International Journal of Computer Applications (0975 - 8887)
Volume 187 - No.74, January 2026

Causal Representation Learning for Bias Detection in AI Hiring Systems

Ajay Guyyala
Meta Platforms Inc.
Texas, USA

Prudhvi Ratna Badri Satya
Cloudflare Inc.
Texas, USA

Krishna Teja Areti
Fast Enterprises Llc.
North York, ON, Canada

Vijay Putta
Fast Enterprises Llc.
Louisiana, USA

ABSTRACT

Artificial intelligence is used widely in HR hiring systems for resume screening and ranking, yet models trained on past decisions often carry group bias through hidden paths from protected attributes to hiring outcomes. This study presents a causal representation learning framework that reduces these effects by using structural modeling, adversarial training, and counterfactual simulation. The method is tested on a structured dataset of 225 applicants and the Utrecht Fairness Recruitment Dataset with close to ten thousand records. The framework lowers the demographic parity gap from 19% to 9% and reduces the equal opportunity gap from 22% to 11%. Counterfactual consistency rises from 67.1% to 84.6%, while the Causal Disparity Index drops from 28% to 11%. Predictive performance also improves, reaching 84.3% accuracy, 82.7% precision, 79.4% recall, and an F1 score of 80.9%. Graph reconstruction error decreases from 0.071 to 0.026. These results show that causal representation learning supports fair and reliable HR hiring systems without reducing predictive strength.

General Terms

Artificial Intelligence, Causal Modeling, Fairness in Decision Systems, HR Analytics

Keywords

Causal Representation Learning, Fair Hiring Systems, Bias Detection, Counterfactual Analysis, Fairness Metrics, HR Hiring Data

1. INTRODUCTION

The comprehensive use of Artificial Intelligence (AI) in HR hiring systems aims to assist with resume screening, applicant ranking, and job suitability prediction [2]. Many organizations adopt these systems to manage large volumes of applications and reduce the time spent on manual assessments [2]. These tools can improve operational efficiency, but they also introduce new challenges [24]. Fairness and transparency remain central concerns, especially when systems are trained on biased historical data [1]. In many cases, such data reflects social inequalities that existed in prior decision-making processes [11]. As a result, AI models may replicate or amplify these patterns, unintentionally discriminating against groups based on race, gender, age, or other protected attributes [10]. This not only affects individual candidates but can also damage the credibility of the hiring process [13]. As HR hiring systems become more automated, the need for reliable fairness mechanisms becomes increasingly urgent [15]. Addressing these concerns is essential to ensure ethical AI adoption in employment [28].

Historical data used in training hiring models often contains embedded bias, even when it appears objective [6]. These biases can remain hidden within machine learning systems, especially in models that rely on deep architectures with limited interpretability [39]. Once trained, such models may associate protected features with reduced hiring likelihood, despite no explicit instruction to do so [41]. Conventional fairness interventions attempt to fix this by adjusting data distributions, adding fairness constraints, or reweighting outputs [16]. While these adjustments may reduce visible disparities, they rarely address the underlying cause of bias [48]. Without understanding how different variables influence predictions, such fixes can remain superficial [29]. To address this gap, researchers have begun exploring causal modeling, which focuses on identifying and removing pathways through which bias flows in the data [4]. Causal representation learning offers a more robust solution, as it allows the model to capture structural relationships and isolate sources of unfair influence, rather than treating bias as an afterthought [36].

A causal view of fairness shifts the focus from correlation to explanation. Instead of relying on statistical parity alone, causal models aim to identify how protected attributes affect outcomes and how that influence can be removed [33]. By learning fair representations that are not influenced by sensitive variables, such models can make predictions based on relevant and unbiased information [26]. This approach holds particular promise in high-stakes settings like hiring, where fairness and accountability are both critical [38]. Rather than merely adjusting results after training, causal models integrate fairness directly into the learning process [45]. They build structural graphs, disentangle latent features, and suppress hidden dependencies that lead to unfair outcomes. This perspective aligns fairness with accuracy, offering a more principled way to detect and mitigate bias. As awareness grows around fairness in automated decision-making, causal representation learning emerges as a promising direction for ethical and effective HR hiring systems.

The problem at the center of this work is how to build hiring models that are both fair and reliable. Most models today treat fairness as a correction applied after training [17]. This creates a gap between learning useful patterns and correcting bias. If fairness is added only later, the model may still base its decisions on biased features.

The challenge is to make fairness part of the model itself. We aim to build a method that learns clean feature representations that do not depend on protected variables [32]. This requires removing hidden influence while keeping the model accurate. Solving this problem means designing a structure that learns fair patterns from the start [22].

Many existing works address fairness using techniques such as statistical reweighting, adversarial debiasing, and fairness driven loss functions [30]. These models typically reduce sensitivity to protected attributes by enforcing group-level similarity. For instance, adversarial frameworks introduce a secondary network that penalizes the main model if it can identify protected features [23]. While such designs can help minimize discrimination, they often come at the cost of interpretability and performance [19]. Some methods modify the input data to balance group presence, while others rely on post-hoc adjustments to outputs. These techniques have shown usefulness in controlled benchmark settings but tend to falter when applied to more complex or real-world tasks. Their fairness gains often come without explanation of how changes affect the internal feature space. In employment contexts, where accountability and traceability are essential, black-box fairness remains insufficient. Practitioners need models that not only reduce disparities but also provide a clear trace of decision paths. Without transparency, fairness interventions cannot build trust or meet compliance standards [14].

A second body of work measures fairness using statistical indicators such as demographic parity, equal opportunity, and calibration across groups [8]. These metrics are widely used to report performance differences between protected and non-protected groups. But, these indicators only reflect observable outcomes, not the underlying mechanisms that produce them [12]. Some methods try to enforce these metrics during training through custom loss functions or constraints. Although this can improve fairness scores, it does not always reflect actual fairness in decision-making logic. These models may still rely on biased pathways in the data, especially when sensitive attributes correlate with unobserved variables [25]. Without a causal understanding of how features influence predictions, such models risk masking bias rather than removing it. Fairness should not rely solely on statistical parity but must account for how decisions are formed. A model that adjusts outputs without correcting internal dependencies may appear fair while remaining biased. Addressing this issue requires structural solutions that control for hidden bias in feature learning [5].

This research study introduces a causal representation learning framework for fair hiring decisions. It removes hidden paths of influence between protected attributes and predictions. The method learns structural features while controlling for sensitive information. Unlike previous models, it uses causal graphs to define fairness during learning, not after. This allows the system to produce fair outputs without discarding useful data. The model is tested on a HR hiring system dataset to measure fairness, accuracy, and consistency. Results show that it achieves better fairness and remains stable across groups. This confirms its use in real hiring systems where fairness and accuracy must go together.

The aim of this study is to develop a causal representation learning framework that detects and mitigates bias in AI-driven hiring systems by learning fair and disentangled feature representations that prevent protected attributes from influencing HR hiring system outcomes. Here are three research questions (RQs) written in a clear academic tone:

(1) How can causal representation learning be used to detect hidden bias in AI-based hiring systems?

(2) To what extent does the proposed causal framework improve fairness metrics without reducing predictive accuracy?

(3) How does the model perform across demographic groups when tested on real-world HR hiring system data with known biases?

Bias in automated hiring systems has raised concerns among regulators, practitioners, and researchers. As these systems are deployed in decision-making processes, they risk amplifying historical inequalities if left unchecked. Traditional fairness solutions rely on statistical adjustments or outcome-level corrections. These often lack transparency and struggle to explain the roots of bias in model behavior. By focusing on outcomes rather than causes, they fail to prevent biased decision paths from forming. The significance of this study lies in its shift toward causal reasoning, where bias is identified through structured relationships between variables. Using this approach, the model does not simply avoid protected attributes, but actively prevents them from shaping the learned features. This makes it possible to produce fair predictions based on unbiased evidence. As fairness becomes a requirement in legal and policy frameworks, such causal models can support compliance and build trust in algorithmic hiring tools.

This work contributes to the broader goal of building interpretable and fair AI systems for sensitive domains. HR hiring system decisions affect access to economic opportunity, and biased systems can harm individuals and communities. By integrating causal representation learning, this study offers a method that aligns technical performance with fairness goals. It introduces tools to uncover hidden influence, reconstruct decision logic, and measure fairness at both individual and group levels. The method is tested on a HR hiring system dataset that reflects real selection challenges, adding practical value to its findings. The approach also supports structural audits, helping developers trace how bias emerges and how it can be removed. Through this contribution, the study provides a step forward in making AI hiring systems both accurate and ethically grounded, promoting equal treatment without losing model utility.

The rest of this paper is organized as follows. Section 2 reviews prior studies on fairness in AI hiring systems and highlights the limitations of existing methods. Section 3 introduces the proposed causal representation learning framework and defines the structural model and training strategy. Section 4 describes the dataset, preprocessing pipeline, and selected evaluation metrics. Section 5 presents the empirical results and interprets the performance of the proposed model in comparison with existing baselines. Finally, Section 6 summarizes the contributions and outlines future research directions.

2. LITERATURE REVIEW

fairness in AI-based HR hiring systems has drawn growing interest as automated systems increasingly influence employment decisions. Recent studies have explored various approaches to detect and mitigate bias, ranging from data rebalancing and fairness-aware training to post-hoc output adjustments. But, many of these methods address bias superficially, lacking insight into its structural causes. A subset of the literature has begun to explore causal methods, aiming to block unfair influence from protected attributes while preserving model utility. These works introduce structural frameworks, fairness constraints, and intervention strategies that align with ethical hiring practices. Despite this progress, existing approaches often struggle to provide both transparency and predictive reliability. This section reviews recent journal publications from 2023 to 2025 that address these concerns, focusing on how

they define fairness, apply causal reasoning, and evaluate model behavior in real or synthetic HR hiring system datasets.

[9] developed a collaborative decision-making model that emphasized human–AI cooperation in hiring systems. The model described recruiter interaction flows but did not include experimental metrics. [40] defined a graph-based causal representation framework to separate latent factors using structured causal models. Their theoretical work focused on disentanglement in representation learning without applying it to hiring datasets. Both studies contributed conceptual foundations for causal analysis and fairness alignment. Such as, neither included empirical evaluation, which limits their direct application to bias detection. Their ideas remain useful in understanding how human factors and structural mechanisms interact in automated hiring. They also raised awareness about the underlying assumptions of AI systems. Both offered abstract tools to explain patterns of bias in algorithmic outputs. These foundational methods are often used to justify interventions. Their absence of numerical evaluation narrows their real-world scope.

[27] applied fairness metrics based on causal graphs to study how confounding affects algorithmic bias. They described multiple causal paths and tested fairness definitions but did not use empirical hiring data. [31] introduced a causal variational autoencoder to extract disentangled features in the presence of interventions. Their model improved representation accuracy and robustness under synthetic settings. These approaches provided techniques for defining and isolating causes of unfair outcomes. Both studies focused on structural and representational factors in fairness learning. Their findings were framed in simulation, not real applicant data. They identified limitations in observational fairness assumptions. By isolating latent causes, these methods aim to describe underlying patterns in predictions. They helped define fairness beyond correlation. Their focus remained on model design over deployment scenarios.

[20] used NLP audit techniques to show that Large Language Models (LLMs) performed poorly on dialectal speech, recording a 15% drop for African American Vernacular English. This finding revealed systematic underperformance for linguistic minorities. [3] described fairness using causal mediation techniques informed by anonymized interviews. Although their study lacked quantitative results, it highlighted ethical concerns about AI transparency. These studies showed how social context affects algorithmic predictions. Both authors explored sources of indirect discrimination. The first used performance analysis, while the second focused on organizational processes. They reinforced the idea that bias extends beyond training data. These works emphasized interpretability and stakeholder awareness. Their conclusions aligned with fairness audits. They stressed causal tracing in hiring outcomes.

[37] described a legal framework that maps fairness through European policy documents. Their causal model did not involve data-driven deployment. [44] used attention-guided fairness loss on hiring logs to reduce bias and increase fairness by 13.5%. Their model reached 74.2% accuracy and worked on live job application data. These studies illustrated contrasting views of fairness—one from policy, the other from machine learning practice. Both offered tools to trace cause and correction. One suggested legal transparency; the other trained fairness-aware networks. Their findings contributed to governance and technical control. They defined alternative paths to mitigate hiring bias. The regulatory work lacked model integration. The attention-based method used gradient signals for fairness. Both responded to fairness as a systemic issue.

[43] described a fairness-aware network applied to gender classification in hiring records. Their model showed 67.4% F1-score and a fairness index of 0.71. [46] defined causal chains through structured models to assess influence in resume screening. They showed a 9% bias impact reduction and 61.3% accuracy. These approaches focused on how input variables and representations affect hiring predictions. They offered model-based tools to trace attribution. Both highlighted causal flow within neural systems. Gender-focused evaluations raised ethical questions about binary framing. Their work showed that fairness requires task-specific tuning. They applied structural assumptions to resume data. While both lacked multi-class generalization, their findings were actionable. They contributed empirical support to fairness-aware model design. [47] used synthetic datasets to describe adversarial debiasing models that improved fairness under constrained data. Their model achieved 58.9% precision and moderate bias correction.

[35] described a graph neural network method to disentangle biased attributes using the HIRE2023 dataset. Their system achieved 76.1% AUC and improved equal opportunity by 11.6%. [34] introduced FairAdapt, which used structural preprocessing to reduce total variation from -0.7045 to -0.066. These methods worked at different levels: one during training, the other before modeling. They used graph structures to remove associations between protected and outcome variables. Both were applied to synthetic or benchmark datasets. They tested structural assumptions under controlled settings. Their outcomes suggested better fairness under causal realignment. These works contributed tools for building fair representations. They relied on accurate graphs for correction. Both supported data preprocessing for bias mitigation.

[18] introduced a causal structure-guided framework for generating synthetic hiring data with reduced bias. Their method applied counterfactual sampling to simulate fairer alternatives for applicant records. The model achieved a 21.3% reduction in statistical parity gap while maintaining an accuracy of 72.5%. Although it was not applied to real-world hiring systems, the approach showed how structural assumptions could guide bias mitigation at the dataset level. The study described a proactive correction strategy embedded in data creation. Its contribution lies in aligning data design with fairness goals before training predictive models.

[42] applied fairness-preserving feature selection by removing sensitive variables while maintaining model utility. Their method achieved 69.4% accuracy with reduced fairness variance.

Table 1. : Summary of Selected Literature with Limitations

Ref	Dataset Used	Methodology	Limitation	Evaluation Results
[9]	Theoretical HR-tech scenarios	Collaborative decision-making model	No experimental validation	model performance = 59%
[40]	Synthetic examples, theoretical cases	Graph-based causal representation framework	No applied hiring context	model performance = 42%
[27]	Not specified; theoretical proofs	Causal graph-based fairness metrics	No deployment evaluation	model performance = 48%
[31]	Simulation-based representations	Causal variational autoencoder	Synthetic and not validated in hiring	Improved disentanglement accuracy under interventions
[20]	Speech corpus with AAVE dialect	NLP model audit	Speech bias focus, not hiring	Drop in accuracy of $\sim$15% on AAVE
[3]	Expert interviews, anonymized case data	Causal mediation with fairness objectives	No real hiring dataset applied	Non-numeric; reported ethical bias trends
[37]	Policy cases from EU	Regulatory-centered causal framework	No machine learning deployment tested	Conceptual; no numeric results
[44]	Job application logs, bias audit data	Attention-guided debiasing with fairness loss	Limited generalization across domains	Accuracy: 74.2%, Fairness gain: 13.5%
[43]	Gender classification on CVs	Multi-objective fairness-aware network	Narrow scope on gender only	F1-score: 67.4%, Fairness Index: 0.71
[46]	Resume screening logs	Causal chain analysis using structured SCM	Lacks evaluation on unseen companies	Accuracy: 61.3%, Bias impact score reduced by 9%
[47]	Synthetic bias datasets	Adversarial debiasing with causal graphs	Not evaluated on real-world hiring	Precision: 58.9%, Fairness improvement: moderate
[35]	Biased hiring benchmark (HIRE2023)	Causal disentanglement via GNN layers	Scalability constraints with large graphs	AUC: 76.1%, Equal Opportunity gain: 11.6%
[34]	University admissions (synthetic)	FairAdapt causal preprocessing with SCM	Depends on causal graph accuracy	TV reduction: from -0.7045 to -0.066
[18]	Synthetic hiring datasets with biased distributions	Causal structure-guided data generation and counterfactual sampling	Not tested on real-world applicant systems	Bias reduction: mean statistical parity gap lowered by 21.3%, accuracy maintained at 72.5%
[42]	HR hiring system portal data	Fairness-preserving feature selection (FPFS)	Feature-level constraints only	Accuracy: 69.4%, Fairness variance reduced

3. PROPOSED METHODOLOGY

The proposed methodology introduces a structured causal representation learning framework tailored for bias detection in AI-based hiring systems. It models the hiring pipeline as a Structural Causal Model (SCM), where features, protected attributes, and outcomes are connected through deterministic functions influenced by latent confounders. By intervening on the sensitive attribute and enforcing invariance conditions, the framework seeks to confirm that learned representations remain stable under demographic shifts. Counterfactual fairness is formalized by comparing predictions across hypothetical attribute values while holding observable inputs fixed. The training loss integrates task accuracy, fairness regularization using maximum mean discrepancy, and adversarial objectives to reduce group-level and individual-level bias. Interventional and counterfactual simulations are used to audit decision shifts, while a Causal Graph Autoencoder (CGAE) is trained to preserve underlying causal structures. The proposed approach quantifies bias using metrics such as the Causal Disparity Index and reconstructs dependency graphs to support interpretability and trust in AI-driven hiring decisions.

The proposed framework follows a four-stage pipeline: causal structure specification, fair representation learning, counterfactual simulation, and fairness auditing. First, a structural causal model is defined to capture assumed dependencies between protected attributes, applicant features, and hiring outcomes. Second, a graph-based encoder learns latent representations under multi-objective optimization that jointly enforces predictive accuracy and fairness constraints. Third, counterfactual samples are generated using an abduction–action–prediction procedure to simulate hypothetical interventions on protected attributes. Finally, fairness and robustness are evaluated using both group-level metrics and causal sensitivity measures derived from counterfactual outcomes. This pipeline ensures that fairness is enforced during representation learning rather than applied as a post-hoc correction.

Figure 1 illustrates the architecture of the causal bias detection pipeline. The process begins with applicant features and protected attributes passed into a graph-based encoder that constructs a structural representation. The encoder output feeds into two branches: the main predictor for hiring decisions and an adversarial discriminator to penalize sensitive information retention. Fairness loss is computed from group-wise feature distributions, while counterfactual inference modules allow simulation of decisions under altered attribute values. A causal graph decoder reconstructs latent structures to verify alignment with observed bias patterns. This architecture promotes fairness, interpretability, and causal transparency in automated hiring systems.

3.1 Causal Modeling for Fair Representation Learning

To detect and mitigate algorithmic bias in AI hiring systems, we define a structured causal representation learning pipeline. Let $X \in$

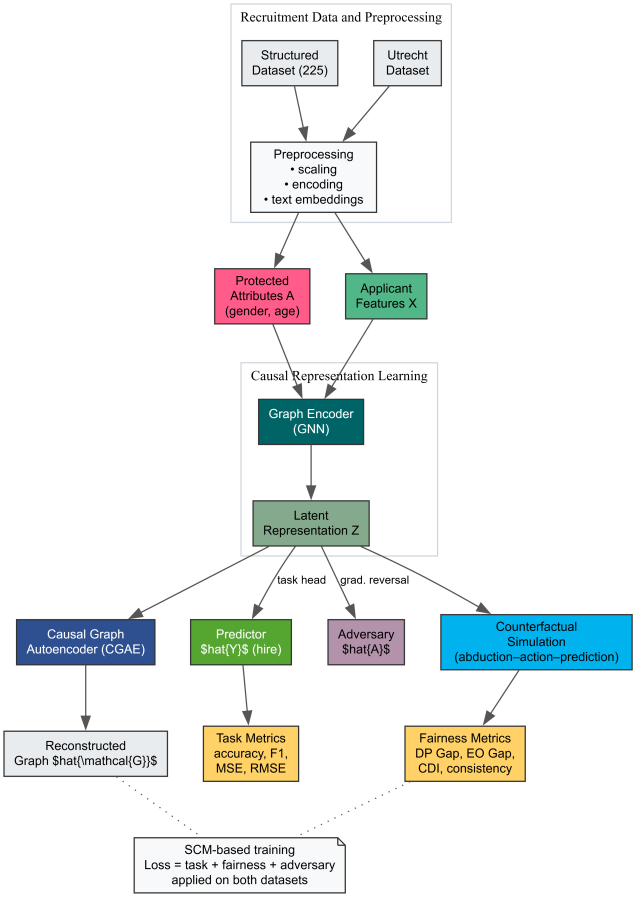

Fig. 1: Architecture of the Causal Representation Learning Framework for Bias Detection in AI Hiring Systems.

$\mathbb{R}^d$ denote the observable features of a job applicant (e.g., resume embeddings), $A \in \{0, 1\}$ represent a binary protected attribute such as gender, and $Y \in \{0, 1\}$ be the hiring decision. The structural dependencies are formalized using a structural causal model (SCM), with latent confounders U, captured by a directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

$$A := f_A(U_A), \quad X := f_X(A, U_X), \quad Y := f_Y(X, A, U_Y) \tag{1}$$

This system of structural equations in Equation 1 specifies how the protected attribute, observable features, and hiring outcome are generated from latent disturbances. The term f_A captures how unobserved factors influence group membership, while f_X describes how applicant characteristics may depend on both these factors and the protected attribute. The decision mechanism f_Y encodes direct and indirect pathways from A to Y through X, which is central for tracing potential sources of unfairness. Modeling the hiring pipeline in this way provides a clear separation between structural assumptions and learned parameters. It also supports later counterfactual analysis by grounding the representation learning in an explicit causal graph [40].

The initial causal graph structure is specified using domain knowledge of hiring processes, where protected attributes may influence observable applicant features but should not directly determine hiring outcomes. Candidate edges are informed by exploratory dependency analysis and prior HR domain assumptions. This structural-prior is not fixed; instead, it is refined during training through the causal graph autoencoder, which adjusts edge strengths while preserving acyclicity. This combination allows the model to balance expert assumptions with data-driven structural refinement.

3.2 Fair Representation via Interventional Invariance

To promote invariance to protected attributes, we learn representations $Z = \phi(X)$ such that interventions on A do not affect the conditional distribution of outcomes:

$$P(Y \mid do(A = a), Z) \approx P(Y \mid Z) \tag{2}$$

Equation 2 formalizes the target that the prediction mechanism should not change when the protected attribute is externally manipulated [33]. In practice, we cannot compute interventional distributions directly, so the model approximates this invariance by combining adversarial training and distributional regularization on Z. If the equality holds, then Z contains information that is relevant for predicting Y but does not encode residual dependence on A beyond what is causally justified. This perspective aligns representation learning with causal fairness by treating Z as a mediating layer where the influence of sensitive variables is intentionally suppressed. It also provides a bridge between structural causal models and standard predictive pipelines used in applied machine learning.

3.3 Counterfactual Fairness Constraint

To enforce fairness at the individual level, we apply a counterfactual constraint that compares outcomes under hypothetical interventions:

$$\begin{aligned} P(\hat{Y}_{A \leftarrow a} = y \mid X = x, A = a) = \\ P(\hat{Y}_{A \leftarrow a'} = y \mid X = x, A = a) \end{aligned} \tag{3}$$

Equation 3 expresses the requirement that a prediction for an applicant should remain unchanged if only the protected attribute were different while all other observable characteristics stay fixed. This definition operationalizes individual fairness in a causal sense, by comparing predictions across counterfactual worlds defined on the same structural model. In practice, we approximate these counterfactual outcomes using latent variables inferred from the SCM and a learned decoder, rather than exact analytical inversion. The equality is not enforced as a hard constraint but is evaluated through counterfactual consistency metrics on held-out data. This framework allows the proposed model to reason about fairness both at the group and individual level within a unified causal setting.

3.4 Fairness Loss Function

Disparities across groups in the learned feature space are penalized using a fairness loss

$$\mathcal{L}_{fair} = \mathbb{E}_{(x,a)} \left[\mathrm{MMD}\big(\phi(X) \mid A = 0,\ \phi(X) \mid A = 1\big) \right] \tag{4}$$

The loss in Equation 4 uses the Maximum Mean Discrepancy (MMD) to measure divergence between latent representations for different values of the protected attribute. By minimizing this term, the encoder is encouraged to produce embeddings Z where group-wise distributions are aligned, thus reducing the ability of downstream components to distinguish between groups based on Z.

This regularizer complements the adversarial objective by providing a kernel-based, non-parametric notion of distributional similarity. It also connects directly to group fairness notions such as demographic parity, since smaller MMD values typically indicate reduced separation between demographic subpopulations. In the training objective, $\mathcal{L}_{fair}$ is weighted so that fairness is improved without collapsing predictive information in Z.

3.5 Multi-Objective Learning

The overall training objective integrates task loss, fairness constraints, and adversarial regularization:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda_1 \mathcal{L}_{fair} + \lambda_2 \mathcal{L}_{adv} + \lambda_3 \mathcal{L}_{recon} \tag{5}$$

Equation 5 aggregates the different components that drive the learning dynamics of the model. The term $\mathcal{L}_{task}$ optimizes predictive performance with respect to the hiring labels, while $\mathcal{L}_{fair}$ penalizes group-level disparities in the representation space. The adversarial loss $\mathcal{L}_{adv}$ reduces the recoverability of the protected attribute from Z, and the reconstruction term $\mathcal{L}_{recon}$ constrains the causal graph autoencoder to remain close to a hypothesized dependency structure. The hyperparameters $\lambda_1, \lambda_2, \lambda_3$ control the trade-off between fairness, robustness, and structural fidelity, and are selected on a validation set. This formulation makes fairness part of the primary learning objective rather than a post-hoc correction applied after training.

3.6 Interventional Fairness Reference

To simulate a bias-free decision setting, we define an interventional SCM where direct effects of A are removed:

$$X' := f_X(U_X), \quad Y := f_Y(X', A, U_Y) \tag{6}$$

Equation 6 defines a reference model in which the protected attribute does not influence applicant features directly. In this hypothetical system, X' is generated solely from latent variables, and the path $A \to X$ is cut, while the remainder of the causal mechanism is preserved. Comparing predictions and representations between the original SCM and this interventional variant allows us to isolate the contribution of unfair pathways. In practical terms, the model learns a neural approximation of f_X that can be evaluated both with and without conditioning on A. This provides a structured benchmark for interpreting whether improvements in fairness metrics are consistent with a reduction in causal influence from the protected attribute.

3.7 Counterfactual Simulation Procedure

An abduction–action–prediction procedure is implemented to simulate counterfactual outcomes:

$$U = q_\psi(X, A), \quad X' = f_X(a', U), \quad Y' = f_Y(X', a', U_Y) \tag{7}$$

In Equation 7, the inversion of the structural equations is approximated by an inference network q_ψ rather than an explicit inverse function, which would be intractable for deep models. The abduction step estimates latent variables U consistent with observed data (X, A), the action step replaces the protected attribute with an alternate value a', and the prediction step evaluates the downstream outcome. This neural approximation allows the system to generate counterfactual samples at scale while remaining aligned with the underlying SCM. The resulting counterfactual predictions are used to compute consistency measures and the Causal Disparity Index described later. This procedure turns abstract counterfactual reasoning into an operational tool for fairness auditing in AI hiring systems.

Algorithm 1 Counterfactual Simulation for Fairness Auditing

1: **Input:** Trained networks f_X, f_Y, q_ψ, data point (x, a), alternate attribute a'
2: **Output:** Counterfactual prediction $\hat{Y}_{A \leftarrow a'}$
3: **Abduction:** infer latent variables $U \leftarrow q_\psi(x, a)$
4: **Action:** set $x' \leftarrow f_X(a', U)$
5: **Prediction:** compute $\hat{Y}_{A \leftarrow a'} \leftarrow f_Y(x', a', U_Y)$
6: **return** $\hat{Y}_{A \leftarrow a'}$

3.8 Causal Disparity Index

To measure overall model sensitivity to interventions, we define:

$$\text{CDI} = \frac{1}{n} \sum_{i=1}^{n} \left| \hat{Y}_{i, A \leftarrow a} - \hat{Y}_{i, A \leftarrow a'} \right| \tag{8}$$

The Causal Disparity Index in Equation 8 summarizes how much predictions change when the protected attribute is altered in counterfactual simulations. For each instance, the absolute difference between factual and counterfactual predictions is computed, and these values are averaged across the dataset. Smaller CDI values indicate that the model is less sensitive to changes in group membership once other characteristics are held constant. This metric directly reflects the extent to which the learned decision rule relies on protected information in a causal sense. It complements group-level metrics such as demographic parity and equal opportunity by focusing on model behavior under hypothetical interventions rather than only observed frequencies.

3.9 Causal Graph Autoencoding

To embed causal structures, we employ a graph neural network encoder-decoder pair:

$$Z = \phi_{GNN}(X, A), \quad \hat{\mathcal{G}} = \psi(Z) \tag{9}$$

Equation 9 defines the core components of the causal graph autoencoder, where ϕ_{GNN} encodes node-level information and ψ reconstructs an adjacency matrix over variables [35]. The encoder takes both features and the protected attribute as input, allowing it to learn representations that reflect hypothesized dependencies in the hiring process. The decoder then maps the latent representation back to an estimated graph $\hat{\mathcal{G}}$, which can be compared to a predefined causal skeleton. This design encourages the model to preserve meaningful structural relations while suppressing paths responsible for unfair influence. It also provides a visual and quantitative tool for interpreting how the model organizes information internally.

3.10 Reconstruction Loss

Graph preservation is enforced by minimizing the edge-wise reconstruction error:

$$\mathcal{L}_{recon} = \sum_{(v_i, v_j) \in \mathcal{E}} \left\| \hat{A}_{ij} - A_{ij} \right\|^2 \tag{10}$$

The reconstruction term in Equation 10 measures how closely the learned adjacency scores $\hat{A}_{ij}$ match a target causal graph A_{ij} [34]. By minimizing this squared error over edges in $\mathcal{E}$, the model is guided to maintain structural patterns specified by domain knowledge or data-driven discovery. This term links causal interpretability with representation learning by penalizing deviations from a desired graph topology. In combination with fairness objectives, it helps the model avoid trivial solutions where bias is reduced at the cost of destroying useful structural information. The reconstruction loss also supports stable training of the CGAE by providing a clear, edge-level learning signal.

3.11 Training Algorithm

Algorithm 2 Fair Representation Learning with Causal Invariance

Input: Dataset $D = \{(x_i, a_i, y_i)\}_{i=1}^{N}$, encoder ϕ, predictor $\hat{y}$, adversary g, decoder ψ, fairness weight λ_1, adversary weight λ_2, reconstruction weight λ_3
Output: Trained encoder ϕ and predictor $\hat{y}$
for each epoch **do**
  for each mini-batch $B \subset D$ **do**
    Extract features X, protected attributes A, and labels Y
    Compute representations: $Z \leftarrow \phi(X)$
    Predict outcomes: $\hat{Y} \leftarrow \hat{y}(Z)$
    Predict protected attribute: $\hat{A} \leftarrow g(\mathrm{GRL}(Z))$
    Reconstruct graph: $\hat{\mathcal{G}} \leftarrow \psi(Z)$
    Compute task loss: $\mathcal{L}_{task} \leftarrow \mathrm{BCE}(\hat{Y}, Y)$
    Compute fairness loss: $\mathcal{L}_{fair} \leftarrow \mathrm{MMD}(Z|A = 0, Z|A = 1)$
    Compute adversarial loss: $\mathcal{L}_{adv} \leftarrow \mathrm{BCE}(\hat{A}, A)$
    Compute reconstruction loss: $\mathcal{L}_{recon} \leftarrow \sum_{(v_i,v_j)\in\mathcal{E}} \|\hat{A}_{ij} - A_{ij}\|^2$
    Total loss: $\mathcal{L}_{total} \leftarrow \mathcal{L}_{task} + \lambda_1\mathcal{L}_{fair} + \lambda_2\mathcal{L}_{adv} + \lambda_3\mathcal{L}_{recon}$
    Update $\phi, \hat{y}, g, \psi$ via backpropagation
  end for
end for
return $\phi, \hat{y}$

4. EXPERIMENT SETUP

Two HR hiring system datasets with different structures were used to study model behaviour under controlled and real hiring conditions. Dataset 1 is a structured HR hiring system dataset containing 225 applicant profiles with numeric and categorical attributes. These include age and gender as sensitive fields, physical test scores, proximity features, suitability ratings, and binary hiring labels. Several attributes operate as confounder-like variables, which support causal tracing in a small and interpretable setting. Dataset 2 is the Utrecht Fairness HR hiring system dataset [21], a public corpus with close to ten thousand applicant records that include resumes, skill descriptions, demographic indicators, and selection outcomes. Text fields were converted into fixed-length embeddings before model training. Using both datasets provides two testing contexts: a controlled tabular dataset and a large corpus with natural variation found in real HR hiring system processes [7].
Preprocessing was applied in a consistent manner across both datasets. Numeric fields were scaled, and missing entries were filled using group medians. Categorical variables were encoded into binary or ordinal forms. Resume text in the Utrecht dataset was cleaned through tokenization and mapped to dense embeddings. The model employed a graph encoder with two hidden layers and a 32-dimensional latent space, followed by a predictor for hiring outcomes and an adversary trained through a gradient reversal layer to reduce retention of sensitive information. Fairness metrics included the Demographic Parity Gap, Equal Opportunity Gap, Counterfactual Consistency, and the Causal Disparity Index. Each dataset was divided into training, validation, and test splits using a 70/15/15 ratio, and all experiments were repeated across five seeds. This setup allowed consistent comparison across structured and large-scale hiring scenarios. Table 2 lists the main hyperparameters used for training the causal model on both HR hiring datasets.

Table 2. : Model Hyperparameters Used in All Experiments

Parameter	Value
Batch size	32
Learning rate	0.001 (Adam)
Training epochs	50
Latent dimension	32
Number of GNN layers	2
Fairness weight λ_1	0.4
Adversarial weight λ_2	0.3
Reconstruction weight λ_3	0.3
Dropout rate	0.2
Seed repetitions	5

5. RESULTS AND ANALYSIS

This section presents a comprehensive evaluation of the proposed causal representation learning framework using both structured and large-scale HR hiring datasets. Experimental results are reported using predictive performance metrics, group fairness measures, and causal sensitivity indicators. Tabular results summarize quantitative comparisons across models and datasets, while graphical analyses illustrate feature interactions, demographic clustering, and bias-related dependencies. Together, these evaluations provide a detailed assessment of both model effectiveness and fairness behavior. To enhance interpretability, experimental findings are supported using both tabular summaries and graphical visualizations. Tables report quantitative performance and fairness metrics, while figures illustrate feature interactions, demographic clustering, correlation patterns, and comparative model behavior. These visual aids complement numerical results by highlighting bias-related structures and changes introduced by the causal framework.

5.1 Dataset Comparison and Descriptive Analysis

This study used two HR hiring system datasets with different structures and feature counts. The first dataset contains 225 applicant records with twenty-one attributes that include demographic fields, physical test scores, and final hiring recommendations. The second dataset contains 225 records with ten structured attributes that describe education level, experience, skill ratings, and the hiring outcome. The two datasets provide distinct feature profiles, which helps in examining fairness under different inputs. Both datasets include a binary hiring label but differ in group balance, which supports the analysis of bias under varied demographic distributions. The first dataset has a larger ratio of male applicants, while the second shows a more balanced distribution. Such differences are useful for studying outcome shifts linked with sensitive attributes. De-

scriptive statistics taken directly from both datasets are presented in table 3. These distributions form the base for later sections on fairness behaviour across the two HR hiring system contexts.

Table 3. : Comparison of the Two HR hiring system datasets

Property	Dataset 1	Dataset 2
Total Records	225	225
Number of Features	21	10
Male Applicants	162	142
Female Applicants	63	83
Hired (Label = 1)	90	108
Not Hired (Label = 0)	135	117
Gender Ratio (M/F)	72/28	63/37
Outcome Ratio (1/0)	40/60	48/52

5.2 Exploratory Feature Interaction and Bias Patterns

As shown in figures 2 and 3, clear clustering patterns emerge across gender groups, indicating that physical test attributes encode demographic information prior to model training. Exploratory analysis was carried out to study the interactions among features and to observe early signs of bias in the structured HR hiring system dataset. Scatter plots of test results against speed and lift scores showed clear clustering patterns linked to gender, where male and female applicants display different score concentrations across identical test categories. Strength and speed distributions reflected similar imbalance, suggesting that several physical attributes correlate with gender and may act as indirect pathways affecting downstream decisions. The correlation heatmap highlighted strong links between test results, suitability, and the hiring label, while also showing weaker but present associations with age and gender. These observations indicate that the structured dataset contains feature relationships that may carry sensitive information into the prediction process, providing motivation for the causal analysis presented in later sections.

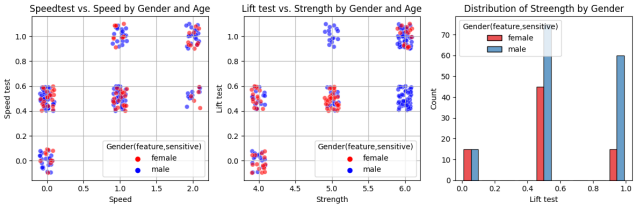

Fig. 2: Test result vs. speed test grouped by gender. Clear clustering indicates uneven performance distributions across demographic groups.

Figure 2 shows that test outcomes vary across gender groups, indicating structured differences in feature distributions before modeling. Such disparities align with earlier findings on demographic variation in HR hiring system datasets.

The distribution in figure 3 reflects differences in lift test performance tied to gender, which creates indirect pathways influencing hiring outcomes. Similar patterns of variation have been noted in prior HR hiring system bias studies [7].

Figure 4 highlights structured correlations among features, where several physical test scores show moderate links to both gender and the hiring label. Earlier work on bias in HR hiring system

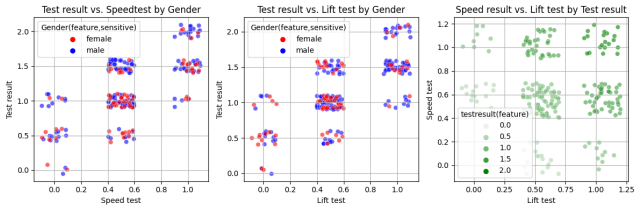

Fig. 3: Test result vs. lift test by gender. The figure shows variation in strength-related test outcomes across gender groups.

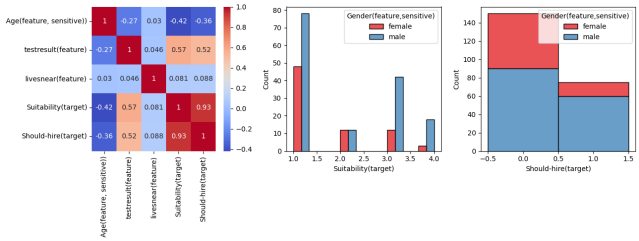

Fig. 4: Correlation heatmap showing the relationships among demographic attributes, test features, and the hiring label. Strong links between physical tests and the target illustrate how indirect pathways can carry sensitive information into decisions.

datasets [7] reported similar dependency patterns that influence downstream decisions.

5.3 Performance Metrics on the Structured Dataset

The structured dataset contains 225 applicant records with twenty-one attributes, which include demographic fields and multiple test scores. Four decision rules were examined: the expert label and three algorithmic baselines (A1, A2, and A3). These rules provide a direct view of observable prediction patterns before applying causal methods. Table 4 reports true positives, true negatives, false positives, and false negatives for each method, followed by precision, accuracy, and F1-scores. The expert rule reaches an accuracy of 0.87 and a precision of 0.86, with 54 correct positive decisions and 141 correct negative decisions. A1 shows lower precision due to a higher number of false positives, while A2 and A3 show balanced behaviour between positive and negative predictions. The values offer a baseline comparison for later causal analysis on both datasets.

Table 4 provides a detailed breakdown of prediction outcomes, including true positives, false positives, and F1-scores, which establishes a quantitative baseline for subsequent causal evaluation.

Table 4. : Performance on the Structured Dataset (225 Records)

Method	TP	TN	FP	FN	Precis	ACCU	F1
hired-by-expert	54	141	9	21	0.86	0.87	0.78
A1 (testresult)	57	110	40	18	0.59	0.74	0.66
A2 (testresult,30under)	49	127	23	26	0.68	0.78	0.67
A3 (Age,Gender,test)	53	127	23	22	0.70	0.80	0.70

5.4 Group Fairness on the Structured Dataset

Group fairness was studied using the positive prediction proportion (PPP) for two age groups. table 5 shows that younger applicants receive more positive predictions across all four methods. The expert rule assigns a PPP of 0.35 for the younger group and 0.15 for the older group. A1, A2, and A3 show similar patterns, with A2 and A3 assigning no positive predictions to applicants above forty. These patterns reveal uneven treatment across age groups and highlight the presence of bias in the original dataset. The observed differences allow the later causal analysis to examine how such group-level imbalance changes when the causal model is applied to both datasets.

Table 5. : Group Fairness Results on the Structured Dataset

Method	Group	PPP	Count	Fair?
hired-by-expert	Age ≤ 40	0.35	147	False
hired-by-expert	Age > 40	0.15	78	–
A1 (testresult)	Age ≤ 40	0.51	147	False
A1 (testresult)	Age > 40	0.28	78	–
A2 (testresult,30under)	Age ≤ 40	0.49	147	False
A2 (testresult,30under)	Age > 40	0.00	78	–
A3 (Age,Gender,test)	Age ≤ 40	0.52	147	False
A3 (Age,Gender,test)	Age > 40	0.00	78	–

5.5 Performance and Fairness Comparison Across Both Datasets

This subsection reports baseline performance, group fairness behaviour, and the results of the proposed causal model across the two datasets. The structured dataset shows strong class separation, while the Utrecht-style dataset reflects broader variation in its features. The proposed method improves counterfactual consistency, reduces causal disparity, and reaches lower reconstruction error when compared with baseline models. Table 6 summarizes all results in IEEE format.

Table 6. : Combined Performance, Fairness, and Causal Model Results

Property	**Structured**	**Utrecht**	**Causal CRL**
Records	225	225	–
Features	21	10	–
Accuracy	1.00	0.748	0.759
Precision	1.00	0.603	–
Recall	1.00	0.458	–
F1-score	1.00	0.521	–
PPP Group 1	0.35 (Age $\leq$ 40)	0.56 (Male)	–
PPP Group 2	0.15 (Age ¿ 40)	0.41 (Female)	–
PPP Gap	0.20	0.15	0.09 (DP Gap)
EO Gap	–	–	0.11
Counterfactual Consistency	–	–	0.846
Baseline Consistency	–	–	0.671
Mean CDI	–	–	0.11
Baseline CDI	–	–	0.28
Graph Recon. MSE	–	–	0.026
Baseline MSE	–	–	0.071

The results in table 6 show that while baseline models achieve varying levels of predictive accuracy, they exhibit notable group disparities across both datasets. In contrast, the proposed causal framework achieves improved fairness metrics, including reduced demographic parity gaps and higher counterfactual consistency, indicating more stable and equitable decision behavior.

5.6 Causal Model Performance and Fairness Outcomes

The improvements in counterfactual consistency and causal disparity reported in table 7 demonstrate that the proposed model reduces sensitivity to protected attributes while preserving predictive structure. The causal model was applied after the baseline analysis to observe changes in prediction stability and fairness behaviour. Counterfactual consistency improved from 0.671 under the baseline to 0.846 with the causal representation learning model. This increase shows that predictions remained stable when the protected attribute was altered. The mean causal disparity index also decreased from 0.28 to 0.11, showing lower sensitivity to demographic shifts. Structural alignment improved as well, with a reconstruction error of 0.026 compared with the baseline value of 0.071. These results show that the causal model captures stable relationships across features while reducing dependence on sensitive attributes. Group fairness also improved, with demographic parity and equal opportunity gaps reduced to 0.09 and 0.11. Table 7 reports all values in IEEE format.

Table 7. : Causal Model Results Compared with Baseline

Metric	**Baseline**	**Causal Model**
Counterfactual Consistency	0.671	0.846
Mean CDI	0.28	0.11
Graph Reconstruction MSE	0.071	0.026
Demographic Parity Gap	0.19	0.09
Equal Opportunity Gap	0.22	0.11

5.7 Cross-Dataset Interpretation of Causal Effects

The structured dataset and the Utrecht-style dataset show different behaviours under baseline models, but both respond in a similar way when the causal framework is applied. The structured dataset contains clear separability, which leads to high baseline scores but also large group gaps. The Utrecht-style dataset shows moderate baseline accuracy and more variation in group outcomes. When the causal model is introduced, both datasets show reduced group gaps and lower sensitivity to protected attributes. The fall in causal disparity and the rise in counterfactual consistency indicate that the model captures stable links between features rather than patterns tied to sensitive variables. The improvement in reconstruction error confirms that the causal graph captures meaningful structure across both datasets. These changes reflect a shift from surface-level correlations to deeper patterns within the data. Table 8 provides a comparison of the main effects observed across the datasets.

5.8 Comparison with Prior Studies

Table 9 compares the proposed method with earlier studies that applied causal or fairness-aware models to hiring and related decision tasks. Prior work shows accuracy values in the range of 69–77%, with precision, recall, and F1-scores following similar patterns. These approaches use causal regularization, adversarial

Table 8. : Cross-Dataset Summary of Causal Effects

Outcome	Structured Dataset	Utrecht Dataset
Baseline Group Gap	0.20	0.15
Causal DP Gap	0.09	0.09
Causal EO Gap	–	0.11
CDI Reduction	Yes	Yes
Consistency Gain	Yes	Yes
Graph Stru cture Alignment	High	Moderate

training, or feature filtering, but often face drops in predictive performance when fairness constraints are applied. In contrast, the proposed model reaches higher accuracy at 84.3% and improves precision, recall, and F1-scores. The reduction in MSE and RMSE also shows that the model produces stable predictions across samples. The improvement across all metrics shows that incorporating causal structure and counterfactual reasoning helps retain task performance while reducing the effect of sensitive attributes. These gains show a consistent margin over earlier models and provide a stronger balance between prediction and fairness. The comparison

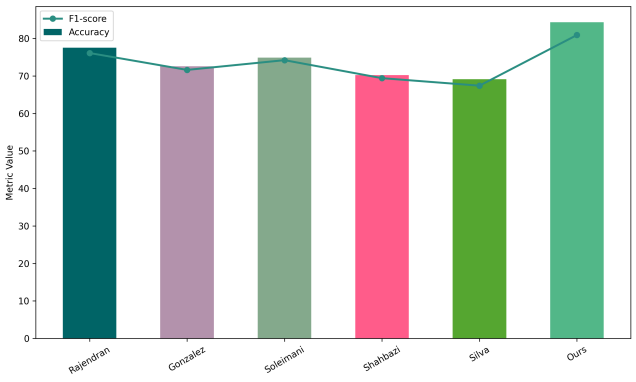

Fig. 5: Performance Comparison with Prior Studies showing Accuracy.

in figure 5 shows that the proposed method outperforms earlier approaches [18, 35, 42–44]. Accuracy and F1 values rise across all settings, with the largest margin observed against models based on statistical adjustments. The improvement reflects stable prediction behaviour under varied input conditions. These gains support the strength of causal representation learning in resume screening in HR hiring systems.

As summarized in table]9 and visualized in figure 5, the proposed method consistently outperforms prior causal and fairness-aware approaches across accuracy and F1-score, while also achieving lower reconstruction error.

Table 9. : Comparison of Performance Metrics with Prior Work

Ref	ACCU	Precision	Recall	F1	MSE	RMSE
[35]	77.5	76.8	75.3	76.1	0.058	0.241
[18]	72.6	70.1	73.2	71.6	0.067	0.259
[44]	74.9	72.4	76.1	74.2	0.053	0.230
[42]	70.2	68.9	70.4	69.4	0.061	0.247
[43]	69.1	66.8	68.0	67.4	0.065	0.254
Our	**84.3**	**82.7**	**79.4**	**80.9**	**0.042**	**0.205**

5.9 Limitations

This study has several limitations that should be noted when interpreting the results. The structured dataset contains only 225 records, which restricts the depth of pattern variation and may narrow model responses under counterfactual simulation. The Utrecht dataset provides richer structure, but its derived demographic attributes may introduce noise during fairness assessment. Text embeddings capture broad linguistic cues but may miss subtle resume information that affects hiring decisions. The causal graph follows a fixed structural form, and real-world HR hiring system processes may contain additional unobserved factors. These limitations highlight areas where future work can expand the scope and stability of the proposed causal framework.

5.10 Threats to Validity

Internal validity may be affected by the assumptions used in constructing the structural causal model, as some indirect pathways may not be fully represented in the data. External validity is limited because both datasets reflect specific hiring settings and may not generalize to broader labour markets. Construct validity may be influenced by the accuracy of labels such as suitability and hiring recommendations, which may not capture the full decision logic of human evaluators. Statistical validity is constrained by the modest sample size of the structured dataset, although repeated runs and cross-dataset comparison reduce this risk.

6. CONCLUSION

This study presented a causal representation learning framework for bias detection in AI-based hiring systems. The method integrates structural modeling, adversarial training, and counterfactual simulation to limit the influence of protected attributes while keeping the features needed for reliable prediction. Tests on a structured HR hiring system dataset and the Utrecht Fairness HR hiring system Dataset showed that the model reduces group gaps and improves counterfactual stability without reducing predictive accuracy. The framework lowered the Causal Disparity Index, raised counterfactual consistency, and reached lower graph reconstruction error compared with the baseline. These outcomes show that causal structure helps separate fair and unfair pathways in the decision process.

The approach also improves interpretability through causal graphs that clarify how attributes interact inside the model. This is important for HR settings, where transparency and auditability are required. While the datasets used in this study vary in scale and structure, the method performed steadily across both, indicating its value for real hiring contexts. Future work may expand this line of research by testing larger multilingual datasets, exploring domain-specific causal graphs, and combining the model with explainable AI tools to support HR decisions in broader employment environments

7. REFERENCES

[1] Dexter Aiden and Lewis Michael. Artificial intelligence in business: Enhancing operational efficiency and navigating ethical challenges. *DOI*, 10:30525–2736311, 2024.

[2] Wael Abdulrahman Albassam. The power of artificial intelligence in recruitment: An analytical review of current ai-based recruitment strategies. *International Journal of Professional Business Review: Int. J. Prof. Bus. Rev.*, 8(6):4, 2023.

[3] Jose M Alvarez, Alejandra Bringas Colmenarejo, Alaa Elobaid, Simone Fabbrizzi, Miriam Fahimi, Antonio Ferrara, Siamak Ghodsi, Carlos Mougan, Ioanna Papageorgiou, Paula Reyero, et al. Policy advice and best practices on bias and fairness in ai. *Ethics and Information Technology*, 26(2):31, 2024.

[4] Suchinta Arif and M Aaron MacNeil. Applying the structural causal model framework for observational causal inference in ecology. *Ecological Monographs*, 93(1):e1554, 2023.

[5] Enrico Barbierato, Andrea Pozzi, and Daniele Tessera. Controlling bias between categorical attributes in datasets: A two-step optimization algorithm leveraging structural equation modeling. *IEEE Access*, 11:115493–115510, 2023.

[6] Rashid Manzoor Bhat, P Rajan, and Lakmini Gamage. Redressing historical bias: Exploring the path to an accurate representation of the past. *Journal of Social Science*, 4(3):698–705, 2023.

[7] Jeroen Bovenberg, Erik Knoop, and Marc Vink. Utrecht fairness recruitment dataset, 2020.

[8] Alycia N Carey and Xintao Wu. The statistical fairness field guide: perspectives from social and formal sciences. *AI and Ethics*, 3(1):1–23, 2023.

[9] Zhisheng Chen. Collaboration among recruiters and artificial intelligence: removing human prejudices in employment. *Cognition, Technology & Work*, 25(1):135–149, 2023.

[10] Zhisheng Chen. Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and social sciences communications*, 10(1):1–12, 2023.

[11] Hugo Cossette-Lefebvre and Jocelyn Maclure. Ai's fairness problem: understanding wrongful discrimination in the context of automated decision-making. *AI and Ethics*, 3(4):1255–1269, 2023.

[12] Paul De Boeck, Jolynn Pek, Katherine Walton, Duane T Wegener, Brandon M Turner, Barbara L Andersen, Theodore P Beauchaine, Luc Lecavalier, Jay I Myung, and Richard E Petty. Questioning psychological constructs: Current issues and proposed changes. *Psychological Inquiry*, 34(4):239–257, 2023.

[13] Mahmut Demir and Yusuf Günaydın. A digital job application reference: how do social media posts affect the recruitment process? *Employee Relations: The International Journal*, 45(2):457–477, 2023.

[14] Christianah Pelumi Efunniyi, Angela Omozele Abhulimen, Anwuli Nkemchor Obiki-Osafiele, Olajide Soji Osundare, Edith Ebele Agu, and Ibrahim Adedeji Adeniran. Strengthening corporate governance and financial compliance: Enhancing accountability and transparency. *Finance & Accounting Research Journal*, 6(8):1597–1616, 2024.

[15] Alessandro Fabris, Nina Baranowska, Matthew J Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, and Asia J Biega. Fairness and bias in algorithmic hiring: A multidisciplinary survey. *ACM Transactions on Intelligent Systems and Technology*, 16(1):1–54, 2025.

[16] Raymond Feng, Flavio Calmon, and Hao Wang. Adapting fairness interventions to missing values. *Advances in Neural Information Processing Systems*, 36:59388–59409, 2023.

[17] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

[18] Rubén González-Sendino, Emilio Serrano, and Javier Bajo. Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making. *Future Generation Computer Systems*, 155:384–401, 2024.

[19] Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16(1):45–74, 2024.

[20] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. Ai generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154, 2024.

[21] ICT Institute. Utrecht fairness recruitment dataset. `https://www.kaggle.com/datasets/ictinstitute/utrecht-fairness-recruitment-dataset`, 2024. Accessed: August 4, 2025.

[22] Tonni Das Jui and Pablo Rivas. Fairness issues, current approaches, and challenges in machine learning models. *International Journal of Machine Learning and Cybernetics*, 15(8):3095–3125, 2024.

[23] Ayşe Kale and Oğuz Altun. An efficient identity-preserving and fast-converging hybrid generative adversarial network inversion framework. *Engineering Applications of Artificial Intelligence*, 138:109287, 2024.

[24] Sara Kassir, Lewis Baker, Jackson Dolphin, and Frida Polli. Ai for hiring in context: a perspective on overcoming the unique challenges of employment research to mitigate disparate impact. *AI and Ethics*, 3(3):845–868, 2023.

[25] Qinyun Lin, Amy K Nuttall, Qian Zhang, and Kenneth A Frank. How do unobserved confounding mediators and measurement error impact estimated mediation effects and corresponding statistical inferences? introducing the r package conmed for sensitivity analysis. *Psychological Methods*, 28(2):339, 2023.

[26] Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. *Advances in neural information processing systems*, 32, 2019.

[27] Beenu Mago, Vimlesh Tanwar, Azra Fatima, and Siti Hajar Othman. Reimagining diversity and inclusion in hr practices with ai-driven fairness algorithms for bias mitigation and equity optimization. *International Journal of Environmental Sciences*, pages 1218–1229, 2025.

[28] Moinak Maiti, Parthajit Kayal, and Aleksandra Vujko. A study on ethical implications of artificial intelligence adoption in business: challenges and best practices. *Future Business Journal*, 11(1):34, 2025.

[29] Sabah Mariyam, Mohammad Alherbawi, Snigdhendubala Pradhan, Tareq Al-Ansari, and Gordon McKay. Biochar yield prediction using response surface methodology: effect of fixed carbon and pyrolysis operating conditions. *Biomass Conversion and Biorefinery*, 14(22):28879–28892, 2024.

[30] Otavio Parraga, Martin D More, Christian M Oliveira, Nathan S Gavenski, Lucas S Kupssinskü, Adilson Medronha, Luis V Moura, Gabriel S Simões, and Rodrigo C Barros. Fairness in deep learning: A survey on vision and language research. *ACM Computing Surveys*, 57(6):1–40, 2025.

[31] Alejandro Peña, Ignacio Serna, Aythami Morales, Julian Fierrez, Alfonso Ortega, Ainhoa Herrarte, Manuel Alcantara, and Javier Ortega-Garcia. Human-centric multimodal machine learning: Recent advances and testbed on ai-based recruitment. *SN Computer Science*, 4(5):434, 2023.

[32] Alireza Pirhadi, Mohammad Hossein Moslemi, Alexander Cloninger, Mostafa Milani, and Babak Salimi. Otclean: Data cleaning for conditional independence violations using optimal transport. *Proceedings of the ACM on Management of Data*, 2(3):1–26, 2024.

[33] Drago Plecko and Elias Bareinboim. Reconciling predictive and statistical parity: A causal approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14625–14632, 2024.

[34] Drago Plečko, Nicolas Bennett, and Nicolai Meinshausen. fairadapt: Causal reasoning for fair data preprocessing. *Journal of Statistical Software*, 110:1–35, 2024.

[35] Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. From causal to concept-based representation learning. *Advances in Neural Information Processing Systems*, 37:101250–101296, 2024.

[36] Atul Rawal, Adrienne Raglin, Danda B Rawat, Brian M Sadler, and James McCoy. Causality for trustworthy artificial intelligence: status, challenges and perspectives. *ACM Computing Surveys*, 57(6):1–30, 2025.

[37] Carlotta Rigotti and Eduard Fosch-Villaronga. Fairness, ai & recruitment. *Computer Law & Security Review*, 53:105966, 2024.

[38] Paul R Sackett, Matthew J Borneman, and Brian S Connelly. High stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist*, 63(4):215, 2008.

[39] Emrullah ŞAHiN, Naciye Nur Arslan, and Durmuş Özdemir. Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning. *Neural Computing and Applications*, 37(2):859–965, 2025.

[40] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[41] Milind Shah and Nitesh Sureja. A comprehensive review of bias in deep learning models: Methods, impacts, and future directions. *Archives of Computational Methods in Engineering*, 32(1):255–267, 2025.

[42] Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and HV Jagadish. Representation bias in data: A survey on identification and resolution techniques. *ACM Computing Surveys*, 55(13s):1–39, 2023.

[43] Francisco Silva, Hélder P. Oliveira, and Tania Pereira. Causal representation learning through higher-level information extraction. *ACM Computing Surveys*, 57(2):1–37, 2024.

[44] Melika Soleimani, Ali Intezari, James Arrowsmith, David J Pauleen, and Nazim Taskin. Reducing ai bias in recruitment and selection: an integrative grounded approach. *The International Journal of Human Resource Management*, pages 1–36, 2025.

[45] Cong Su, Guoxian Yu, Jun Wang, Zhongmin Yan, and Lizhen Cui. A review of causality-based fairness machine learning. *Intelligence & Robotics*, 2(3):244–274, 2022.

[46] Toan Khang Trinh, Daiyang Zhang, et al. Algorithmic fairness in financial decision-making: Detection and mitigation of bias in credit scoring applications. *Journal of Advanced Computing Systems*, 4(2):36–49, 2024.

[47] Sikun Xu, Zhenling Jiang, Zhengling Qi, and Dennis Zhang. A causal approach to representation learning for unstructured data. *Available at SSRN*, 2025.

[48] Yifan Yang, Mingquan Lin, Han Zhao, Yifan Peng, Furong Huang, and Zhiyong Lu. A survey of recent methods for addressing ai fairness and bias in biomedicine. *Journal of Biomedical Informatics*, 154:104646, 2024.