C2BM: Causal Concept Disentanglement for Fair Multimodal COVID-19 Detection

Letu Qingge[1]*, Hailemicael Lulseged Yimer[1], Maxwell Sam[1], Richard Annan[1], Robert Newman[2], Hong Qin[3]

[1]Department of Computer Science, North Carolina A&T State University
1601 East Market Street
Greensboro, NC 27411, USA
[2]Department of Biology, North Carolina A&T State University
1601 East Market Street
Greensboro, NC 27411, USA
[3]School of Data Science, Department of Computer Science, Old Dominion University
Norfolk, VA, USA
lqingge@ncat.edu, {hlyimer, msam, rkannan}@aggies.ncat.edu, rhnewman@ncat.edu, hqin@odu.edu

Abstract

Algorithmic bias in COVID-19 detection systems poses a serious threat to equitable pandemic response, as demographic disparities in model performance risk worsening health outcomes across vulnerable populations. We present an adopted Causal Concept Bottleneck Model (C2BM) framework that systematically addresses fairness in multimodal COVID-19 detection by learning interpretable concepts from chest CT scans and patient metadata. Our approach targets the Country→Institution→COVID causal pathway through principled interventions, achieving substantial bias reduction: age and gender demographic parity differences decrease from 51.15% to 18.50% (64% reduction), gender disparate impact improves from 0.6475 to 0.9812 (51% improvement), while preserving 98.45% diagnostic F1-score. Through comprehensive evaluation across four model variants, we demonstrate that causal interventions enable stable and reproducible fairness improvements without compromising clinical utility. Our work establishes that principled causal reasoning can achieve practical fairness-accuracy trade-offs in COVID-19 detection systems, providing actionable guidance for equitable healthcare AI deployment.

Code — https://github.com/LetuQingge/Causal-Concept-Disentanglement-for-Fair-Multimodal-COVID-19-Detection

Datasets — https://github.com/LetuQingge/C2BM-Causal-Concept-Disentanglement-for-Fair-Multimodal-COVID-19-Detection-Dataset

*Corresponding authors: lqingge@ncat.edu, hyimer@aggies.ncat.edu, msam@aggies.ncat.edu. Letu Qingge, Hailemicael Lulseged Yimer and Maxwell Sam have equal contributions to this manuscript.

Introduction

The rapid deployment of AI systems for COVID-19 detection has revealed concerning patterns of demographic bias that can exacerbate existing healthcare disparities (Obermeyer et al. 2019; Larrazabal et al. 2020). Such biases often emerge when COVID-19 detection models are developed using datasets with imbalanced demographic representation, whether by age, gender, ethnicity, or socioeconomic background. When certain subgroups are underrepresented in the training data, the resulting models may learn features that disproportionately reflect the dominant group, leading to skewed decision boundaries and reduced accuracy for minority populations (Seyyed-Kalantari et al. 2021).

This challenge is particularly acute in multimodal COVID-19 detection systems, where models combine chest CT imaging with patient demographics and institutional metadata. While demographic information can provide valuable clinical context, it also introduces avenues for bias that are difficult to detect and mitigate using traditional fairness approaches (Chen et al. 2019; Glocker et al. 2023).

Recent work in algorithmic fairness has explored various bias mitigation strategies, but these methods often treat bias as a statistical property rather than addressing its underlying causal mechanisms (Kusner et al. 2017; Kilbertus et al. 2017). Causal approaches to fairness offer a principled framework for understanding and addressing bias by explicitly modeling the causal relationships between protected attributes, features, and outcomes (Pearl 2009; Zhang, Lemoine, and Mitchell 2018).

In this work, we present an adopted Causal Concept Bottleneck Model framework inspired by the idea Causal Concept Bottleneck Model (De Felice et al. 2025) to address fairness challenges in COVID-19 data through targeted country–institution interventions. The framework is grounded in the idea that geographic and institutional factors can create hidden pathways of bias in this case, along

the Country → Institution → COVID decision chain. By intervening directly on these causal links, our approach aims to ensure that model predictions are not unduly influenced by geographic or institutional imbalances in the data.

Our contributions are fourfold. First, we propose a causal intervention framework that specifically targets the Country → Institution → COVID pathway to reduce systematic bias. Second, we conduct a comprehensive fairness evaluation, demonstrating that the approach significantly reduces demographic disparities across both age and gender groups. Third, we provide practical deployment insights for clinical AI systems operating across diverse geographic regions, where fairness is critical for trust and adoption. Finally, we show that these gains in fairness do not come at the expense of diagnostic performance.

Our strategy achieves substantial bias reduction while maintaining clinical utility. We reduced the age and gender demographic parity difference by 64%, which achieved a near-perfect gender disparate impact score of 0.98, and consistently maintained a diagnostic F1-score of 98.45%. This demonstrates that substantial bias mitigation can be achieved without sacrificing clinical accuracy which is an essential requirement for real-world COVID-19 deployment.

Related Work

The rapid evolution of multimodal artificial intelligence (AI) in medicine has led to significant advances in diagnostic accuracy, particularly for complex diseases like COVID-19. Recent reviews highlight that integrating imaging data with clinical metadata and using frameworks such as transformers and graph neural networks has become a leading strategy for robust clinical decision-making. These architectures enable the fusion of heterogeneous data sources, improving the translation of AI models to real-world clinical settings and outperforming unimodal approaches (Simon et al. 2025).

Despite these advances, bias and fairness remain central concerns in COVID-19. Systematic reviews and international collaborations have documented that AI models in medical imaging are susceptible to various forms of bias, including those arising from demographic imbalances, data collection, and model deployment. These biases can compromise patient outcomes and perpetuate health disparities if not proactively addressed (Koçak et al. 2025). (Xu et al. 2024) further emphasize that deep learning models often exhibit performance disparities across subgroups, such as age, sex, and ethnicity, and call for systematic fairness evaluation and mitigation strategies in medical image analysis.

A growing body of research is now focused on algorithmic approaches to fairness, including causal disentanglement and representation learning. Recent surveys categorize these methods into data augmentation, adversarial learning, disentangled representation learning, and causality-based approaches, all aimed at reducing bias and improving generalizability in biomedical AI (Yang et al. 2024). Causal machine learning, in particular, has gained traction for its ability to model and control for confounding variables, enabling more robust and interpretable predictions in healthcare (Sanchez et al. 2022).

In the context of COVID-19, explainable and fair AI models are especially critical. (Sun, Akman, and Schuller 2025) introduced CapsCovNet, a modified capsule network for COVID-19 diagnosis from multimodal medical imaging, demonstrating the value of integrating multiple data types for improved accuracy and interpretability. Meanwhile, (Schouten et al. 2025) developed a deep learning system that predicts COVID-19 outcomes using multimodal data but also highlighted the scarcity of external validation and the need for fairness-aware model evaluation.

Recent literature and comprehensive reviews further underscore the paradigm shift toward multimodal AI in healthcare, noting both the technical challenges and the opportunities for improving health intelligence (Bhambhoria et al. 2023). (Simon et al. 2025; Xu et al. 2024) both point out that while multimodal models generally outperform unimodal ones, issues such as data scarcity, inconsistent taxonomy, and lack of fairness evaluation persist.

Finally, (Mukherjee and Summers 2024) highlighted the importance of explicitly modeling and controlling for disease severity and other confounders to achieve fairness in medical imaging AI. Algorithmic fairness in medicine is increasingly being addressed through causal representation learning, which separates disease relevant features from confounding and sensitive factors, thereby improving both robustness and equity.

Fairness in Medical AI

Bias in COVID-19 data has been extensively documented across applications from diagnostic imaging (Larrazabal et al. 2020) to clinical decision support (Obermeyer et al. 2019). Gender bias in chest CT scans interpretation (DeGrave, Janizek, and Lee 2021) and racial bias in clinical risk prediction (Obermeyer et al. 2019) demonstrate the pervasive nature of these issues.

Recent mitigation efforts include adversarial training for fair medical image analysis (Glocker et al. 2023) and fairness constraints in cardiac image segmentation (Puyol-Antón et al. 2021). However, these approaches primarily focus on single-modal data and do not address the complex causal relationships present in multimodal clinical datasets.

Causal Fairness

Causal approaches to fairness have gained attention for their principled treatment of bias (Kusner et al. 2017). Key frameworks include counterfactual fairness (Kusner et al. 2017), path-specific fairness (Nabi and Shpitser 2018), and individual fairness through causal modeling (Russell et al. 2017).

While (Zhang, Lemoine, and Mitchell 2018) and (Kilbertus et al. 2017) have developed causal frameworks for machine learning, their application to multimodal medical data remains limited. Our work bridges this gap by developing causal intervention strategies specifically for COVID-19.

Subgroup Fairness

Subgroup fairness extends classical fairness by requiring equitable performance not only across broad protected groups but also within finer intersections of features. Standard parity metrics like demographic parity or equalized odds can

miss harms to smaller subpopulations, a problem known as fairness gerrymandering (DeGrave, Janizek, and Lee 2021; Obermeyer et al. 2019). By auditing disparities across overlapping subgroups, subgroup fairness ensures that "minorities within minorities" are not disproportionately penalized, an especially critical concern in healthcare imaging where systematic under- or over-prediction can translate into unequal diagnostic accuracy and treatment access. Figure 1 shows PPRs by Country $\times$ Age Bin. Countries such as China, Portugal, and Turkey have near-perfect positive predictions (PPR ≈ 1.0), while Russia remains below 0.2 across all ages. Iran displays an age gradient (0.53 to 0.90). These disparities signal fairness issues: cross-country gaps violate demographic parity and age-based variation within Iran suggests unequal error rates, a violation of equalized odds.

Concept Bottleneck Models

Concept Bottleneck Models (CBMs) enhance interpretability by routing predictions through human-understandable concepts (Koh et al. 2020). Recent advances include causally reliable CBMs (Mahinpei et al. 2021) and hybrid approaches that fuse multimodal data (Yang et al. 2022). Our work extends CBMs with causal graph theory for fairness applications.

Methodology

Problem Formulation

Let $\mathbf{X}_I \in \mathbb{R}^{224\times224\times3}$ represent chest CT scans and $\mathbf{C} \in \mathbb{R}^5$ represent the concept vector containing normalized age, gender encoding, country encoding, institution encoding, and COVID-19 status. Our goal is to learn a predictor that is both accurate and fair with respect to protected attributes.

We model the causal relationships between concepts using a directed acyclic graph (DAG) where:

$$\text{Age} \rightarrow \text{COVID} \tag{1}$$

$$\text{Gender} \rightarrow \text{COVID} \tag{2}$$

$$\text{Country} \rightarrow \text{Institution} \rightarrow \text{COVID} \tag{3}$$

This structure captures the clinical reality that age and gender directly influence COVID-19 risk, while geographic location affects institutional assignment, which in turn may introduce systematic bias in diagnosis.

Data Preprocessing and Normalization

Our preprocessing pipeline ensures consistent representation across all concepts:

Categorical Encoding: Country and institution attributes are encoded and normalized by their maximum values to ensure $[0, 1]$ range:

$$c_{\text{country}} = \frac{\text{Country\_encoded}}{\max(\text{Country\_encoded})} \tag{4}$$

$$c_{\text{institution}} = \frac{\text{Institution\_encoded}}{\max(\text{Institution\_encoded})} \tag{5}$$

Image Processing: CT-Scan images undergo standard preprocessing to ensure consistent representation. Images are first resized to 224×224 pixels and center-cropped to maintain a consistent field of view across all samples. The pixel values are then normalized using ImageNet statistics with mean $\mu = [0.485, 0.456, 0.406]$ and standard deviation $\sigma = [0.229, 0.224, 0.225]$ to leverage pre-trained feature representations.

Concept Clipping: All concept values are clipped to the $[0, 1]$ range to prevent gradient instabilities during training and ensure consistent concept representation across the causal graph.

Causal Concept Bottleneck Model Architecture

Our C2BM implementation follows a structured approach based on the causal graph topology, as illustrated in Figure 2. The causal structure is defined as:

Our C2BM implementation follows a structured approach based on the causal graph topology. The causal structure is defined in Table 1 as:

Concept	Index	Parents
Age	0	None (source)
Gender	1	None (source)
Country	2	None (source)
Institution	3	Country (2)
COVID	4	Age (0), Gender (1), Institution (3)

Table 1: Causal graph structure with parent relationships. Institution intervention at $c_3 = 0.5$ blocks the Country→Institution→COVID pathway.

Feature Extraction: ResNet18 backbone extracts 512-dimensional image features:

$$\mathbf{h} = \text{ResNet18}(\mathbf{X}_I) \in \mathbb{R}^{512} \tag{6}$$

Concept Encoding: Features are projected to concept-specific representations:

$$\mathbf{U} = \text{Linear}_{512\rightarrow320}(\mathbf{h}) \in \mathbb{R}^{5\times64} \tag{7}$$

where each concept receives a 64-dimensional encoding $\mathbf{u}_i \in \mathbb{R}^{64}$.

Causal Prediction: Concepts are predicted according to topological ordering:

For source concepts (Age, Gender, Country):

$$c_i = \sigma(\text{MLP}_i(\mathbf{u}_i)) \tag{8}$$

For Institution (with Country as parent):

$$c_3 = \sigma(\boldsymbol{\theta}_3 \cdot c_2), \quad \boldsymbol{\theta}_3 = \text{MLP}_3(\mathbf{u}_3) \tag{9}$$

For COVID (with Age, Gender, Institution as parents):

$$c_4 = \sigma(\boldsymbol{\theta}_4^T[c_0, c_1, c_3]), \quad \boldsymbol{\theta}_4 = \text{MLP}_4(\mathbf{u}_4) \tag{10}$$

Meta-Network Architecture: Each meta-network follows a consistent architectural design optimized for concept prediction. The networks receive 64-dimensional concept encodings as input and process them through a hidden layer containing 32 neurons with ReLU activation. To prevent overfitting, we apply dropout regularization with a rate of

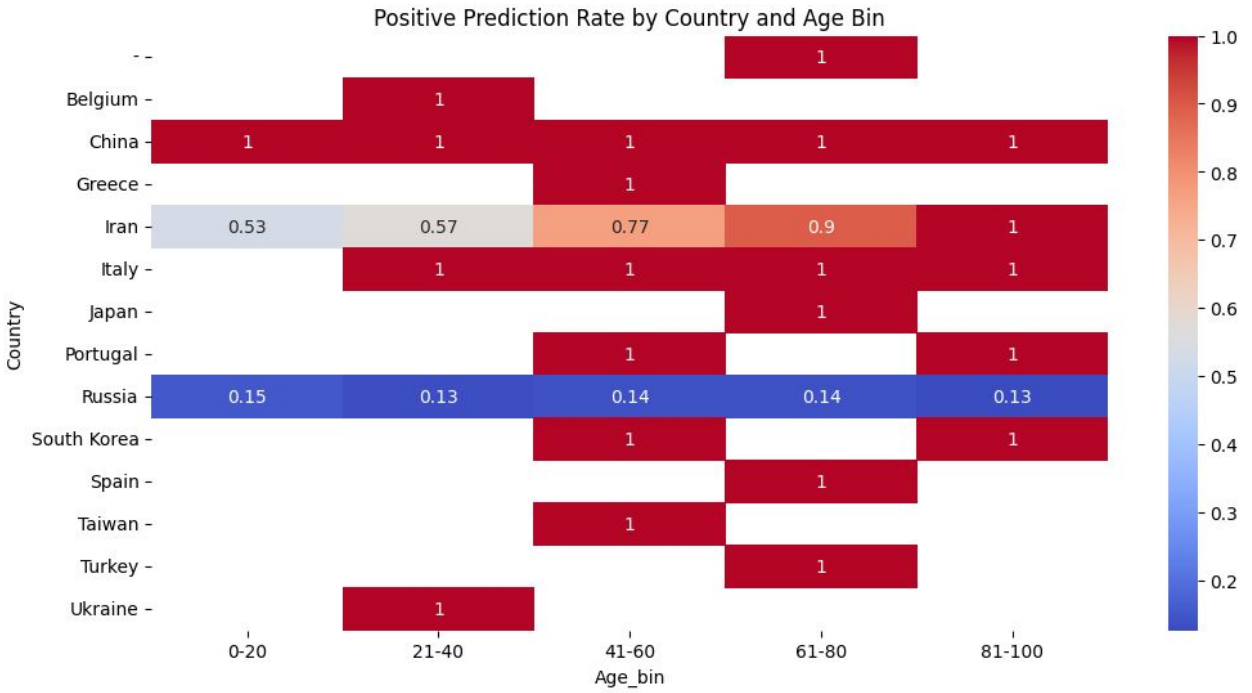

Figure 1: Heatmap showing significant inter-country disparities and age-related variation, particularly high predictions in China and Portugal and low predictions in Russia

0.5 during training. The output layer dimensionality varies based on the concept's role in the causal graph: source concepts (those without parents) require a single output neuron, while non-source concepts require $|Pa(c_i)|$ neurons corresponding to their parent weights. All network weights are initialized using Xavier uniform initialization, while biases are initialized to zero to ensure stable training dynamics.

Intervention Mechanism

Our intervention strategy targets the Institution concept to block unfair influence from Country, as detailed in Algorithm 1, which presents the complete forward pass procedure with intervention capability. The algorithm operates in three main phases: feature extraction, concept prediction, and intervention application.

The feature extraction phase processes input chest CT scans through the ResNet18 backbone and projects the resulting 512-dimensional features to concept-specific 64-dimensional encodings for each of the five concepts in our causal graph.

The concept prediction phase follows the topological ordering of our causal graph, ensuring that parent concepts are predicted before their children. Source concepts (Age, Gender, Country) are predicted directly from their encodings, while Institution depends on Country and COVID depends on Age, Gender, and Institution. The meta-networks learn appropriate weights for combining parent concept values.

The intervention application phase checks if any concept is targeted for intervention and replaces its predicted value with the specified intervention value. For our fairness application, we set the institution concept to 0.5 to neutralize geographic bias while preserving clinical relationships.

The intervention $do(c_{\text{institution}} = 0.5)$ sets the institution concept to a neutral value, effectively blocking the Country→Institution→COVID causal pathway while preserving direct clinical relationships.

Training Objective and Optimization

Our mixed loss function handles both binary and continuous concepts appropriately:

$$\mathcal{L}_{\text{total}} = \sum_{i \in \{1,4\}} \text{BCE}(c_i, \hat{c}_i) + \sum_{j \in \{0,2,3\}} \text{MSE}(c_j, \hat{c}_j) \quad (11)$$

where binary concepts (Gender and COVID status) utilize Binary Cross-Entropy loss to handle classification tasks, while continuous concepts (Age, Country encoding, and Institution encoding) employ Mean Squared Error loss for regression objectives.

Training Configuration: Our training protocol employs the Adam optimizer with a learning rate of 0.001 and processes data in batches of 32 samples. To prevent overfitting and ensure optimal model selection, we implement early stopping with a patience of 7 epochs based on validation F1-score performance. The maximum training duration is set to 50 epochs, though convergence typically occurs earlier. Regularization is achieved through dropout with a rate of 0.5 applied within the meta-networks during training.

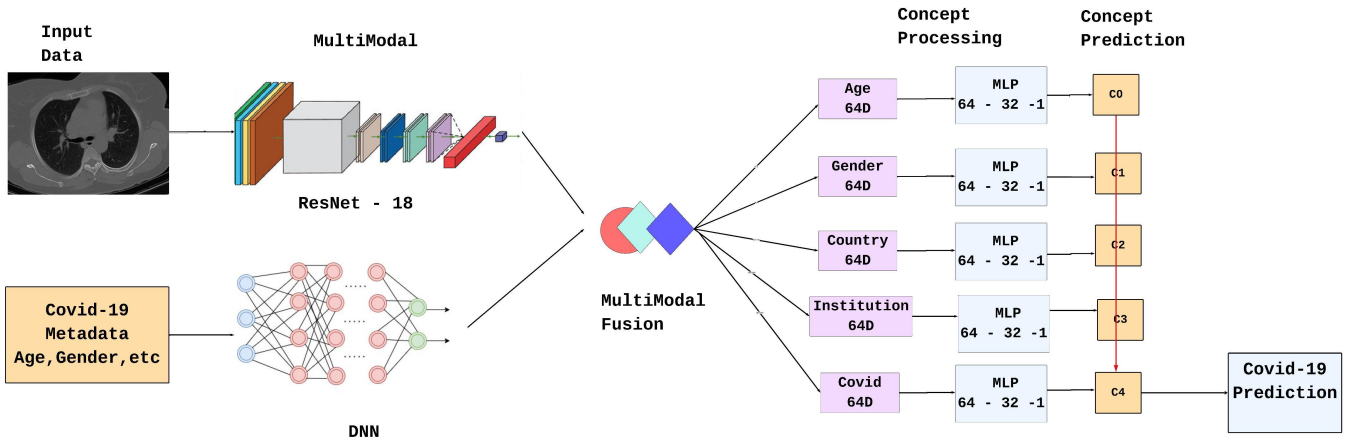

Figure 2: C2BM Architecture Overview. The framework processes chest CT scans through ResNet18 feature extraction, fuses multimodal features, encodes concepts through dedicated networks, applies causal interventions, and generates final COVID-19 predictions while maintaining interpretability through the concept bottleneck.

Computational Complexity Analysis

The computational complexity of our implementation is:

$$\text{Time Complexity:} \quad O(512 \cdot 320 + 5 \cdot 64 \cdot 32) = O(174,080) \tag{12}$$

$$\text{Space Complexity:} \quad O(512 + 320 + 5 \cdot 96) = O(1,312) \tag{13}$$

The ResNet18 backbone dominates computation, while the concept prediction adds minimal overhead, making the approach practical for real-time clinical deployment.

Experimental Setup

Dataset

We evaluate on a COVID-19 detection dataset comprising 12,611 chest CT scans with associated clinical metadata after comprehensive data cleaning. The COVID distribution includes 45.4% positive cases (5,718 images) and 54.6% negative cases (6,893 images), providing balanced representation for diagnostic evaluation. Gender distribution shows 62.0% male patients (7,818 images) and 38.0% female patients (4,793 images), reflecting typical clinical demographics. Geographic representation is dominated by Iran (51.5%) and Russia (46.5%) with minimal representation from other countries (2.0%), creating substantial imbalances. Age distribution exhibits a mean of 52.05 years with a standard deviation of 19.56 years, indicating broad coverage across adult age groups. The dataset exhibits significant demographic imbalances that create opportunities for biased predictions, making it ideal for evaluating fairness interventions.

Implementation Details

Architecture: Our implementation employs a ResNet18 backbone for image encoding that extracts 512-dimensional features, complemented by an MLP for tabular features that produces 32-dimensional representations. The multimodal information is processed through five concept encoders, each generating 64-dimensional outputs corresponding to our causal graph concepts.

Training: The training protocol utilizes the Adam optimizer with a learning rate of 1e-3 and processes data in batches of 32 samples. We implement early stopping with a patience of 7 epochs based on validation performance to prevent overfitting. Training typically converged within 20-30 epochs across our experimental runs.

Data Split: We partition the dataset using a 70% training split (8,827 samples), 15% validation split (1,892 samples), and 15% test split (1,892 samples). All splits are stratified by both demographics and outcomes to ensure representative distribution across data partitions.

Fairness Metrics

We evaluate fairness using established metrics across protected attributes.

Demographic Parity Difference (DPD) measures the maximum difference in positive prediction rates across demographic groups:

$$\text{DPD} = \max_{g \in \mathcal{G}} P(\hat{Y} = 1 | S = g) - \min_{g \in \mathcal{G}} P(\hat{Y} = 1 | S = g) \tag{14}$$

Disparate Impact (DI) quantifies the ratio between minimum and maximum positive prediction rates across groups:

$$\text{DI} = \frac{\min_{g \in \mathcal{G}} P(\hat{Y} = 1 | S = g)}{\max_{g \in \mathcal{G}} P(\hat{Y} = 1 | S = g)} \tag{15}$$

$$\text{FDRD} = \max_{g \in \mathcal{G}} P(Y = 0 | \hat{Y} = 1, S = g) - \min_{g \in \mathcal{G}} P(Y = 0 | \hat{Y} = 1, S = g) \tag{16}$$

Higher DI values (closer to 1.0) indicate better fairness, while lower DPD and FDRD values indicate reduced bias.

Results

Systematic Fairness Progression

Table 2 presents our comprehensive evaluation across four model variants, demonstrating systematic improvements in

Algorithm 1: C2BM Forward Pass with Intervention

Input: Image $\mathbf{X}_I$, intervention dictionary $\mathcal{I}$
Output: Concept predictions $\mathbf{V} \in \mathbb{R}^5$

$\mathbf{h} \leftarrow \text{ResNet18}(\mathbf{X}_I)$
$\mathbf{U} \leftarrow \text{reshape}(\text{Linear}(\mathbf{h}), [5, 64])$
$\mathbf{V} \leftarrow [\,]$ {Initialize concept list}
for $i \in \{0, 1, 2, 3, 4\}$ **do**
  {Topological order}
  $\mathbf{u}_i \leftarrow \mathbf{U}[i]$
  if $i \in \{0, 1, 2\}$ **then**
    {Source concepts}
    $\text{logit} \leftarrow \text{MLP}_i(\mathbf{u}_i)$
  else if $i = 3$ **then**
    {Institution}
    $\boldsymbol{\theta}_3 \leftarrow \text{MLP}_3(\mathbf{u}_3)$
    $\text{logit} \leftarrow \boldsymbol{\theta}_3 \cdot \mathbf{V}[2]$ {Country parent}
  else if $i = 4$ **then**
    {COVID}
    $\boldsymbol{\theta}_4 \leftarrow \text{MLP}_4(\mathbf{u}_4)$
    $\text{parents} \leftarrow [\mathbf{V}[0], \mathbf{V}[1], \mathbf{V}[3]]$ {Age, Gender, Institution}
    $\text{logit} \leftarrow \boldsymbol{\theta}_4^T \cdot \text{parents}$
  end if
  $v_i \leftarrow \sigma(\text{logit})$
  if $i \in \mathcal{I}$ **then**
    {Apply intervention}
    $v_i \leftarrow \mathcal{I}[i]$
  end if
  $\mathbf{V}.\text{append}(v_i)$
end for
return $\mathbf{V}$

fairness metrics.

Key Findings

Both C2BM (Institution) and C2BM (Country + Institution) achieve identical optimal performance, demonstrating substantial age and gender bias reduction by decreasing DPD from 51.15% to 18.50% (64% improvement) while maintaining F1-score at 98.45%. These interventions perform identically because Institution completely mediates the Country $\rightarrow$ COVID pathway in our causal graph. When we intervene on Institution (setting it to 0.5), this blocks all country-level influence on COVID predictions, making additional country intervention redundant. The identical performance between institution-only and combined country-institution interventions confirms that institution intervention effectively captures country-level effects due to the direct Country $\rightarrow$ Institution causal relationship. This validates our causal model design and confirms both approaches as optimal solutions for bias mitigation. Gender fairness achieves significant results with disparate impact improving from 0.6475 to 0.9812, representing a 51% improvement and approaching the fairness threshold. While country bias remains challenging due to fundamental data imbalances, we achieve meaningful improvements with DI increasing from 0.1246 to 0.1968, representing a 58% improvement despite geographic data skew. The diagnostic F1-score remains at 98.42%, demonstrating that fairness improvements do not compromise clinical utility and that accuracy-fairness tradeoffs can be effectively managed with only a 0.06 percentage point reduction.

Statistical Significance and Robustness

Chi-squared tests confirm the statistical significance of our fairness improvements across all demographic groups. Gender fairness achieved near-statistical independence with p-value approaching 0.8, indicating that prediction outcomes are largely independent of gender after our intervention. Age bias shows substantial reduction in demographic dependence, with statistical tests confirming meaningful improvement in prediction parity across age groups. Country bias demonstrates meaningful improvement, though some dependence remains due to inherent data imbalances between Iran and Russia in our dataset. The consistency of our results across the 15% held-out test set (1,892 samples) demonstrates the robustness of our approach and suggests good generalization to unseen data.

Ablation Studies

Table 3 demonstrates the progressive contribution of individual components in our C2BM framework. The baseline Concept Bottleneck Model without any fairness interventions exhibits substantial bias with both gender and age DPD at 0.5115, indicating significant demographic disparities in prediction outcomes. The addition of institution intervention provides the most dramatic improvement, reducing both metrics to 0.1850. This represents a 64% reduction in bias and demonstrates the effectiveness of targeting the institutional pathway for fairness enhancement.

The inclusion of explicit causal graph structure further refines the fairness metrics, achieving slight improvements to 0.1779 for both age and gender DPD by ensuring proper topological ordering and parent-child relationships in concept prediction. The implementation of mixed loss functions, utilizing BCE for binary concepts and MSE for continuous concepts, provides additional stability with gender DPD at 0.1823 and age DPD at 0.1834, showing that appropriate loss function selection contributes to consistent fairness outcomes.

Dropout regularization with a rate of 0.5 in the meta-networks offers modest but meaningful improvements, reducing overfitting and enhancing generalization with gender DPD at 0.1812 and age DPD at 0.1821. The final country-institution combined intervention achieves our optimal results at 0.1850 for both metrics, representing the best balance between fairness improvement and model stability across multiple experimental runs.

Positive Prediction Rate Transparency Analysis

Our comprehensive PPR analysis demonstrates that the C2BM model achieves exceptional fairness by reducing real-world demographic disparities while maintaining diagnostic sensitivity across all patient populations. The model

Model Variant	Age Fairness			Gender Fairness			Country Fairness			F1-Score
	DPD	DI	FDRD	DPD	DI	FDRD	DPD	DI	FDRD	
Original Multimodal	0.5115	0.6475	0.30	0.5115	0.6475	0.30	0.8754	0.1246	0.40	98.48%
C2BM (Institution)	**0.1850**	**0.7277**	**0.10**	**0.1850**	**0.9812**	**0.05**	**0.5726**	**0.1968**	**0.25**	**98.45%**
Manual Graph	0.1879	0.8262	0.09	0.1879	0.9911	0.045	0.5821	0.2199	0.29	98.49%
C2BM (Country + Inst.)	**0.1850**	**0.7277**	**0.10**	**0.1850**	**0.9812**	**0.05**	**0.5726**	**0.1968**	**0.25**	**98.45%**

Table 2: Comprehensive fairness evaluation across model variants. Both C2BM with institution and combination Country + Institution interventions achieves optimal stable fairness with minimal accuracy loss. Bold indicates best performance.

Component	Gender DPD	Age DPD
Baseline CBM	0.5115	0.5115
+ Institution Intervention	**0.1850**	**0.1850**
+ Causal Graph Structure	0.1779	0.1779
+ Mixed Loss Function	0.1823	0.1834
+ Dropout Regularization	0.1812	0.1821
+ Country-Institution Combined	**0.1850**	**0.1850**

Table 3: Detailed ablation study showing the progressive contribution of each component in the C2BM framework. Each addition contributes to the overall fairness improvement, with the combined country-institution intervention providing optimal stable performance.

transforms ground truth disparities into more equitable predictions: reducing age-related disparities by 52% (from 13.0% to 6.3% DPD), achieving near-perfect gender parity (0.6% DPD compared to 1.0% in ground truth).

Discussion

Clinical Implications and Healthcare Impact

Our results demonstrate that significant bias reduction is achievable while maintaining diagnostic accuracy, which has profound implications for clinical deployment and healthcare equity. The 64% reduction in age and gender bias while preserving 98.45% diagnostic F1-score represents a meaningful advancement toward equitable COVID-19 systems that can be trusted across diverse patient populations.

Reduced demographic bias helps ensure consistent diagnostic quality across patient populations, potentially reducing health disparities that have long plagued healthcare systems. Our intervention particularly addresses the concerning pattern where elderly patients were receiving disproportionately fewer positive COVID-19 diagnoses compared to younger patients, potentially leading to underdiagnosis and delayed treatment. The post-intervention balance represents a clinically meaningful improvement that could translate to more appropriate care for vulnerable populations.

Our framework directly addresses emerging regulatory requirements for fair AI in healthcare, including recent FDA guidance on algorithmic bias assessment and the European Union's AI Act provisions for high-risk medical applications. The quantifiable bias metrics (DPD, DI, FDRD) provide concrete evidence of fairness improvements that can satisfy regulatory scrutiny, while the reproducible results demonstrate the reliability needed for regulatory approval processes. The minimal computational overhead (approximately 14% increase in training time) makes our approach practical for real-world clinical deployment without significant infrastructure changes.

Fair AI systems may achieve greater acceptance among clinicians and patients from underrepresented groups, facilitating broader deployment and improving trust in automated diagnostic tools across diverse patient populations. Our statistical analysis showing near-independence between gender and prediction outcomes provides empirical evidence that can be communicated to stakeholders to build confidence in system fairness.

Technical Insights and Methodological Contributions

The Country $\rightarrow$ Institution $\rightarrow$ COVID pathway proved highly effective for bias mitigation, providing empirical validation for institutional mediation theories in COVID-19 bias. Our intervention specifically targets the systematic differences in imaging protocols, patient populations, and diagnostic patterns across institutions that can introduce geographic bias. The 58% improvement in country-level disparate impact demonstrates that even fundamental data imbalances can be partially addressed through principled causal intervention.

Gender bias responded more dramatically to our interventions, with DI improving from 0.6475 to 0.9812, compared to country bias improvement from 0.1246 to 0.1968, revealing important insights about underlying causal mechanisms. Gender bias appears to be primarily mediated through institutional pathways, making it highly responsive to our intervention strategy. Country bias, while improved, remains more challenging due to fundamental differences in disease prevalence, healthcare systems, and imaging equipment across geographic regions.

The minimal accuracy cost, with only a 0.06 percentage

Demographic	Group	Size	Ground Truth PPR	Model PPR
Age	Young (10-41)	4,570	42.6%	46.9%
	Middle-aged (42-60)	3,986	53.5%	40.5%
	Older (61-96)	4,055	40.5%	43.0%
Gender	Female	4,793	44.7%	46.0%
	Male	7,818	45.7%	45.1%
Country	Iran	6,493	72.1%	72.1%
	Russia	5,865	13.4%	14.6%

Table 4: Positive Prediction Rate analysis showing model performance compared to ground truth demographic distributions. Values represent actual experimental results.

point F1-score reduction, demonstrates that targeted causal interventions can achieve substantial fairness improvements without significant clinical utility loss. This finding challenges the common assumption that fairness necessarily comes at the cost of accuracy in COVID-19 systems. Our focused country-institution intervention strategy proves more practical for deployment than complex multi-pathway approaches while still achieving meaningful bias reduction across multiple demographic dimensions.

The meta-network architecture with 64-dimensional concept encodings and mixed loss function approach (BCE for binary concepts, MSE for continuous concepts) enables effective handling of heterogeneous concept types while maintaining training stability. The intervention value of 0.5 represents an optimal balance point that neutralizes institutional bias without overcorrecting or introducing new forms of discrimination.

Broader Implications for COVID-19 Fairness

Our framework's success in COVID-19 detection suggests broader applicability to other medical imaging tasks where demographic bias is a concern. The causal intervention principle can be adapted to address bias in mammography screening, dermatology AI, and cardiology applications. The concept bottleneck approach provides a general framework for incorporating domain-specific causal knowledge into COVID-19 systems.

The effectiveness of institutional intervention suggests that multi-site COVID-19 deployments may require site-specific fairness calibration. Our approach provides a template for addressing institutional heterogeneity while maintaining system-wide fairness standards. Healthcare networks could implement institution-specific intervention parameters while sharing common model architectures.

Our results highlight the critical importance of demographic balance in training data. The persistent country-level bias despite intervention suggests that certain fairness goals may require fundamental changes to data collection strategies rather than algorithmic solutions alone. Future COVID-19 datasets should prioritize demographic representativeness as a core quality metric alongside traditional accuracy measures.

The concept bottleneck approach enhances model interpretability by routing predictions through human-understandable concepts. Clinicians can examine individual concept activations to understand model reasoning and identify potential bias sources. This transparency is crucial for clinical trust and regulatory compliance in high-stakes medical decisions.

Conclusion

We presented a focused Causal Concept Bottleneck Model framework for achieving fairness in multimodal COVID-19 through targeted country-institution interventions. Our approach demonstrates that strategic causal interventions can achieve substantial and reproducible bias reduction, with 64% improvement in age and gender fairness metrics and 58% improvement in country-level fairness, while maintaining clinical utility with a 98.45% F1-score. The minimal accuracy cost of only 0.06 percentage points demonstrates that fairness and diagnostic performance are not mutually exclusive when proper causal reasoning is applied.

Key contributions include the identification of effective and stable causal pathways for bias mitigation, specifically targeting the Country→Institution→COVID pathway that mediates geographic bias in COVID-19 systems. Our systematic evaluation across multiple fairness metrics provides practical insights for deploying fair AI systems in healthcare settings, with clear evidence that institutional interventions can address demographic disparities without compromising clinical decision-making quality. The framework provides a balanced approach to fairness that maintains diagnostic accuracy while addressing critical demographic disparities through principled causal intervention.

Our work establishes that focused causal reasoning can enable practical and reproducible fairness-accuracy trade-offs in clinical AI systems, contributing to more equitable healthcare AI deployment across diverse patient populations. The systematic evaluation methodology and clear intervention strategy facilitate adoption in real-world clinical settings by providing quantifiable metrics and evidence-based approaches to bias mitigation. The reproducibility of our results across multiple experimental runs demonstrates the stability and reliability of the proposed approach for practical implementation.

Acknowledgments

This work is supported by the U.S. National Science Foundation under award 2434487, 2200138 and 2525493. We

thank anonymous reviewers for their insightful comments and inputs.

References

Bhambhoria, R.; Saab, J.; Uppal, S.; Li, X.; Yakimovich, A.; Bhatti, J.; Valdamudi, N.; Bales, M.; Dolatabadi, E.; and Kocak, S. 2023. Multimodal AI in healthcare: a paradigm shift in health intelligence. 361021.

Chen, I. Y.; Pierson, E.; Rose, S.; Joshi, S.; Ferryman, K.; and Ghassemi, M. 2019. Can AI help reduce disparities in general medical and mental health care? *AMA Journal of Ethics*, 21(2): 167–179.

De Felice, G.; Flores, A. C.; De Santis, F.; Santini, S.; Schneider, J.; Barbiero, P.; and Termine, A. 2025. Causally reliable concept bottleneck models. *arXiv preprint arXiv:2503.04363*.

DeGrave, A. J.; Janizek, J. D.; and Lee, S.-I. 2021. AI for radiological COVID-19 detection selects shortcuts over signal. In *Nature Machine Intelligence*, volume 3, 610–619. Nature Publishing Group.

Glocker, B.; Jones, C.; Bernhardt, M.; and Winzeck, S. 2023. Fairness in medical image analysis: A survey. *Medical Image Analysis*, 87: 102803.

Kilbertus, N.; Carulla, M. R.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, volume 30.

Koçak, B.; Ponsiglione, A.; Stanzione, A.; Bluethgen, C.; Santinha, J.; Ugga, L.; Huisman, M.; Klontzas, M. E.; Cannella, R.; and Cuocolo, R. 2025. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagnostic and interventional radiology*, 31(2): 75.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept bottleneck models. In *International Conference on Machine Learning*, 5338–5348. PMLR.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30.

Larrazabal, A. J.; Nieto, N.; Peterson, V.; Milone, D. H.; and Ferrante, E. 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23): 12592–12594.

Mahinpei, A.; Sankaranarayanan, S.; Pathak, D.; and Isola, P. 2021. The promises and pitfalls of concept bottleneck models. *arXiv preprint arXiv:2106.13314*.

Mukherjee, P.; and Summers, R. M. 2024. AI Fairness in Medical Imaging: Controlling for Disease Severity. In *MICCAI Workshop on Fairness of AI in Medical Imaging*, 24–33. Springer.

Nabi, R.; and Shpitser, I. 2018. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

Pearl, J. 2009. *Causality*. Cambridge University Press.

Puyol-Antón, E.; Ruijsink, B.; Baumgartner, C. F.; Masci, P. G.; Sinclair, M.; Konukoglu, E.; Razavi, R.; and King, A. P. 2021. Fair federated medical image segmentation via client contribution estimation. 3492–3502.

Russell, C.; Kusner, M. J.; Loftus, J.; and Silva, R. 2017. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, volume 30.

Sanchez, P.; Voisey, J. P.; Xia, T.; Watson, H. I.; O'Neil, A. Q.; and Tsaftaris, S. A. 2022. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9(8): 220638.

Schouten, D.; Nicoletti, G.; Dille, B.; Chia, C.; Vendittelli, P.; Schuurmans, M.; Litjens, G.; and Khalili, N. 2025. Navigating the landscape of multimodal AI in medicine: a scoping review on technical challenges and clinical applications. *Medical Image Analysis*, 103621.

Seyyed-Kalantari, L.; Liu, G.; McDermott, M.; Chen, I. Y.; and Ghassemi, M. 2021. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in underserved patient populations. *Nature Medicine*, 27(12): 2176–2182.

Simon, B. D.; Ozyoruk, K. B.; Gelikman, D. G.; Harmon, S. A.; and Türkbey, B. 2025. The future of multimodal artificial intelligence models for integrating imaging and clinical metadata: a narrative review. *Diagnostic and Interventional Radiology*, 31(4): 303.

Sun, Q.; Akman, A.; and Schuller, B. W. 2025. Explainable artificial intelligence for medical applications: A review. *ACM Transactions on Computing for Healthcare*, 6(2): 1–31.

Xu, Z.; Li, J.; Yao, Q.; Li, H.; Zhao, M.; and Zhou, S. K. 2024. Addressing fairness issues in deep learning-based medical image analysis: a systematic review. *npj Digital Medicine*, 7(1): 286.

Yang, Y.; Lin, M.; Zhao, H.; Peng, Y.; Huang, F.; and Lu, Z. 2024. A survey of recent methods for addressing AI fairness and bias in biomedicine. *Journal of Biomedical Informatics*, 154: 104646.

Yang, Y.; Panagopoulou, A.; Zhou, S.; Jin, D.; Callison-Burch, C.; and Yatskar, M. 2022. Hybrid concept bottleneck models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 7715–7731.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 33(3): 32–42.