieves Accuracy = 0.866, AUROC = 0.603, and AUPRC = 0.181. Fairness analysis shows that, among the four protected attributes considered, the baseline model exhibits the clearest disparity with respect to marital status, with a DIR score of 1.308. By contrast, age, gender, and insurance remain relatively close to parity in this setup, and the differences in false negative rates are also small across all protected labels (Table 3). We therefore focus our experiments on marital status.

Table 3: MIMIC: Fairness Scores for Baseline Model

Label	**DI Ratio Score**	**Diff in FN Scores**
Age	1.008	-0.006
Marital status	1.308	0.005
Gender	1.182	-5.645e^{-05}
Insurance	1.021	-0.004

6.2.2 Methods Implementation

Because the baseline model shows the most pronounced disparity for marital status, we apply the proposed checkpoint-screening framework using IHM as the target label and marital status as the protected attribute. Figure 5 shows the resulting Pareto front for DIR versus AUROC. For visualization purposes, we also use

the absolute deviation from parity, $|1 - \text{DIR}|$, when discussing the trade-off between fairness and predictive quality.

The models along the Pareto front exhibit relatively small variation in accuracy and more substantial variation in AUROC and DIR. Because AUROC is a central metric for this task and is emphasized in prior work (Harutyunyan et al., 2019), we prioritize AUROC when selecting among Pareto-optimal candidates. In Figure 5, models 23 and 40 both move DIR closer to parity, but model 40 provides the stronger overall improvement in AUROC, AUPRC, and accuracy. We therefore choose model 40 as the final checkpoint.

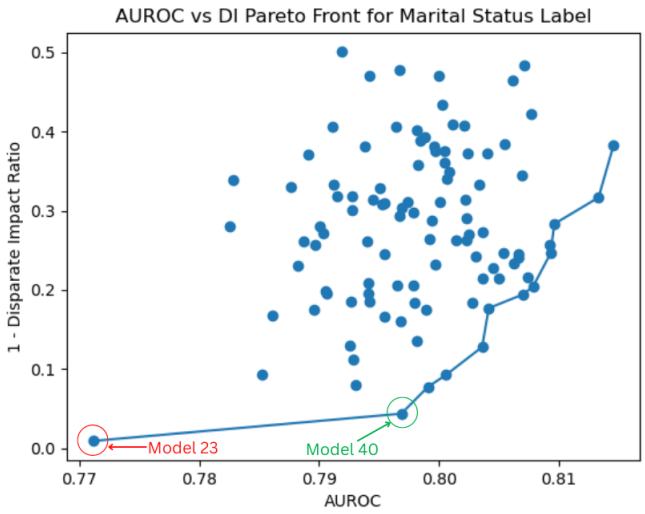


Figure 5: MIMIC-III: Pareto Front Plot for AUROC vs DI Ratio Score

Table 4 compares the selected checkpoint against the baseline and alternative fairness baselines. Reweighing improves fairness and predictive performance relative to baseline, but the selected checkpoint from our method yields a DIR closer to parity together with the strongest AUROC and AUPRC among the compared methods. ARL and FairRF achieve DIR values closer to 1 than the baseline as well, but do so with substantially larger reductions in predictive performance. In this setting, the proposed framework appears to recover a favorable fairness-performance trade-off: it improves fairness relative to the baseline while preserving or improving predictive quality more effectively than the alternative baselines considered here.

Table 4: MIMIC: Performance and Fairness of Models from Proposed Method, Reweighing Method, and Baseline Model

Method	Acc	AUROC	AUPRC	DIR	Diff in FN
Base Model	0.886	0.603	0.181	1.308	0.005
Reweighing	0.883	0.777	0.332	1.188	-0.05
ARL	0.719	0.528	0.136	0.954	0.064
FairRF	0.669	0.520	0.118	0.982	-0.035
Proposed Method	0.883	0.797	0.367	1.043	0.005

6.2.3 Saliency-Based Qualitative Analysis

Additional saliency maps for MIMIC-III are provided in Appendix A. These maps offer a qualitative view of which clinical features are most influential for the baseline IHM model and how those patterns differ in the selected checkpoint. As throughout this paper, we interpret these visualizations as descriptive diagnostics rather than causal explanations of fairness changes.

For the baseline model, highly weighted features include Glasgow Coma Scale (GCS) motor-response variables, Fraction inspired oxygen (FiO_2), and pH. These variables are clinically plausible for mortality predic-

tion. Prior work has linked low GCS motor-response scores to greater mortality risk (Jain & Iverson, 2018; Yumoto et al., 2019; Sadaka et al., 2012; Settervall et al., 2011), higher FiO_2 to more severe respiratory support and worse outcomes (Fuentes & Chowdhury, 2020; de Jonge et al., 2008), and abnormal pH levels to increased mortality in critically ill patients (Arias-Oliveras, 2016; Park et al., 2020). The saliency analysis therefore suggests that the baseline model emphasizes medically relevant variables, while the selected checkpoint may redistribute feature sensitivity somewhat differently. We treat these differences as qualitative evidence of a different representational profile rather than as proof of a specific fairness mechanism.

6.3 SNAPSHOT Dataset

6.3.1 Fairness Analyses

We first perform 32 fairness analyses on the original SNAPSHOT data across the protected attributes and target labels considered in this study. In contrast to ADULT and MIMIC-III, the data itself shows relatively limited disparity overall. Figure 6 indicates that the most pronounced imbalance occurs for the combination of *Race* as the protected label and *evening-sad-happy* as the target label (marked by the red circle), while most other settings remain closer to parity. The baseline models across the four SNAPSHOT targets achieve F1 scores between approximately 0.7 and 0.8.

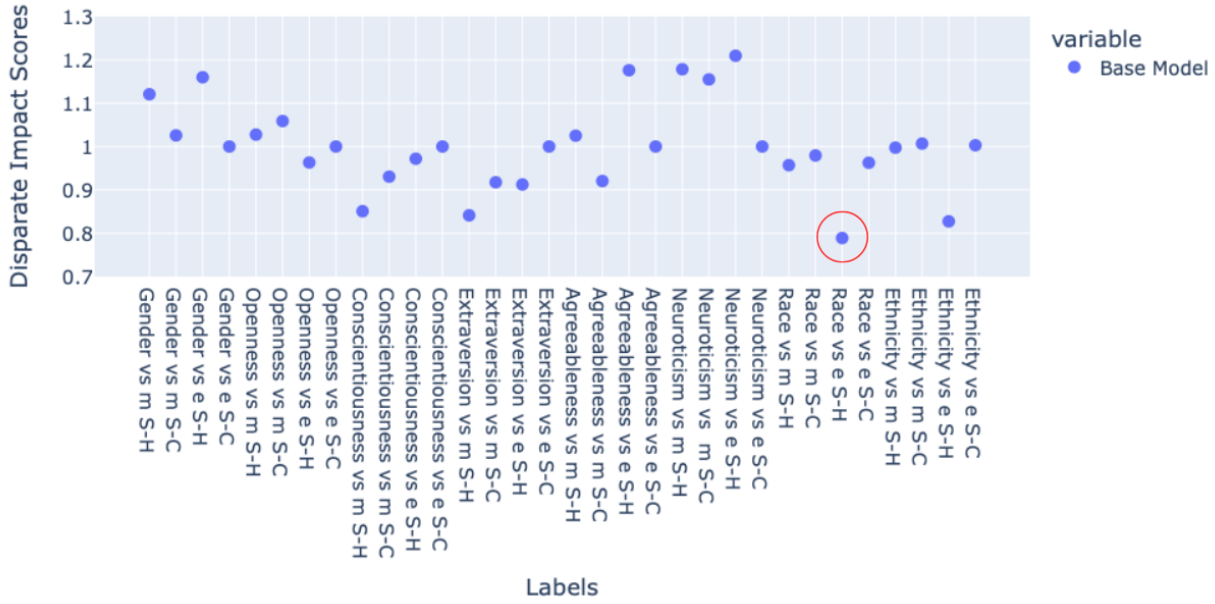


Figure 6: SNAPSHOT: Scatter Plot of DIR Scores from Baseline Model

6.3.2 Methods Implementation

We apply the proposed method to the most imbalanced original SNAPSHOT setting: Race as the protected label and evening-sad-happy as the target label. The resulting checkpoint ensemble yields accuracy scores in the range of 0.76 to 0.77 and DIR scores between 0.80 and 0.88. The selected checkpoint, shown in Figure 7, achieves a DIR of 0.882, improving over the baseline DIR of 0.78 while preserving approximately the same F1 score (about 0.76). Reweighing produces a perfect DIR of 1 in this setting, but at a large predictive cost, reducing F1 to 0.39. This contrast illustrates the type of trade-off the proposed framework is designed to navigate: it does not necessarily achieve the most extreme fairness correction, but it can recover a checkpoint with a more favorable balance between fairness and predictive performance.

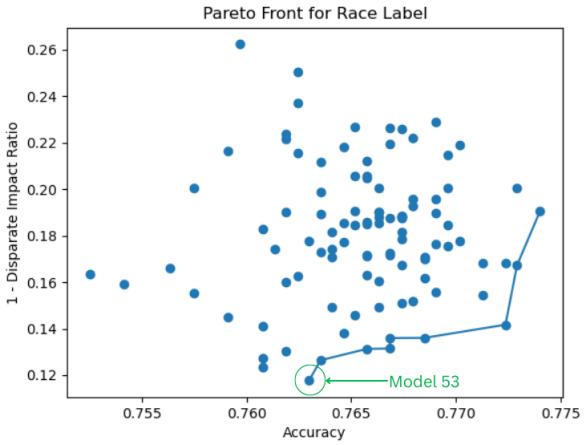


Figure 7: SNAPSHOT: Pareto Front Plot for Race vs Evening-sad-happy Label

6.3.3 Data Manipulation on SNAPSHOT

To probe the method under a stronger imbalance, we construct an altered SNAPSHOT setting by introducing a larger gender imbalance (see Appendix A). After retraining the baseline model and reanalyzing fairness across all target labels, the most substantial disparity again appears for evening-sad-happy, now with respect to gender. Table 5 summarizes the resulting F1 and fairness metrics for one representative altered-data setting. The evening-sad-happy label exhibits the clearest deviation from parity, while the other labels remain comparatively less biased.

Table 5: SNAPSHOT: F1 and Fairness Scores for Altered Data Experiments

Label	F1 Score	DI Ratio Score	Diff in FN	Diff in FP
Morning-Sad-Happy	0.749	1.153	-0.041	0.0979
Morning-Stressed-Calm	0.803	1.045	-0.009	0.073
Evening-Sad-Happy	0.798	1.291	-0.123	0.088
Evening-Stressed-Calm	0.788	1.0536	-0.0281	0.048

We then apply the proposed checkpoint-screening framework to evening-sad-happy with gender as the protected attribute. Figure 8 shows the resulting Pareto front. Across the Pareto-optimal models, the average F1 score is 0.74, corresponding to a 5.8% decrease relative to the baseline evening-sad-happy model, while the average DIR improves from 1.291 to 1.034. Two candidate checkpoints, models 89 and 43, lie particularly close to parity. Model 89 attains a DIR of 1.001 with an F1 score of 0.722, while model 43 attains a DIR of 0.995 with slightly better performance (0.733). These results suggest that, under stronger imposed imbalance, the framework can still recover checkpoints that move substantially closer to parity, albeit with a moderate reduction in predictive performance.

6.3.4 Saliency-Based Qualitative Analysis

Figure 9 shows a subset of the SNAPSHOT saliency maps (additional maps are provided in Appendix A). Darker red values indicate higher feature sensitivity across the seven time steps. The left panels correspond to the baseline evening-sad-happy model and the right panels to the selected checkpoint from the proposed method.

In the baseline model, some of the most prominent features include `screen_0H-3H_mean_duration`, `screen_10H-17H_stdev_duration`, `Mob_5_minutes_distance_median`, and `Mob_number_of_ROIs_visited`. Relative to the baseline, the selected checkpoint appears to redistribute feature sensitivity more broadly across the available features and time steps rather than placing as much emphasis on a smaller number of highly weighted variables. As with the other saliency analyses, we interpret

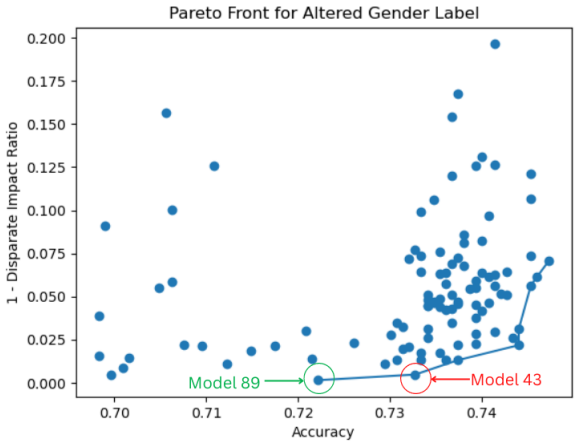


Figure 8: SNAPSHOT: Pareto Front Plot for Gender vs Evening-sad-happy Label with Altered Data

this pattern as qualitative evidence of a different feature-sensitivity profile rather than as direct evidence of a specific debiasing mechanism.

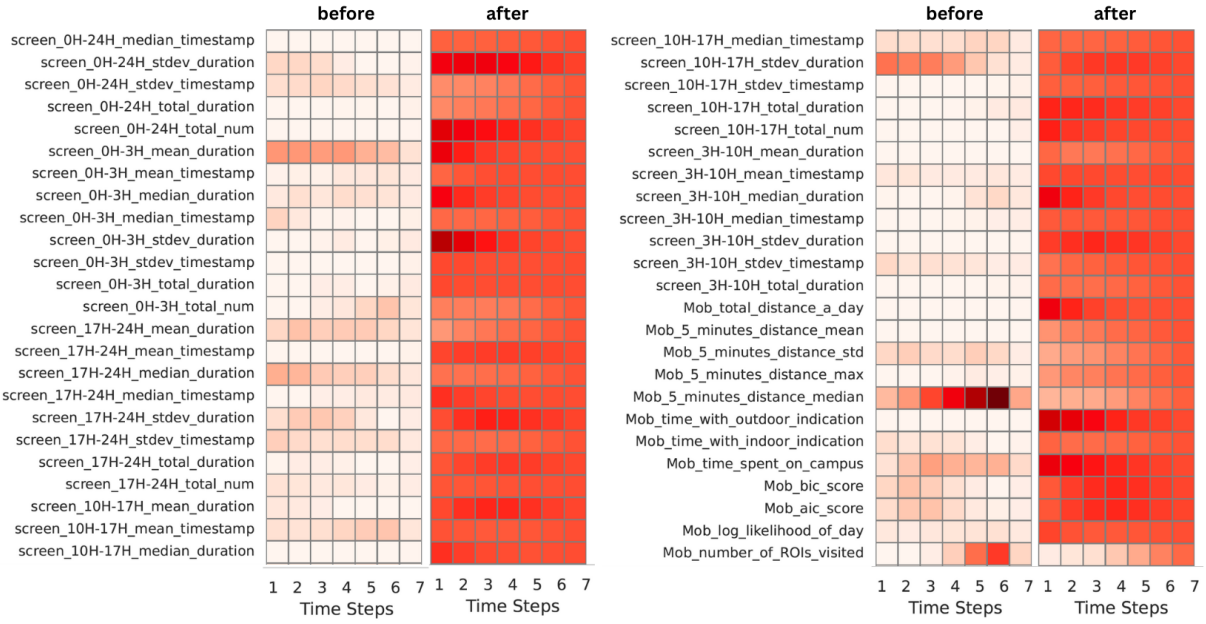


Figure 9: SNAPSHOT: Saliency Maps from Baseline Model and Model from Proposed Method.

7 Discussion

7.1 Fairness and Performance

Across the ADULT, MIMIC-III, and SNAPSHOT datasets, our results suggest that the proposed framework can serve as a practical method for identifying neural checkpoints with improved fairness-performance trade-offs relative to baseline training runs. Importantly, we show that the empirical benefit appears to come from treating fairness-aware model selection as part of the training pipeline: by screening saved checkpoints on a

validation set using both predictive performance and a fairness criterion, the method can surface candidate models that would likely be missed by standard final-epoch selection.

On the ADULT dataset, the baseline models exhibit substantial disparities across age, sex, and race, consistent with prior literature on demographic bias in income prediction. In this setting, the validation-selected checkpoints generally move the DIR closer to parity across the protected attributes studied, while preserving broadly competitive predictive performance relative to the baseline neural models. The strongest improvements are observed for fairness, whereas changes in predictive performance are typically more modest. Compared with reweighting, ARL, and FairRF, our method does not uniformly dominate across all protected attributes or metrics, but it often yields a favorable compromise when fairness improvement is prioritized without incurring severe degradation in predictive quality. This is consistent with the intended use of the framework as a screening-based trade-off mechanism rather than a universally superior fairness intervention.

For MIMIC-III, the baseline in-hospital mortality model shows measurable disparity with respect to marital status in our setup. The proposed checkpoint-screening procedure again identifies checkpoints that move DIR substantially closer to parity while maintaining competitive predictive performance. In this dataset, the fairness gains are accompanied by improvements or near-preservation in several predictive metrics, suggesting that some checkpoints along the training trajectory may offer better joint generalization and fairness than the final checkpoint or baseline model.

On the SNAPSHOT dataset, baseline models generally exhibit smaller fairness disparities than in ADULT or MIMIC-III, with most protected attributes already lying relatively close to parity. In this lower-disparity setting, the proposed method still identifies checkpoints that improve fairness for the labels and protected attributes that remain most imbalanced, typically with modest changes in predictive performance. This is a useful stress test for the framework: when the baseline model is already near parity, the method does not necessarily produce dramatic shifts, but can still recover checkpoints with slightly improved fairness profiles without requiring retraining under an explicit fairness constraint.

We also considered an artificially biased SNAPSHOT setting created by selectively removing samples from female-identifying participants to induce a stronger gender imbalance. In this more adverse setting, the baseline model shows a clearer deviation from parity, and the proposed checkpoint-screening procedure again identifies a checkpoint that moves DIR substantially closer to 1. This result suggests that the method can be responsive when stronger group imbalance is present, although the accompanying reduction in predictive performance underscores the broader point that fairness improvements are not free and remain application-dependent.

Taken together, these experiments suggest that the proposed method is best understood as a practical checkpoint-selection framework for navigating fairness-performance trade-offs in neural models. Its main strength is not that it consistently outperforms every dedicated fairness baseline, but that it offers a simple and interpretable way to inspect, compare, and select among candidate checkpoints using a pre-specified validation-set protocol. In settings where fairness and predictive quality evolve differently across training, this can provide a useful alternative to selecting the final epoch or optimizing a single scalarized objective.

7.2 Saliency-Based Qualitative Analysis

We use saliency maps as a qualitative post hoc diagnostic to compare baseline models with the validation-selected checkpoints. These visualizations are intended to provide descriptive insight into whether the selected checkpoints exhibit different feature-sensitivity patterns, rather than to establish a causal explanation for the observed fairness changes.

Across datasets, the saliency maps suggest that the selected checkpoints can differ meaningfully from baseline models in how importance is distributed across input features and, where applicable, across time steps. In some cases, the selected checkpoints appear to place less emphasis on features plausibly associated with protected attributes or their proxies; in others, the differences are more diffuse and reflect broader shifts in feature sensitivity. We interpret these patterns as evidence that the checkpoint-screening procedure may select models with different internal representations, rather than as proof of a specific debiasing mechanism.

These observations are consistent with the broader framing of our method as a selection procedure over saved checkpoints. Because the final model is chosen using validation-set fairness and predictive performance, it is reasonable that the selected checkpoint may differ from the default final checkpoint in its saliency profile. At the same time, saliency maps have well-known limitations and do not establish causal feature importance. We therefore treat them as complementary qualitative evidence rather than as independent confirmation of the mechanism underlying fairness improvements.

7.3 Limitations

Our study has several important limitations. First, the proposed method is not fully model-agnostic: it is designed for neural architectures that support MTL and MC dropout-based stochastic inference. As a result, its applicability is narrower than that of some pre-processing methods such as reweighing. Second, our baseline comparisons, while representative, are not exhaustive. More specialized fairness interventions tailored to particular notions of group or individual fairness may provide stronger performance in settings aligned with their design assumptions. Third is the restriction to binary outcomes imposed by the tools and techniques utilized for bias mitigation and the outcome labels.

Fourth, the Pareto frontier in our framework is constructed over saved training checkpoints, rather than over independently optimized models. Consequently, the frontier should be interpreted as a practical approximation to the fairness–performance trade-off landscape induced by training dynamics, rather than as a globally optimized Pareto frontier over all possible models.

Fifth, while our results suggest that uncertainty-aware checkpoint screening can recover candidate models with improved fairness–performance trade-offs, the relationship between protected-label uncertainty and downstream fairness is currently empirical rather than theoretically guaranteed. Formal analysis of when and why this relationship holds remains an important direction for future work.

Additionally, working with deep learning models, we have also encountered limitations associated with the computational complexity of using MC Dropout. Using this method along with the aspect of saving and loading model weights results in a higher than normal computational cost and time.

A further limitation is that the effectiveness of checkpoint screening depends on the chosen fairness objective. Because different fairness criteria can conflict, conclusions drawn under DIR may not transfer directly to settings where equal opportunity, equalized odds, or individual-fairness criteria are primary.

8 Conclusion

We presented a neural in-processing framework for uncertainty-aware checkpoint selection that combines MTL, MC dropout, and Pareto-based screening to navigate fairness–performance trade-offs in predictive modeling. Across three application domains, our results show that checkpoint-level screening can often identify candidate models with improved demographic-parity trade-offs relative to standard baselines while maintaining competitive predictive performance.

Rather than proposing a universally applicable fairness intervention, we view this work as a practical model-selection framework for neural pipelines in which multiple checkpoints naturally arise during training and fairness–performance trade-offs must be made explicit. Our results suggest that checkpoint-level predictive variability may provide a useful signal for identifying promising fairness-aware candidates, but the mechanism linking uncertainty and fairness remains an open question.

The present study is limited by its reliance on neural architectures with MC dropout and by its primary focus on a demographic-parity-style fairness criterion. Future work should evaluate the framework under alternative fairness objectives, stronger baseline families, and more formal analyses of when uncertainty-aware screening yields reliable fairness gains.

Broader Impact Statement

This work introduces a fairness-aware checkpoint-screening framework for neural models, designed to help identify model configurations that better balance predictive performance and group fairness. By making fairness-performance trade-offs explicit at the model-selection stage, the approach may support more transparent decision-making in applications such as healthcare, finance, and social systems.

However, several limitations and risks should be noted. The framework optimizes for a specific group fairness metric (disparate impact ratio), and improvements under this metric do not guarantee fairness under alternative definitions. The method also depends on the availability and quality of protected-attribute labels, which may be incomplete or imperfect in practice. Additionally, because fairness is addressed at the model-selection stage, the approach does not mitigate upstream sources of bias in data collection or problem formulation.

Finally, the method should be viewed as a tool for exploring fairness-performance trade-offs rather than a guarantee of equitable outcomes. Responsible deployment requires domain-specific evaluation, careful metric selection, and ongoing monitoring.

References

Soyed Tuhin Ahmed, Kamal Danouchi, Michael Hefenbrock, Guillaume Prenat, Lorena Anghel, and Mehdi B Tahoori. Scale-dropout: Estimating uncertainty in deep neural networks using stochastic scale. *arXiv preprint arXiv:2311.15816*, N/A(N/A):N/A, 2023.

Mohammadsina Almasi, Nazanin Nezami, Francesco Di_Carlo, Abolfazl Asudeh, and Hadis Anahideh. Adaptive pareto exploration (apex) for fairness-aware hyperparameter optimization in fairpilot. *Information and Software Technology*, 189(C), 2026.

Ana Arias-Oliveras. Neonatal blood gas interpretation. *Newborn and Infant Nursing Reviews*, 16(3):119–121, 2016.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

Stefan Buijsman. Navigating fairness measures and trade-offs. *AI and Ethics*, X(Y):1–12, 2023.

Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, 21(2):277–292, 2010.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 13–18. IEEE, 2009.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7), April 2024. ISSN 0360-0300. doi: 10.1145/3616865. URL `https://doi.org/10.1145/3616865`.

Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021.

Pádraig Cunningham and Sarah Jane Delany. Underestimation bias and underfitting in machine learning. In *International Workshop on the Foundations of Trustworthy AI Integrating Learning, Optimization and Reasoning*, pp. 20–31, Cham, 2021. Springer, Springer.

Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data*, 5(2):120–134, 2017.

API Dark Sky. Dark sky api, 2015.

Evert de Jonge, Linda Peelen, Peter J Keijzers, Hans Joore, Dylan de Lange, Peter HJ van der Voort, Robert J Bosman, Ruud AL de Waal, Ronald Wesselink, and Nicolette F de Keizer. Association between administered oxygen, arterial partial oxygen pressure and mortality in mechanically ventilated intensive care unit patients. *Critical care*, 12(6):1–8, 2008.

Devinder Singh Dhindsa, Jay Khambhati, William M Schultz, Ayman Samman Tahhan, and Arshed A Quyyumi. Marital status and outcomes in patients with cardiovascular disease. *Trends in cardiovascular medicine*, 30(4):215–220, 2020.

Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Awadallah, and Xia Hu. Fairness via representation neutralization. *Advances in Neural Information Processing Systems*, 34:12091–12103, 2021.

Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pp. 973–978. Lawrence Erlbaum Associates Ltd, 2001.

Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6):2074–2152, 2022.

Centers for Disease Control, Prevention, et al. Identifying vulnerable older adults and legal options for increasing their protection during all-hazards emergencies: A cross-sector guide for states and communities. *Atlanta: US Department of Health and Human Services*, 15, 2012.

Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 329–338, New York, 2019. ACM.

Stepfany Fuentes and Yuvraj S Chowdhury. Fraction of inspired oxygen. *Journal Name*, 1(1):xx–yy, 2020.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059, New York, NY, USA, 2016. PMLR, PMLR.

L Gao, H Zhan, A Chen, and V Sheng. Mitigate gender bias using negative multi-task learning, 2022. *URL: https://doi. org/10.21203/rs*, 3:–, 2022.

Cristina Gorrostieta, Reza Lotfian, Kye Taylor, Richard Brutti, and John Kane. Gender de-biasing in speech emotion recognition. In *INTERSPEECH*, pp. 2823–2827, Graz, Austria, 2019. ISCA.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18, 2019.

Maria Heuss, Daniel Cohen, Masoud Mansoury, Maarten de Rijke, and Carsten Eickhoff. Predictive uncertainty-based bias mitigation in ranking. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 762–772, New York, NY, USA, 2023. ACM.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Shobhit Jain and Lindsay M Iverson. Glasgow coma scale. *Journal Name*, 1(1):1–5, 2018.

Oliver P John, Sanjay Srivastava, et al. The big-five trait taxonomy: History, measurement, and theoretical perspectives. 1999.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Norman J Johnson, Eric Backlund, Paul D Sorlie, and Catherine A Loveless. Marital status and mortality: the national longitudinal mortality study. *Annals of epidemiology*, 10(4):224–238, 2000.

Mohammad Mahdi Kamani, Rana Forsati, James Z Wang, and Mehrdad Mahdavi. Pareto efficient fairness in supervised learning: From extraction to tracing. *arXiv preprint arXiv:2104.01634*, arXiv(Preprint), 2021.

Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*, pp. 1–6. IEEE, 2009.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 35–50, Cham, 2012. Springer, Springer.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 100–109, 2019.

Ron Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pp. 202–207, Menlo Park, CA, 1996. AAAI Press.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

Can Li, Sirui Ding, Na Zou, Xia Hu, Xiaoqian Jiang, and Kai Zhang. Multi-task learning with dynamic re-weighting to achieve fairness in healthcare predictive modeling. *Journal of biomedical informatics*, 143: 104399, 2023.

Annie Liang, Jay Lu, and Xiaosheng Mu. Algorithmic design: Fairness versus accuracy. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 58–59, New York, NY, USA, 2022. ACM.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Camille Olivia Little, Michael Weylandt, and Genevera I Allen. To the fairness frontier and beyond: Identifying, quantifying, and optimizing the fairness-accuracy pareto frontier. *arXiv preprint arXiv:2206.00074*, arXiv(Preprint):xx–yy, 2022.

Suyun Liu and Luis Nunes Vicente. Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *Computational Management Science*, XX(YY):1–25, 2022.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Tom M Mitchell. *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research ..., New Brunswick, NJ, 1980.

Rashmi Nagpal, Rasoul Shahsavarifar, Vaibhav Goyal, and Amar Gupta. Optimizing fairness and accuracy: a pareto optimal approach for decision-making. *AI and Ethics*, 5(2):1743–1756, 2025.

Luca Oneto, Michele Doninini, Amon Elders, and Massimiliano Pontil. Taking advantage of multitask learning for fair classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 227–237, New York, NY, USA, 2019. ACM.

Gökhan Özbulak, Oscar Jimenez-del Toro, Maíra Fatoretto, Lilian Berton, and André Anjos. A multi-objective evaluation framework for analyzing utility-fairness trade-offs in machine learning systems. *arXiv preprint arXiv:2503.11120*, 2025.

Tiago Palma Pagano, Rafael Bessa Loureiro, Maira Matos Araujo, Fernanda Vitoria Nascimento Lisboa, Rodrigo Matos Peixoto, Guilherme Aragao de Sousa Guimaraes, Lucas Lisboa dos Santos, Gustavo Oliveira Ramos Cruz, Ewerton Lopes Silva de Oliveira, Marco Cruz, et al. Bias and unfairness in machine learning models: a systematic literature review. *arXiv preprint arXiv:2202.08176*, N/A(N/A):1–10, 2022.

Sunghoon Park, Kyeongman Jeon, Dong Kyu Oh, Eun Young Choi, Gil Myeong Seong, Jeongwon Heo, Youjin Chang, Won Gun Kwack, Byung Ju Kang, Won-Il Choi, et al. Normothermia in patients with sepsis who present to emergency departments is associated with low compliance with sepsis bundles and increased in-hospital mortality rate. *Critical care medicine*, 48(10):1462–1470, 2020.

Yoonyoung Park, Jianying Hu, Moninder Singh, Issa Sylla, Irene Dankwa-Mullan, Eileen Koski, and Amar K Das. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. *JAMA network open*, 4(4):e213909–e213909, 2021.

Jessica K Paulus and David M Kent. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ digital medicine*, 3(1):1–8, 2020.

Annunziata Paviglianiti and Eros Pasero. Vital-ecg: a de-bias algorithm embedded in a gender-immune device. In *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT*, pp. 314–318, Online, 2020. IEEE, IEEE.

Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30:xx–yy, 2017.

Xiaohan Qin, Xiaoxing Wang, and Junchi Yan. Revisiting fairness in multitask learning: A performance-driven approach for variance reduction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20492–20501, 2025.

Ricardo Trainotti Rabonato and Lilian Berton. A systematic review of fairness in machine learning. *AI and Ethics*, 5(3):1943–1954, 2025.

Eliane Röösli, Selen Bozkurt, and Tina Hernandez-Boussard. Peeking into a black box, the fairness and generalizability of a mimic-iii benchmarking model. *Scientific Data*, 9(1):1–13, 2022.

Lucas Rosenblatt and R. Teal Witter. Fairlyuncertain: A comprehensive benchmark of uncertainty in algorithmic fairness, 2024. URL `https://arxiv.org/abs/2410.02005`.

Farid Sadaka, Darshan Patel, and Rekha Lakshmanan. The four score predicts outcome in patients after traumatic brain injury. *Neurocritical care*, 16:95–101, 2012.

Akane Sano, Andrew J Phillips, Z Yu Amy, Andrew W McHill, Sara Taylor, Natasha Jaques, Charles A Czeisler, Elizabeth B Klerman, and Rosalind W Picard. Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. In *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pp. 1–6, Boston, MA, USA, 2015. IEEE, IEEE.

Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 99–106, New York, NY, USA, 2019. ACM.

Matthias Schmal and Patrick Mäder. Reliable uncertainty estimates in deep learning with efficient metropolis-hastings algorithms. *Nature Communications*, 2026.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31:xx–yy, 2018.

Cristina Helena Costanti Settervall, Regina Marcia Cardoso de Sousa, and Silvia Cristina Fürbringer Silva. In-hospital mortality and the glasgow coma scale in the first 72 hours after traumatic brain injury. *Revista latino-americana de enfermagem*, 19:1337–1343, 2011.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, N/A(N/A), 2013. arXiv preprint.

Hua Tang, Lu Cheng, Ninghao Liu, and Mengnan Du. A theoretical approach to characterize the accuracy-fairness trade-off pareto frontier. *arXiv preprint arXiv:2310.12785*, 2023.

Sara Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Transactions on Affective Computing*, 11(2):200–213, 2017.

Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, and Ed H Chi. Understanding and improving fairness-accuracy trade-offs in multi-task learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1748–1757, New York, NY, USA, 2021. ACM.

Susan Wei and Marc Niethammer. The fairness-accuracy pareto front. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(3):287–302, 2022.

Shen Yan, Di Huang, and Mohammad Soleymani. Mitigating biases in multimodal personality assessment. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 361–369, Virtual Event, Netherlands, 2020. ACM.

Yu Yang, Aayush Gupta, Jianwei Feng, Prateek Singhal, Vivek Yadav, Yue Wu, Pradeep Natarajan, Varsha Hedau, and Jungseock Joo. Enhancing fairness in face detection in computer vision systems by demographic bias mitigation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 813–822, Oxford, United Kingdom, 2022. ACM.

Tetsuya Yumoto, Hiromichi Naito, Takashi Yorifuji, Toshiyuki Aokage, Noritomo Fujisaki, and Atsunori Nakao. Association of japan coma scale score on hospital arrival with in-hospital mortality among trauma patients. *BMC emergency medicine*, 19(1):1–7, 2019.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pp. 962–970, Proceedings of Machine Learning Research, 2017. PMLR, PMLR.

Khadija Zanna, Kusha Sridhar, Han Yu, and Akane Sano. Bias reducing multitask learning on mental health prediction. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–8, New York, NY, 2022. IEEE, IEEE.

Yu Zhang and Qiang Yang. An overview of multi-task learning. *National Science Review*, 5(1):30–43, 2018.

Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. Towards fair classifiers without sensitive attributes: Exploring biases in related features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pp. 1433–1442, New York, NY, USA, 2022. ACM.

A Appendix

A.1 Data Description

In this appendix we provide further description of the ADULT, MIMIC-III and SNAPSHOT datasets and our data processing procedures.

A.1.1 ADULT Data

The ADULT dataset was drawn from the 1994 United States Census Bureau data and involves using personal details such as education level to predict whether an individual will earn more or less than $50,000 per year (Kohavi et al., 1996). This dataset has been extensively utilized in ML bias and fairness research (Zafar et al., 2017; Friedler et al., 2019; Kearns et al., 2019; Du et al., 2021; Fabris et al., 2022). The extracted features from the dataset used in this work are summarized in Table 6.

Table 6: ADULT: List of extracted features

Feature	Classes
Age	Continuous
Education	Preschool, 1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th, HS-grad, Some-college, Assoc-voc, Assoc-acdm, Prof-school, Bachelors, Masters, Doctorate
Marital Status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
Occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
Race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
Relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
Sex	Female, Male
Workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, State-gov, Without-pay, Never-worked

This dataset contains 14 categorical and continuous integer attributes from which we dropped columns with noisy data (over 100 missing values). We were left with 8 features: Age, Workclass, Education, Marital status, Occupation, Relationship, Race, and Sex.

The prediction task for this dataset is a binary income/salary prediction, where a label of 1 represents participants who earn $>$ 50k and 0 for those who earn $<=$ 50k in a year. We converted all the string features to integers, and we used 45222 of 48842 instances after dropping a total of 3620 samples with noisy data. We utilized age, sex (gender), and race as the protected labels in our experiments.

For fairness analysis, we binarized the 3 protected labels to 0 as unprivileged and 1 as privileged classes. For the age label, 12% of the female class earned more than 50k in salary, while for the male class, 32% earned more than 50k. 36% of participants above the age of 40 earned more than 50k, while 17% of those below the age of 40 earned more than 50k. As for the race label, less than 13% of the participants who identified as Black, Amer-Indian-Eskimo, and "Other" earned more than 50k, while about 30% of those who identified as White or Asian-Pacific -Islander earned more than 50k. We chose the class corresponding to higher income/salary as the privileged class for each protected label. These privileged classes are also the majority classes in the dataset. For the age label, since the difference in the number of participants between classes is larger than the difference in average salary earned, we chose the class with more participants (less than or equal to 40) as the privileged class (1), while the class with fewer participants (over 40) as the unprivileged class (0). For race, Black(0), Amer-Indian-Eskimo(4), and other(2) became 0, and White(3) and Asian-Pac-Islander(1) became 1, and for sex, male became 1 and female became 0. We excluded the protected labels as features in the target-prediction model. For the checkpoint-screening experiments, we partitioned the data into training, validation, and held-out test splits. The validation split was used for checkpoint screening and final checkpoint selection, while the test split was reserved for final reporting.

A.1.2 MIMIC-III Data Description

The MIMIC-III ('Medical Information Mart for Intensive Care') data set is a large, freely-available database that contains de-identified health-related data associated with over 40,000 patients who stayed in critical

care units of the Beth Israel Deaconess Medical Center. It contains detailed information regarding the care of patients, which includes vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, patient demographics and in-hospital mortality (IHM), laboratory test results, billing information, etc. The complete data description can be found in the publication (Johnson et al., 2016).

The risk of mortality is often formulated as binary classification using observations recorded from a limited window of time following admission (Harutyunyan et al., 2019), and for our experiments on MIMIC-III, we predicted IHM as a binary label, so the target label indicates whether the patient died before hospital discharge. We processed and cleaned the data according to the method in (Harutyunyan et al., 2019), where there are two major steps in the data processing pipeline. First, the root cohort is extracted based on the following exclusion criteria: Hospital admissions with multiple ICU stays or ICU transfers are excluded to reduce any possible ambiguity between outcomes, and ICU stays. Patients younger than 18 are excluded as well. Lastly, event entries are only retained if they can be assigned to a hospital and ICU admission and are part of the list of 17 physiological variables used for modeling, namely: Capillary refill rate, Diastolic blood pressure, Fraction inspired oxygen, Glasgow coma scale, eye opening, motor response, verbal response, glucose, heart rate, height, mean arterial pressure, Oxygen saturation, respiratory rate, Systolic blood pressure, temperature, weight, and pH (Röösli et al., 2022). From the root cohort, the IHM cohort was extracted by further excluding all ICU stays for which "length of stay" is unknown or less than 48 hours or for which there are no observations in the first 48 hours. The complete process of patient filtering is described in Figure 10 (Harutyunyan et al., 2019). This provided final training and test sets of 17,903 and 3,236 ICU stays, respectively. Similar to (Harutyunyan et al., 2019), We determined IHM by comparing patient date of death with hospital admission and discharge times. The resulting mortality rate is 13.23% (2,797 of 21,139 ICU stays) (Harutyunyan et al., 2019).

The MIMIC-III dataset has some missing data in some instances. For these, we imputed the missing values using the most recent measurement value if it exists and a pre-specified “normal” value otherwise. These pre-specified values are provided by the authors in (Harutyunyan et al., 2019). For categorical values, we encoded them using a one-hot vector and standardized the numeric inputs by subtracting the mean and dividing by the standard deviation. We calculated the statistics per variable after the imputation of missing values. After imputing the missing values and standardizing, we obtained 17 pairs of time series for each ICU stay. We have a binary target label for IHM for each stay, determining whether or not a patient dies in the hospital.

From this dataset, the demographic data we utilized in our experiments include age, gender, marital status, and insurance, referred to as the protected labels in this paper. Table 7 shows the distribution of patients, ICU stays, and IHM rates in our data based on the different demographic groups. The age distribution is skewed towards older patients, with over 80% above 50 years. As for ethnicity, two-thirds of the patients are non-Hispanic White. In terms of health insurance, according to (Röösli et al., 2022), health insurance type is included as a socioeconomic proxy, and this led to us including it in our experiments. In general, three major insurance types are available in the United States: The public programs Medicare and Medicaid, and private insurance providers. Medicare is a federal social insurance program, and anyone over 65 years or with certain disabilities qualifies for it, whereas Medicaid provides health insurance to very low-income children and their families. Insurance data was therefore mapped to the four distinct categories Medicare, Medicaid, private, and other insurance. In this dataset, roughly a third of the patients are covered by private insurance, and more than 50% are enrolled in Medicare. The overall IHM rate in the MIMIC-III dataset is 13.23%, and IHM increases with age. As for gender, female patients appear to have a slightly higher IHM risk, and there is also a large range of variability for ethnic and socioeconomic groups.

With this dataset, we predicted in-hospital mortality (IHM) as a binary label. We determined IHM by comparing the patient’s date of death with hospital admission and discharge times. The resulting mortality rate is 13.23% (2,797 of 21,139 ICU stays). We processed data by reproducing the method in (Harutyunyan et al., 2019). A full description can be found in the original publication. To process the data for fairness evaluation, we binary-encoded each protected label and assigned them as privileged and unprivileged classes to analyze them using our chosen fairness metrics. We encoded all participants with age $<= 60$ as the unprivileged class (0), and those > 60 as the privileged class (1). The reasoning behind this encoding is that

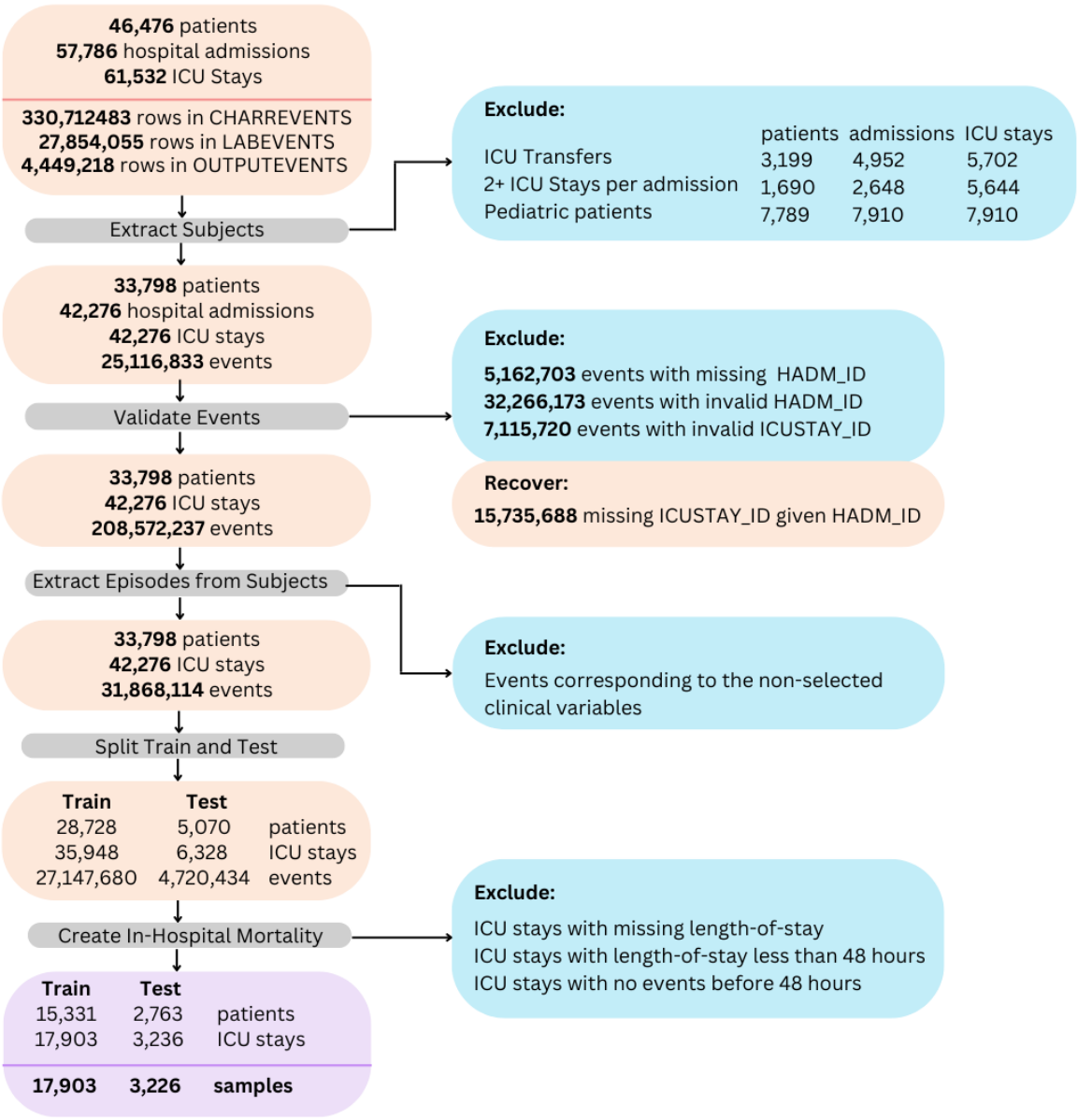


Figure 10: MIMIC-III: Data Preparation Workflow

Table 7: MIMIC-III: Demographic Distributions

	Patients n(%)	**IHM rate (%)**
Totals	18,094	13.23
Age		
0-17	0 (0.0)	0
18-29	782 (4.3)	5.6
30-49	2,680 (14.8)	9.3
50-69	6,636 (36.7)	11.1
70-89	7,043 (38.9)	16.5
90+	953 (5.3)	21.8
Gender		
Female	8,090 (44.7)	13.5
Male	10,004 (55.3)	13
Insurance		
Medicare	10,337 (57.1)	15.3
Medicaid	1,489 (8.2)	10.3
Private	5,601 (31.0)	10.2
Other	667 (3.7)	11.6
Marital Status		
Married	8,564 (45.3)	46.4
Single	4,422 (24.4)	19.3
Widowed	2,654 (14.7)	17.4
Divorced	1,119 (6.2)	5.4
Separated	194 (1.0)	1.1
Unknown (default)	302 (1.7)	1.1
Life partner	5 (0.03)	0

the Centers for Disease Control and Prevention (CDC) defines an "older adult" as someone who is at least 60 years old (for Disease Control et al., 2012). For the Gender label, male was encoded as the privileged class (1) due to its higher number of samples and because female patients appear to have a slightly higher IHM risk, making the female class the unprivileged class (0). We classified the participants according to whether they have private or government-assisted insurance. Those with private or self-care insurance represented the privileged class (1), and those with Medicare, Medicaid, and Government represented the unprivileged class (0). The reasoning behind this is the connection between government-aided insurance and socioeconomic conditions and age as mentioned earlier in this section. For marital status, we categorized married patients, those with life partners as privileged, and all others (no partner) as unprivileged. We did this in accordance with research showing that unmarried patients, including those who are divorced, separated, widowed, or never married, show elevated rates of certain health complications and death compared to those with partners (Dhindsa et al., 2020; Johnson et al., 2000).

A.1.3 SNAPSHOT Data Description

The SNAPSHOT dataset includes multi-modal physiological, behavioral, and survey data collected from 350 college students in one university, totaling over 7500 days of data (Taylor et al., 2017). The study was conducted between 2013 and 2017, within which 30 days of consecutive data were recorded for 228 participants and 90 days for 15 participants. The data recorded includes wrist-worn wearable sensor data (acceleration, skin conductance, and skin temperature), mobile phone data (call, SMS, screen on/off logs, GPS locations), weather data (obtained using DarkSky API (Dark Sky, 2015)), self-reported daily morning and evening well-being (non-numeric scales or mood, health, and stress later scored 0-100), genders, and Big Five Personality scores (John et al., 1999).

From the SNAPSHOT data, we computed a total of 378 features, including 173 physiological features that include Electrodermal activity (EDA) features, skin conductance Level (SCL), skin temperature features,

accelerometer features, 165 mobile phone features, 43 weather features from the raw data, and statistical features calculated from each of these classes of features. Tables 8, 9, 10, 11, 12, and 13 summarize these features. For more details on the features extracted, refer to (Taylor et al., 2017).

Table 8: SNAPSHOT: Descriptions of Sleep and Nap Related Survey Data

Features	**Description**
State Score	State anxiety score
no_sleep_24	Number of times participant slept in 24 hours.
sleep_latency	Time taken to fall asleep
pre_sleep_activity	What activity the participant performed before going to sleep.
awakening	If the participant woke up through the night (yes or no)
awakening_occations	Number of awakenings
wake_reason	How the participant woke up (spontaneously, alarm, or disturbance).
count_awakening	How many times participant woke up through the night.
awakening_duration	How long was the participant awake for if they woke up through the night.
nap	If the participant took a nap (yes or no)
nap_occations	Number of naps
count_nap	How many times the participant took a nap.
nap_duration	How long the nap was.
first_event_none	If an event is scheduled that day (yes or no)
time_in_bed	How long the participant spent in bed for the day.
sleep_try_time_mins_since_midnight	What time the participant tries to go to sleep
wake_time_mins_since_midnight	Wake time in minutes since midnight
first_event_mins_since_midnight	first event time in minutes since midnight
presleep_media_interaction	If the participant has presleep media interaction (yes or no)
presleep_personal_interaction	If the participant has presleep personal interaction (yes or no)
positive_interaction	If the participant had any positive interactions with someone for the day (yes or no).
negative_interaction	If the participant had any negative interactions with someone for the day (yes or no).

From the SNAPSHOT data, we predicted 4 self-reported mood labels, namely morning and evening happiness and calmness levels, using physiological features, mobile phone usage, and weather. We utilized the previous seven (7) days of data to predict the current day's label. These labels appear as a score of 1 to 100 in the data set, and we binarized them with all scores less than or equal to 50 to mean sad or stressed and above 50 means happy or calm.

The SNAPSHOT dataset contains numerous demographic information on the participants, and we used 3 of these: gender, race, and ethnicity as protected labels. For these labels, we chose the class with more participants as the privileged class, following the concept of negative legacy or data bias (Cunningham & Delany, 2021). We also used personality types, including openness, conscientiousness, extraversion, agreeableness, and neuroticism, as protected labels in our experiments. Figure 11 shows a distribution of all the protected labels across this dataset, where non-white refers to all other races other than white, which include Asian (77 participants), Black or African American (38 participants), American Indian or Alaskan Native (6), and other (13 participants). H\L refers to Hispanic and Latino in the figure.

Some participants in the SNAPSHOT dataset do not have race and ethnicity information, and we dropped samples for such participants when running experiments for race and ethnicity labels. In the end, we dropped 314 samples from the training data, reducing it from 4030 to 3716 samples, and the test samples were reduced from 1866 to 1715. For the checkpoint-screening experiments, we partitioned the data into training, validation, and held-out test splits. The validation split was used for checkpoint screening and final checkpoint selection, while the test split was reserved for final reporting.

Table 9: SNAPSHOT: Descriptions of Activity Related Survey Data

Features	**Description**
academic	If the participant attended any academic activities (yes or no).
count_academic	How many academic activities the participant attended.
academic_duration	How long the participant spent on academic activities
study_duration	How many hours participant studied for, outside of academic activities.
exercise	If the participants engaged in any exercise-based activities (yes or no)
exercise_occations	How many times the participant engaged in exercise-based activities.
exercise_duration	For how long the participant exercised.
extracurricular	If the participant attended any other extracurricular activities (yes or no).
count_extracurricular	How many extracurricular activities the participant attended in the day.
extracurricular_duration	How long the participant attended extracurricular activities for.
overslept	If the participant overslept and missed any scheduled events.
caffeine_count	Total servings of caffeine participant had for the day.
drugs	If the participant had any other drugs or medication besides (yes or no).
drugs_alcohol	If the participant had any alcohol (yes or no).
drugs_alert	If the participant had any drugs to keep them alert (yes or no).
drugs_sleepy	If the participant had any drugs that made them sleepy (yes or no).
drugs_tired	If the participant had any drugs that made them tired (yes or no).

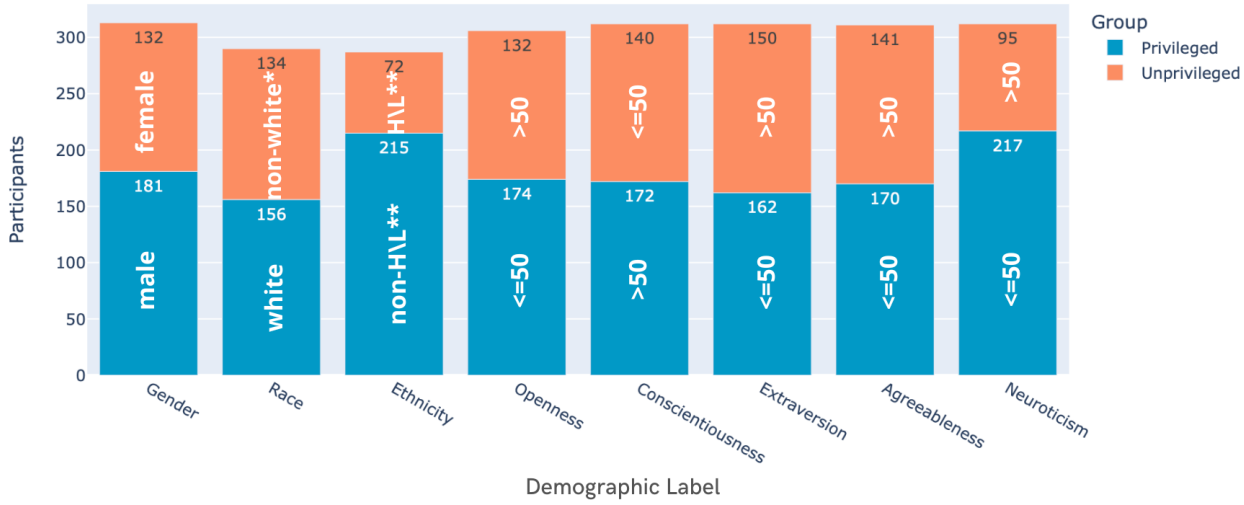


Figure 11: SNAPSHOT Demographic Distribution

Table 10: SNAPSHOT: Descriptions of Electrodermal Activity and Skin Conductance Features Extracted from Wearable Sensor Data

Features	**Description**
Electrodermal activity (EDA) Peak Features	
Sum AUC	Sum of the AUC of all peaks for this period where the amplitude of the peak is calculated as the difference from base tonic signal
Sum AUC Full	Sum of AUC of peaks where amplitude is calculated as difference from 0
Median RiseTime	Median rise time for peaks (seconds)
Median Amplitude	Median amplitude of peaks (μS)
Count Peaks	Number of detected peaks
SD Peaks 30 min	Compute number of peaks per 30 minute epoch, take standard deviation of this signal
Med Peaks 30 min	Compute number of peaks per 30 minute epoch, take median of this signal
Percent Med Peak	Percentage of signal containing 1 minute epochs with greater than 5 peaks
Percent High Peak	Same as Percent Med Peak
Skin Conductance Level (SCL) Features	
Percent Off	Percentage of period where sensor was off
Max Unnorm	Maximum level of un-normalized EDA signal
Med Unnorm	Median of un-normalized EDA signal
Mean Unnorm	Mean of un-normalized EDA signal
Median Norm	Median of z-score normalized EDA signal
SD Norm	Standard Deviation of z-score normalized EDA signal
Mean Deriv	Mean derivative of z-score normalized EDA signal (μS/second)

A.2 Experiments

In this appendix, we provide more information on the experimental process from Section 5. We provide details on baseline model implementations, model structures, and parameters, and elaborate on the Reweighting method.

A.2.1 Baseline Model Implementations

For the ADULT dataset, we built a Convolutional Neural Network (CNN) with two dense layers and a dropout layer with a dropout rate of 0.25 to predict binary income labels; <= 50k as 0 and > 50k as 1. We used a 3-fold cross-validation method and binary cross entropy as the loss function with an Adam optimizer. The structure of this model is shown below.

For the MIMIC-III dataset, we used a Long-Short Term Memory (LSTM)-based model proposed in (Harutyunyan et al., 2019). LSTM is a Recurrent Neural Network (RNN) which is a type of Artificial Neural Network (ANN) designed to capture long-term dependencies in sequential data (Harutyunyan et al., 2019). To implement our model, we resampled the time series data into regularly spaced intervals. In the case of multiple measurements of the same variable in the same interval, we utilized the value of the last measurement. We used a 48-hour window for prediction, enabling the detection of patterns that may indicate changes in patient acuity, as proposed by (Harutyunyan et al., 2019). We selected the channel-wise LSTM model without deep supervision for our study based on its superior reported area under the receiver operating characteristic (AUROC) performance among the five developed non-MTL models by Harutyunyan et al. (Harutyunyan et al., 2019). The selected network consists of a bi-directional layer, an LSTM layer with 16

Table 11: SNAPSHOT: Descriptions of Accelerometer and Skin Temperature Features Extracted from Wearable Sensor Data

Features	**Description**
Accelerometer Features	
Step Count	Number of steps detected
Mean Movement Step Time	Average number of samples (at 8Hz) between two steps (aggregated first to 1 minute, then we take the mean of only the parts of this signal occurring during movement)
Stillness Percent	Percentage of time the person spent nearly motionless
Sum Stillness weighted AUC	Sum the weights of the peak AUC signal by how still the user was every 5 minutes
Sum Steps Weighted AUC	Sum the weights of the peak AUC signal by the step count over every 5 minutes
Sum Stillness Weighted Peaks	Multiply the number of peaks every 5 minutes by the amount of stillness during that period
Max Stillness Weighted Peaks	Max value for the number of peaks multiplied by the stillness for any five minute period
Sum Steps Weighted Peaks	Divide number of peaks every five minutes by step count and sum
Med Steps Weighted Peaks	Average value of number of peaks divided by step count for every 5 mins
Skin Temperature (ST) Features	
Max Raw Temp	Maximum of the raw temperature signal (C)
Min Raw Temp	Minimum of the raw temperature signal (C)
SD Raw Temp	Standard deviation of the raw temperature signal
Med Raw Temp	Standard deviation of the raw temperature signal
Sum Temp Weighted AUC	Sum of peak AUC divided by the average temp for every 5 minutes
Sum Temp Weighted Peaks	Number of peaks divided by the average temp for every 5 minutes
Max Temp Weighted Peaks	Maximum number of peaks in any 5 minute period divided by the average temperature
SD Stillness Temp	Standard deviation of the temperature recorded during periods when the person was still
Med Stillness Temp	Median of the temperature recorded during periods when the person was still

neurons, and a dropout layer with a dropout rate of 0.3 (structure shown in Appendix A). We used a batch size of 8, binary cross entropy as the loss function, and Adam optimizer. We executed this for 100 epochs, reproducing the experiments from (Harutyunyan et al., 2019).

The baseline model we used for predicting the labels for the SNAPSHOT dataset is a gated recurrent unit (GRU) model with 2 GRU layers and 2 dropout layers with a dropout rate of 0.5, 30 neurons, and focal loss with a gamma value of 2, and an alpha value of 4. This model predicts 4 labels, namely: “morning happiness", “morning calmness", “evening happiness", and “evening calmness". We trained this model for 300 epochs with a batch size of 1000 and a focal loss function since there is a high class imbalance in the target label for this dataset. Focal loss is an extension of the cross-entropy loss function that would down-weight easy examples and focus training on hard negatives (Lin et al., 2017). It reduces the loss contribution from easy examples and increases the importance of correcting misclassified examples. The structure of this model is also shown in Appendix A. With 8 protected labels, which are gender, race, ethnicity, openness, conscientiousness, extraversion, agreeableness, and neuroticism, and 4 target labels, we conducted 32 sets of experiments to analyze the bias in this dataset.

Table 12: SNAPSHOT: Descriptions of Features Extracted from Weather API 1

Features	**Description**
Sunrise	Time of sunrise (UTC)
Moon_phase	Moon phase value on a scale of 01(new moon-full moon)
Apparent_temp_max	Maximum apparent temperature of the day (°F)
Apparent_temp_min	Minimum apparent temperature of the day (°F)
Temperature_max	Maximum temperature of the day (°F)
Temperature_min	Minimum temperature of the day (°F)
Avg_cloud_cover	Percentage of sky covered by cloud on a scale of 0-1
Avg_dew_point	Average dew point temperature
Avg_humidity	Daily average value of humidity on a scale of 0-1
Avg_pressure	Average atmospheric pressure on the sea level (hPa)
Morning_pressure_change	Trinary value of pressure difference between midnight and noon (rising, falling, steady)
Evening_pressure_change	Trinary value of pressure difference between noon and midnight (rising, falling, steady)
Avg_visibility	Average visibility (meters)
weather_precip_probability	Precipitation probability
Temperature_rolling_mean	Rolling average of temperature
Temperature_rolling_std	Rolling standard deviation of temperature
apparentTemperature_rolling_mean	Rolling average of apparent temperature
apparentTemperature_rolling_std	Rolling standard deviation of apparent temperature
apparentTemperature_today_vs_avg_past	Difference between today's apparent temperature and its rolling average
pressure_rolling_mean	Rolling average of pressure
pressure_rolling_std	Rolling standard deviation of pressure
apparentTemperature_today_vs_avg_past	Difference in today's apparent temperature and its rolling average
pressure_rolling_mean	Rolling average of pressure
pressure_rolling_std	Rolling standard deviation of pressure
pressure_today_vs_avg_past	Difference between today's pressure and its rolling average
cloudCover_rolling_mean	Rolling average of cloud cover
cloudCover_rolling_std	Rolling standard deviation of cloud cover
cloudCover_today_vs_avg_past	Difference between today's cloud cover and its rolling average
humidity_rolling_mean	Rolling average humidity
humidity_rolling_std	Rolling standard deviation humidity
humidity_today_vs_avg_past	Difference between today's humidity and its rolling average

A.2.2 Model Structures

A.2.3 Proposed Method Implementation Details

We applied the proposed fairness-aware checkpoint-screening framework by training MTL models that jointly predict the target label and the selected protected attribute. The MTL stage generates a trajectory of saved checkpoints, which are later evaluated and selected using validation-set fairness and predictive-performance metrics. We assigned task-loss weights to prioritize the primary prediction task while still allowing the protected-attribute task to shape the shared representation. For ADULT, the target/protected loss-weight ratio was 4.5/0.25; for MIMIC-III, 5/0.5; and for SNAPSHOT, 6/0.25. These ratios were chosen through preliminary tuning to maintain stable target-task learning while producing a usable checkpoint trajectory for validation-based fairness-performance screening. For the SNAPSHOT dataset, we used focal loss for

Table 13: SNAPSHOT: Descriptions of Features Extracted from Weather API 2

Features	**Description**
windSpeed_rolling_mean	Rolling average of wind speed
windSpeed_rolling_std	Rolling standard deviation of wind speed
windSpeed_today_vs_avg_past	Difference between today's wind speed and its rolling average
precipProbability_rolling_mean	Rolling average of precipitation probability
precipProbability_rolling_std	Rolling standard deviation of precipitation probability
precipProbability_today_vs_avg_past	Difference between current precipitation probability and its rolling average
sunlight	Duration of sunlight (sec)
quality_of_day	Quality of the day defined in terms of 8 categories in the range 4, 4: clear=4, partly-cloudy=3, cloudy=2, wind=1, fog=1, rain=2, sleet=3, snow=4
avg_quality_of_day	Average value for quality_of_day
precipType	Type of precipitation as integer: None=0, Rain=1, Hail=2, Sleet=3, Snow=4, Other=5
max_precip_intensity	Maximum Precipitation volume (mm)
median_wind_speed	Median wind speed of the day (meter/sec)
median_wind_bearing	Median wind bearing of the day (degrees)

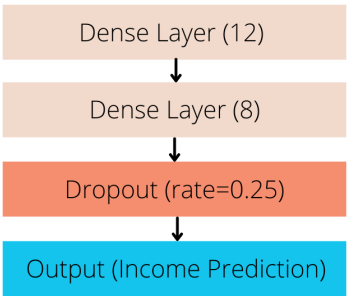


Figure 12: ADULT: Income Prediction Model Structure

both prediction heads and assigned gamma values of 2 and 0 for the target and protected-attribute heads, respectively.

A.2.4 Data Manipulation on SNAPSHOT Dataset

To create a relatively higher imbalance for the gender label in the SNAPSHOT dataset, we manipulated the data in the following way:

Originally, the data contained samples from 181 male and 132 female participants. There was a total of 4030 train and 1866 test samples, and we removed the samples from 75 participants who identified as female from the training data, making it a total of 2660 samples for training. We removed participants with 14 or more samples in the train data (experiment S-1). We also experimented with removing samples from participants with 5 or more samples in the train data (experiment S-2) and those with 30 or more samples (experiment S-3), leaving 2349 and 3685 samples, respectively.The motivation was to induce an artificial representation imbalance, sometimes described as a form of negative legacy bias, so that we could evaluate whether validation-based checkpoint screening could recover a more favorable fairness-performance trade-off under stronger group imbalance. Because the held-out test data still contains a more comparable number of female samples, this manipulation allows us to examine whether a model trained on the altered training

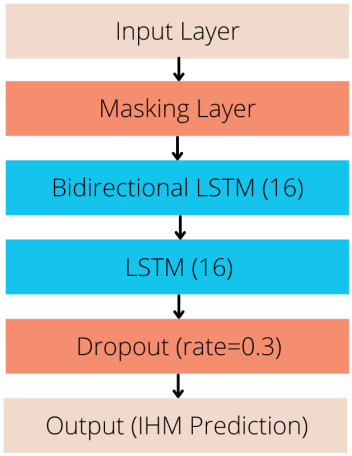


Figure 13: MIMIC-III: IHM prediction Model Structure

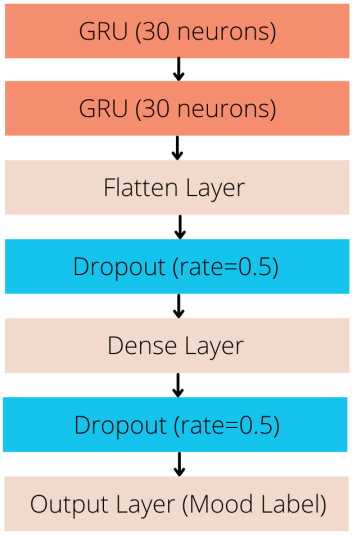


Figure 14: SNAPSHOT: Mood Labels Prediction Model Structure

distribution exhibits stronger group disparity at evaluation time, consistent with prior work on representation imbalance (Gorrostieta et al., 2019).

A.2.5 Reweighting

We compare the selected checkpoints from our proposed framework to a standard pre-processing fairness baseline, reweighting, introduced by Kamiran et al. (Kamiran and Calders, 2012). Reweighting is a pre-processing bias mitigation technique. It involves applying appropriate weights to different tuples in the training dataset to make it discrimination free with respect to the protected label (Park et al., 2021). It generates weights for the training examples in each (group, label) combination differently to ensure fairness before classification. (Bellamy et al., 2018).

We apply the AI Fairness 360 implementation of this method to our processed data for all datasets, pass the resulting weight vectors to the corresponding baseline model, and compute fairness and performance metrics on the held-out test split for comparison with the selected checkpoint from our framework.

The Reweighting bias mitigation technique works by applying different weights to each group-label combination according to the conditional probability of the label by the protected label. It assigns different weights to various subgroups, determined by the relationship between the protected label and the outcome label. For example, if certain groups are underrepresented in positive outcomes, Reweighting increases their influence in the training process by assigning higher weights to these groups. This process aims to equalize the predictive performance across groups, mitigating any potential bias that the model may learn from imbalanced or skewed data.

Assume T and “Label" represent the protected and target labels simultaneously. Samples with T=t and Label = + will get higher weights than samples with T=t and Label = -, and samples with T≠t and Label = + will get lower weights than samples with T≠t and Label = -, where T represents the protected label, and Label represents the target label. According to these weights, the objects will be sampled (with replacement), leading to a dataset without dependency (Calders et al., 2009). These weights are calculated as follows using some basic notions of probability theory.

Assuming the dataset D is unbiased, in the sense that T and Label are independent, the expected probability $P_{exp}(t \wedge +)$ would be:

$$P_{exp}(t \wedge +) := t \times +$$

where t is the fraction of objects having T=t and ‘+’ the fraction of tuples having Label == +. In reality, however, the actual probability

$$P_{act}(t \wedge +) := t \wedge +$$

might be different. If the expected probability is higher than the actual probability value, it shows the bias towards label ‘-’ for T=t. We assign weights to t with respect to label ‘+’. The weight will be:

$$W(T = t|x(Label) = +) := \frac{P_{exp}(t \wedge +)}{P_{act}(t \wedge +)}$$

This weight of t for label ’+’ will over-sample objects with T=t for the label ’+’. The weight of t for label ’-’ becomes:

$$W(T = t|x(Label) = -) := \frac{P_{exp}(t \wedge -)}{P_{act}(t \wedge -)}$$

When applied to for example the ADULT dataset taking sex as the protected label, to remove the dependency between sex (T) and the target label (Label), we calculate a weight for each data object according to its T and Label value. In this example where both T=sex and Label are both binary attributes, only 4 combinations between the values of T and Label are possible, i.e., T=female(f) or T=male(m) can have Label values ’+’ or ’-’. We can then calculate the weight of a data sample with T=f and Label ’+’. Let’s assume that approximately 50% samples have T=m and 60% have Label value ’-’. so the expected probability of the sample can be computed as:

$$P_{exp}(Sex = m|x(label) = -) = 0.50 \times 0.60$$

but its actual probability is 20%. So the weight W will be:

$$W(Sex = m|x(Label) = -) = \frac{0.50 \times 0.60}{0.2} = 1.5$$

The weights for the other combinations are also calculated in a similar manner.

We assign to every tuple a weight according to its T and Label values. The balanced dataset is then created by sampling the original training data, replaced with the assigned weights. On this balanced dataset, the dependency-free classifier is learned. The reweighting technique can be seen as an instance of cost-sensitive learning (Elkan, 2001) in which, e.g., an object of label '+' with T=t gets a higher weight, making an error for this object more "expensive". The pseudocode of the algorithm describing this reweighting technique can be found in (Calders et al., 2009) and (Kamiran & Calders, 2009).

A.2.6 Additional Comparison Baselines

In addition to reweighting, we compare the selected checkpoints from our framework against two additional fairness baselines on the ADULT and MIMIC-III datasets: FairRF (Zhao, Dai, Shu, and Wang, 2022) and Adversarial Reweighted Learning (ARL) (Lahoti, Beutel, Chen, Lee, Prost, Thain, Wang, and Chi, 2020). These methods were selected because they represent distinct fairness-intervention strategies beyond standard pre-processing.

FairRF is a fairness-aware feature-reweighting method that leverages features related to, but not explicitly equal to, the protected attribute. It adjusts feature importance to reduce the association between these related features and model predictions while preserving predictive performance. We compare against the reported FairRF results on the ADULT and MIMIC-III datasets using the experimental settings described in the original study.

ARL is an in-processing fairness method based on adversarial reweighting. It learns sample weights during training so that examples associated with higher unfairness receive greater emphasis in the optimization process. We use the AI Fairness 360 implementation of ARL and evaluate it on the ADULT and MIMIC-III datasets under the same held-out test protocol used for the selected checkpoints from our framework.

Together, these baselines provide additional comparison across pre-processing, in-processing, and feature-level fairness interventions. Final comparisons are reported on the held-out test split after validation-based checkpoint screening for the proposed framework.

A.3 Saliency Maps for MIMIC-III and SNAPSHOT Datasets

In this appendix, we provide representative saliency maps for the MIMIC-III and SNAPSHOT experiments. These saliency maps are provided as qualitative post hoc diagnostics and should not be interpreted as causal evidence that the selected checkpoint removes or suppresses protected-attribute information. Figure 15 shows a subset of features for the SNAPSHOT experiments. In this figure, the image on the left represents the saliency map for the initial baseline model for evening-sad-happy prediction, and the image on the right represents the saliency map for the selected checkpoint from the proposed framework. Figure 16 was generated from the baseline model for IHM prediction in the MIMIC-III experiments, and Figure 17 corresponds to the selected checkpoint from the proposed framework using marital status as the protected attribute.

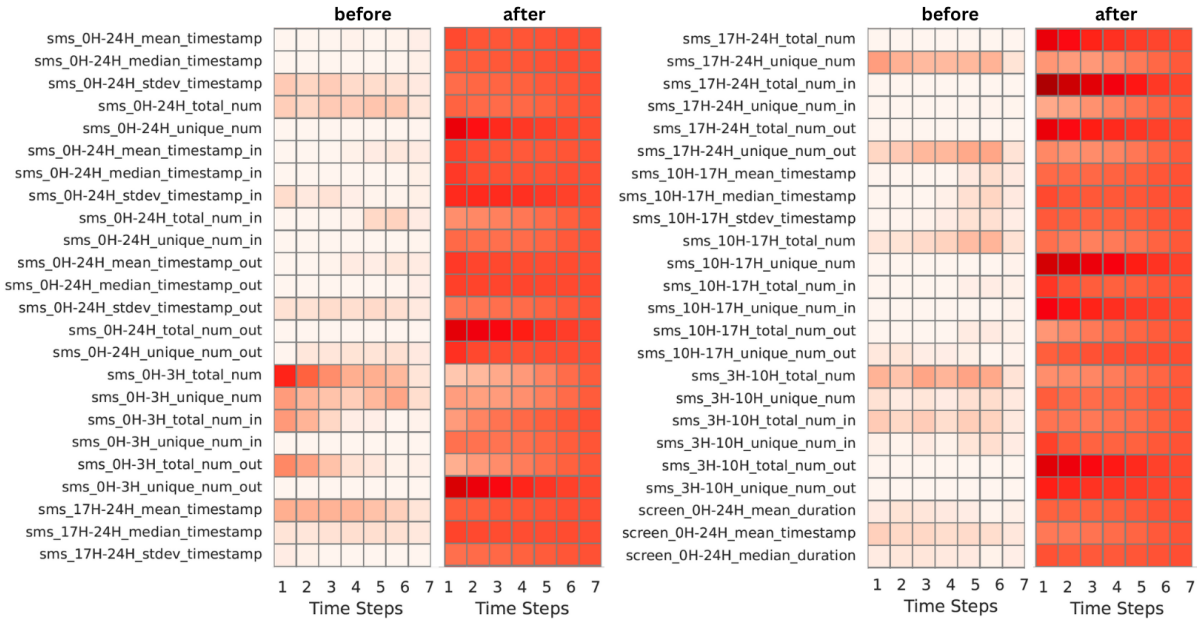


Figure 15: SNAPSHOT: Saliency Maps from Baseline Model and Model from Selected Checkpoint using Proposed Method.

Figure 16: MIMIC-III: Saliency Maps From Baseline Model for IHM prediction.

Figure 17: MIMIC-III: Saliency Maps from the Selected IHM Prediction Checkpoint Using Marital Status as the Protected Attribute.