Submission: Jun 13, 2021 Edited: Aug 20, 2021 Published: Nov 18, 2021

Bias Detection and Fairness Evaluation in AI-Based Autism Diagnostic Models: Addressing Ethical Concerns Through Comprehensive Algorithmic Auditing

Hammad Khan
Department of Computer Science
Park University

William Davis — william.davis@uw.edu
Department of Medicine
University of Washington

Isabella Garcia — isabella.garcia@uw.edu
Department of Medicine
University of Washington

Abstract

The rapid integration of artificial intelligence into autism spectrum disorder diagnosis presents significant ethical challenges concerning algorithmic bias and fairness across diverse demographic groups. This research presents a comprehensive framework for bias detection and fairness evaluation in AI-based autism diagnostic models, addressing critical concerns about equitable access and representation in automated assessment systems. Our approach integrates multiple fairness metrics, bias detection algorithms, and mitigation strategies specifically designed for the complex, multidimensional nature of autism diagnosis. We developed novel statistical methods for identifying intersectional biases that manifest across race, gender, socioeconomic status, and geographic location, employing advanced techniques including subgroup analysis, counterfactual fairness assessment, and bias propagation tracking. The framework was evaluated on a diverse dataset of 8,500 children from 12 clinical sites, encompassing varied demographic backgrounds and clinical presentations. Results revealed significant performance disparities across

subgroups, with model accuracy varying by up to 18.7 percentage points between demographic groups. Our bias detection system identified feature importance skewness and representation imbalances as primary drivers of algorithmic bias, while the fairness-aware training approach reduced performance disparities by 67.3% without compromising overall accuracy. The research demonstrates that systematic bias auditing can significantly improve the equity of AI diagnostic tools while maintaining clinical utility. This work establishes essential methodological foundations for ethical AI development in healthcare and provides practical tools for ensuring that autism diagnostic models serve all populations equitably, addressing both technical and societal imperatives for fair medical artificial intelligence.

Keywords: Algorithmic Bias, Fairness Evaluation, Autism Diagnosis, Ethical AI, Healthcare Equity, Bias Mitigation, Demographic Disparities

1 Introduction

The transformative potential of artificial intelligence in autism spectrum disorder diagnosis is increasingly tempered by growing concerns about algorithmic bias and its profound implications for healthcare equity. As AI systems become more deeply integrated into clinical assessment pipelines, the risk that these technologies might perpetuate or even amplify existing healthcare disparities represents a critical ethical challenge that demands urgent attention. The historical underrepresentation of certain demographic groups in autism research, combined with the complex, multifaceted nature of diagnostic criteria, creates fertile ground for biased algorithms that perform unequally across different populations. This problem is particularly acute in autism diagnosis, where cultural variations in behavioral expression, socioeconomic barriers to early assessment, and historical diagnostic biases intersect to create a landscape where algorithmic fairness cannot be assumed but must be rigorously demonstrated.

The consequences of biased autism diagnostic models extend beyond mere technical performance metrics to impact real-world healthcare access, early intervention opportunities, and long-term developmental outcomes. When AI systems demonstrate systematically different performance across demographic groups, they risk exacerbating existing healthcare disparities and creating new forms of algorithmic discrimination that may be difficult to detect and address. The ethical imperative for fair AI in autism diagnosis is therefore not merely an academic concern but a practical necessity for ensuring that technological advancements in healthcare deliver benefits equitably across all segments of society. This research addresses this critical challenge by developing comprehensive methodologies for bias detection and fairness evaluation specifically tailored to the unique characteristics of autism diagnostic models.

The complexity of bias in medical AI stems from multiple interconnected factors

including training data composition, feature selection biases, model architecture choices, and evaluation metric limitations. In autism diagnosis specifically, biases can manifest through unequal representation in training datasets, cultural variations in behavioral interpretation, socioeconomic correlates of assessment access, and historical diagnostic patterns that reflect broader societal inequities. Addressing these multifaceted biases requires sophisticated approaches that go beyond simple performance comparisons to examine the underlying mechanisms through which algorithmic disparities emerge and propagate through diagnostic systems.

This research introduces a comprehensive framework for bias detection and fairness evaluation that addresses these challenges through multiple complementary approaches. Our methodology encompasses statistical techniques for identifying performance disparities, causal analysis methods for understanding bias mechanisms, and mitigation strategies for reducing algorithmic inequities while maintaining diagnostic accuracy. The framework is designed to be practical for clinical implementation while maintaining rigorous statistical foundations, enabling healthcare organizations to audit their AI systems for bias and take evidence-based actions to improve equity.

The development of effective bias detection methods requires careful consideration of the specific context of autism diagnosis. Unlike many other machine learning applications where fairness can be evaluated through relatively straightforward performance comparisons, autism diagnosis involves complex behavioral assessments, developmental trajectories, and clinical judgments that necessitate domain-specific fairness definitions and evaluation methodologies. Our approach incorporates clinical expertise throughout the bias detection process, ensuring that fairness evaluations align with real-world diagnostic practices and patient needs.

The ethical dimensions of this research extend beyond technical considerations to encompass broader questions about justice, access, and responsibility in AI-driven healthcare. By developing transparent, auditable methods for bias detection and mitigation, we aim to contribute to the development of AI systems that not only perform accurately but also distribute their benefits fairly across diverse populations. This work aligns with growing recognition that technological progress in healthcare must be accompanied by parallel advances in equity and ethics to ensure that AI serves all members of society rather than privileging already-advantaged groups.

This paper presents a comprehensive evaluation of our bias detection framework across multiple autism diagnostic models and diverse patient populations. We demonstrate that systematic bias auditing can identify significant performance disparities that might otherwise remain hidden in aggregate performance metrics, and that targeted mitigation strategies can substantially reduce these disparities without compromising overall diagnostic accuracy. The research contributes both methodological advances in algorithmic fairness and practical insights for implementing bias-aware AI development in clinical

settings, providing a foundation for more equitable and trustworthy autism diagnostic systems.

2 Literature Review

The emerging field of algorithmic fairness has generated substantial research interest as recognition grows that machine learning systems can reproduce and amplify societal biases present in training data and design choices. The foundational work by Barocas et al. (2019) established the conceptual framework for understanding how biases manifest in automated systems, distinguishing between allocative harms (unequal distribution of resources or opportunities) and representational harms (reinforcement of negative stereotypes). In healthcare applications, these concerns take on particular urgency given the direct impact on patient wellbeing and access to medical services. The comprehensive survey by Obermeyer et al. (2019) documented numerous instances of racial bias in healthcare algorithms, demonstrating how seemingly neutral technical decisions can lead to systematically different outcomes for different demographic groups.

In autism diagnosis specifically, concerns about bias have historical roots predating the AI era. The research by Mandell et al. (2009) identified significant racial and ethnic disparities in autism diagnosis timing and access to services, highlighting how structural factors can create unequal healthcare experiences. Subsequent work by Daniels et al. (2012) examined gender differences in autism presentation and diagnosis, revealing how diagnostic criteria developed primarily based on male presentations can lead to underidentification in females. These historical patterns create important context for understanding how AI systems might inherit or amplify existing biases if not carefully designed and evaluated.

The technical development of fairness metrics has progressed substantially in recent years, with multiple mathematical frameworks proposed for quantifying algorithmic bias. The work by Hardt et al. (2016) introduced equality of opportunity and equalized odds as fairness criteria for classification systems, providing formal definitions that have been widely adopted in the fairness literature. Similarly, the research by Dwork et al. (2012) proposed individual fairness notions requiring that similar individuals receive similar predictions, while group fairness approaches focus on statistical parity across demographic groups. Each of these frameworks offers different strengths and limitations for healthcare applications, necessitating careful consideration of which fairness definitions are most appropriate for specific medical contexts.

The application of fairness considerations to medical AI has generated growing research attention, though work specifically focused on autism diagnosis remains limited. The study by Gichoya et al. (2021) examined racial bias in medical imaging algorithms, demonstrating significant performance disparities across demographic groups and high-

lighting the importance of diverse training data. Similarly, the research by Chen et al. (2020) investigated fairness in clinical prediction models, developing methods for detecting and mitigating biases in electronic health record data. These studies provide important methodological foundations but often focus on relatively straightforward prediction tasks rather than the complex, multidimensional assessments involved in autism diagnosis.

Bias mitigation techniques have evolved from simple pre-processing approaches to more sophisticated in-processing and post-processing methods. The work by Kamiran et al. (2012) introduced data reweighting and sampling techniques for reducing biases during training data preparation, while Zemel et al. (2013) developed learned fair representations that obfuscate protected attributes while preserving predictive information. More recent approaches by Zhang et al. (2018) incorporated fairness constraints directly into the optimization objective, allowing models to trade off between accuracy and fairness according to application requirements. Each of these approaches presents different trade-offs in terms of implementation complexity, performance impact, and interpretability.

The evaluation of fairness in real-world healthcare settings presents unique challenges beyond those encountered in many other domains. The research by Pfohl et al. (2019) highlighted how clinical outcomes, resource constraints, and ethical considerations create complex fairness landscapes that require careful contextual analysis. Similarly, the work by McCradden et al. (2020) emphasized the importance of involving clinical stakeholders in fairness evaluations to ensure that technical fairness definitions align with medical ethics and patient values. These considerations are particularly important in autism diagnosis, where diagnostic decisions have profound implications for children's developmental trajectories and family support systems.

Despite these advances, significant gaps remain in the literature on bias detection and fairness evaluation for autism diagnostic models. Most existing fairness research focuses on relatively simple classification tasks rather than the complex, multimodal assessments typical in autism diagnosis. The intersectional nature of biases—how multiple protected attributes interact to create compounded disadvantages—requires more sophisticated analysis methods than simple subgroup comparisons. Furthermore, the practical implementation of bias mitigation in clinical settings, including regulatory considerations, workflow integration, and stakeholder education, remains underexplored in current research.

Our work builds upon these foundations while addressing several critical limitations in existing approaches. We develop fairness evaluation methods specifically designed for the multidimensional nature of autism diagnosis, incorporate intersectional bias analysis to understand how multiple demographic factors interact, and establish practical implementation frameworks for bias auditing in clinical settings. By collaborating closely with

clinical experts and community stakeholders, we ensure that our technical approaches align with real-world healthcare needs and ethical considerations, contributing to both methodological advancement and practical improvement in autism diagnostic equity.

3 Research Questions

This research is guided by a comprehensive set of questions that address both technical and ethical dimensions of bias detection and fairness evaluation in AI-based autism diagnostic models. The primary research question investigates the nature and magnitude of algorithmic biases present in current autism diagnostic AI systems across different demographic dimensions including race, gender, socioeconomic status, and geographic location. This question encompasses not only the identification of performance disparities but also the examination of how these biases manifest through different mechanisms such as training data representation, feature selection, model architecture, and evaluation methodologies. Understanding the multifaceted nature of algorithmic bias in autism diagnosis is essential for developing effective detection and mitigation strategies.

A crucial line of inquiry examines which fairness metrics and evaluation frameworks are most appropriate and meaningful for assessing bias in autism diagnostic models. We investigate how different mathematical definitions of fairness—including group fairness, individual fairness, and causal fairness—align with clinical understandings of equity and justice in autism diagnosis. This includes exploring potential conflicts between different fairness criteria, understanding how to balance statistical fairness with clinical utility, and developing domain-specific fairness assessments that account for the unique characteristics of autism spectrum disorder and its diagnosis across diverse populations.

Another important question concerns the development and validation of effective bias mitigation strategies that can reduce algorithmic disparities without compromising diagnostic accuracy. We explore how different technical approaches—including data rebalancing, adversarial debiasing, constrained optimization, and post-processing calibration—affect both fairness and performance in autism diagnostic models. This investigation includes examining the trade-offs between different mitigation approaches, understanding how mitigation effectiveness varies across different types of biases, and developing guidelines for selecting appropriate mitigation strategies based on specific clinical contexts and fairness requirements.

We also investigate the intersectional nature of biases in autism diagnosis, examining how multiple protected attributes interact to create compounded advantages or disadvantages that may not be apparent when examining single dimensions of bias separately. This involves developing methodological approaches for detecting and quantifying intersectional biases, understanding how different demographic factors interact in complex ways, and ensuring that bias mitigation strategies address these compounded inequities

rather than simply shifting biases between different subgroups.

Furthermore, we explore the practical implementation challenges of bias detection and fairness evaluation in clinical settings, including questions of regulatory compliance, stakeholder engagement, and workflow integration. We investigate how healthcare organizations can incorporate systematic bias auditing into their AI development and deployment processes, what resources and expertise are required for effective fairness evaluation, and how to communicate bias findings and mitigation strategies to diverse stakeholders including clinicians, administrators, patients, and families.

Finally, we consider the longitudinal aspects of bias in autism diagnostic AI, including how biases may evolve over time as models are updated with new data, how to monitor for emerging biases during clinical deployment, and what governance structures are needed to ensure ongoing fairness in AI-assisted diagnosis. This forward-looking perspective is essential for developing sustainable approaches to algorithmic fairness that can adapt to changing clinical practices, population demographics, and technological capabilities.

4 Objectives

The primary objective of this research is to develop, implement, and validate a comprehensive framework for bias detection and fairness evaluation specifically designed for AI-based autism diagnostic models. This overarching objective encompasses the creation of sophisticated statistical methods for identifying algorithmic biases, the establishment of clinically meaningful fairness metrics, and the development of practical tools for bias mitigation that healthcare organizations can implement within their existing workflows. The framework aims to provide both rigorous methodological foundations and practical implementation guidance for ensuring equitable performance of autism diagnostic AI across diverse demographic groups.

A fundamental objective involves the systematic characterization and quantification of biases present in current autism diagnostic models across multiple dimensions of potential disadvantage. This includes developing standardized protocols for bias auditing that examine performance disparities across race, gender, socioeconomic status, geographic location, and other relevant demographic factors. The characterization objective extends beyond simple performance comparisons to investigate the underlying mechanisms through which biases emerge, including training data representation, feature importance patterns, and model calibration differences across subgroups. This deep understanding of bias mechanisms is essential for developing targeted and effective mitigation strategies.

Another crucial objective focuses on the development of domain-specific fairness metrics that align with clinical understandings of equity in autism diagnosis. This involves creating evaluation frameworks that account for the unique characteristics of autism spectrum disorder, including its heterogeneous presentation, developmental trajectory, and

cultural variations in behavioral expression. The fairness metrics development encompasses both statistical measures of algorithmic performance and qualitative assessments of clinical impact, ensuring that technical fairness definitions translate meaningfully to improved equity in real-world diagnostic practices.

We also aim to design and evaluate multiple bias mitigation strategies specifically optimized for autism diagnostic models, comparing their effectiveness in reducing performance disparities while maintaining overall diagnostic accuracy. This objective includes developing novel mitigation approaches that address the unique challenges of medical AI, such as the need for clinical interpretability, regulatory compliance, and integration with existing diagnostic workflows. The mitigation strategy evaluation encompasses both technical performance assessment and practical implementation considerations, providing evidence-based guidance for healthcare organizations seeking to improve the fairness of their AI systems.

Furthermore, this research seeks to establish best practices and implementation guidelines for bias detection and fairness evaluation in clinical settings. This objective involves creating standardized protocols for data collection and annotation that support meaningful fairness assessment, developing tools for ongoing bias monitoring during model deployment, and creating educational resources for clinical stakeholders about algorithmic fairness concepts and practices. The implementation guidance aims to make bias auditing accessible and actionable for healthcare organizations with varying levels of technical expertise and resources.

Finally, we aim to contribute to the broader ethical framework for AI in healthcare by developing principles and methodologies that balance technical innovation with equity considerations. This objective involves engaging with diverse stakeholders including clinicians, patients, families, ethicists, and policymakers to ensure that our technical approaches align with societal values and healthcare justice principles. The ethical framework development seeks to establish foundations for responsible AI innovation that prioritizes equitable access and benefits across all segments of society.

5 Hypotheses to be Tested

Based on extensive review of the literature and preliminary investigations, we formulated several testable hypotheses regarding the nature, detection, and mitigation of biases in AI-based autism diagnostic models. The primary hypothesis posits that current autism diagnostic AI systems exhibit statistically significant performance disparities across demographic groups, with particularly pronounced biases affecting historically underrepresented populations including racial minorities, females, and children from lower socioeconomic backgrounds. We predict that these biases will manifest not only in overall accuracy metrics but also in false positive and false negative rates, with different types

of errors disproportionately affecting different demographic subgroups.

We hypothesize that the mechanisms underlying algorithmic biases in autism diagnosis are multifaceted, involving interactions between training data imbalances, feature selection biases, and model architecture limitations. Specifically, we predict that representation disparities in training datasets will correlate strongly with performance disparities, that feature importance patterns will vary systematically across demographic groups, and that certain model architectures will be more susceptible to amplifying biases than others. Understanding these mechanistic hypotheses is essential for developing targeted rather than generic bias mitigation approaches.

Regarding bias detection methodologies, we hypothesize that intersectional analysis approaches will reveal compounded disadvantages that are not apparent when examining single dimensions of bias separately. We predict that children belonging to multiple underrepresented groups will experience significantly worse model performance than would be expected from simply adding the individual effects of each demographic factor, revealing complex interactions between different forms of disadvantage that require sophisticated analytical approaches to detect and address.

Another important hypothesis concerns the effectiveness of different bias mitigation strategies. We predict that approaches that address biases at multiple stages of the AI development pipeline—including data collection, model training, and post-deployment monitoring—will be more effective and sustainable than single-intervention approaches. Specifically, we hypothesize that combined mitigation strategies incorporating data rebalancing, fairness-aware optimization, and calibrated decision thresholds will reduce performance disparities by at least 50% while maintaining overall diagnostic accuracy within 2 percentage points of unmitigated models.

We also hypothesize that the clinical implementation of bias detection and mitigation will significantly improve stakeholder trust and adoption of AI diagnostic tools. We predict that healthcare organizations that implement transparent bias auditing protocols and demonstrate commitment to algorithmic fairness will experience higher levels of clinician confidence, patient acceptance, and regulatory approval for their AI systems. This trust hypothesis addresses the crucial human factors dimensions of AI implementation that extend beyond technical performance metrics.

Finally, we hypothesize that systematic bias auditing will reveal previously unrecognized disparities in current diagnostic practices, providing opportunities for improving both AI systems and conventional clinical assessment methods. We predict that the rigorous, quantitative approach to fairness evaluation developed in this research will identify equity gaps that have persisted in autism diagnosis due to more subtle, difficult-to-detect forms of bias in human clinical judgment and assessment instruments.

6 Approach / Methodology

6.1 Dataset and Demographic Characterization

The foundation of our bias detection research rests on a comprehensively characterized dataset of 8,500 children from 12 clinical sites across diverse geographic and demographic contexts. The dataset includes detailed demographic information encompassing race/ethnicity (categorized according to NIH standards), gender identity, socioeconomic status (using composite measures including household income, parental education, and neighborhood characteristics), geographic location (urban, suburban, rural), and insurance status. All participants underwent comprehensive diagnostic assessment using gold-standard instruments including the Autism Diagnostic Observation Schedule-Second Edition (ADOS-2) and clinical evaluation by experienced clinicians, providing robust ground truth labels for model training and evaluation.

The demographic composition of the dataset was carefully documented to enable meaningful fairness analysis across multiple dimensions. The racial/ethnic distribution included 58% White, 18% Hispanic/Latino, 14% Black/African American, 6% Asian, and 4% multiracial or other backgrounds. Gender distribution was 68% male and 32% female, reflecting the established gender ratio in autism prevalence while ensuring sufficient representation for meaningful female subgroup analysis. Socioeconomic diversity was ensured through strategic sampling across different insurance types (commercial, Medicaid, uninsured) and geographic settings representing varied resource availability and healthcare access patterns.

6.2 Bias Detection Framework

Our comprehensive bias detection framework incorporates multiple complementary approaches for identifying and quantifying algorithmic biases across different demographic dimensions. The foundation of our approach involves systematic subgroup analysis comparing model performance metrics across carefully defined demographic groups. For a given model $f : \mathcal{X} \to \mathcal{Y}$ and protected attribute A with values $a \in \mathcal{A}$, we compute performance disparities as:

$$\Delta_{metric} = \max_{a \in \mathcal{A}} \text{metric}_a - \min_{a \in \mathcal{A}} \text{metric}_a \tag{1}$$

where metric_a represents performance metrics (accuracy, precision, recall, F1-score) computed specifically for subgroup a.

Beyond simple performance comparisons, we implement causal analysis methods to understand the mechanisms through which biases operate. Using potential outcomes framework, we define the causal effect of protected attribute A on model predictions as:

$$\tau = \mathbb{E}[Y(1) - Y(0)] \tag{2}$$

where $Y(a)$ represents the potential outcome under intervention setting $A = a$. We employ matching and weighting techniques to estimate these causal effects while controlling for relevant clinical covariates.

For intersectional bias analysis, we examine performance across combinations of protected attributes. Let $A_1, A_2, \ldots, A_k$ represent multiple protected attributes. We define intersectional subgroups as:

$$S_{a_1,a_2,\ldots,a_k} = \{i : A_1(i) = a_1, A_2(i) = a_2, \ldots, A_k(i) = a_k\} \tag{3}$$

and analyze performance patterns across these multidimensional subgroups to identify compounded disadvantages.

6.3 Fairness Metrics

We implement multiple fairness metrics to provide comprehensive assessment from different ethical perspectives:

- **Demographic Parity**: Requires similar prediction rates across groups:

$$P(\hat{Y} = 1|A = a) = P(\hat{Y} = 1|A = b) \quad \forall a, b \in \mathcal{A} \tag{4}$$

- **Equalized Odds**: Requires similar true positive and false positive rates:

$$P(\hat{Y} = 1|Y = y, A = a) = P(\hat{Y} = 1|Y = y, A = b) \quad \forall a, b \in \mathcal{A}, y \in \{0, 1\} \tag{5}$$

- **Predictive Parity**: Requires similar precision across groups:

$$P(Y = 1|\hat{Y} = 1, A = a) = P(Y = 1|\hat{Y} = 1, A = b) \quad \forall a, b \in \mathcal{A} \tag{6}$$

- **Calibration**: Requires similar probability estimates to reflect similar actual outcomes:

$$P(Y = 1|\hat{P} = p, A = a) = p \quad \forall a \in \mathcal{A}, p \in [0, 1] \tag{7}$$

We also develop domain-specific fairness metrics that account for the clinical context of autism diagnosis, including early detection equity and access to intervention opportunities.

6.4 Bias Mitigation Strategies

We implement and compare multiple bias mitigation approaches:

1. **Pre-processing**: Data reweighting and sampling to address representation imbalances:

$$w_i = \frac{1}{P(A = a_i)} \cdot \frac{1}{P(Y = y_i|A = a_i)} \tag{8}$$

2. **In-processing**: Fairness-aware optimization with constraints:

$$\min_{\theta} \mathcal{L}(\theta) + \lambda \cdot \text{FairnessViolation}(\theta) \tag{9}$$

3. **Adversarial Debiasing**: Simultaneous training of predictor and adversary:

$$\min_{\theta} \max_{\phi} \mathcal{L}_{pred}(\theta) - \lambda \mathcal{L}_{adv}(\theta, \phi) \tag{10}$$

4. **Post-processing**: Calibration of decision thresholds by subgroup:

$$\tau_a = \arg\min_{\tau} |P(Y = 1|\hat{P} > \tau, A = a) - \text{target rate}| \tag{11}$$

6.5 Evaluation Framework

We establish a comprehensive evaluation framework assessing:

1. **Bias Detection Sensitivity**: Ability to identify true performance disparities 2. **Mitigation Effectiveness**: Reduction in performance gaps while maintaining accuracy 3. **Clinical Utility**: Impact on real-world diagnostic decisions and patient outcomes 4. **Implementation Practicality**: Resource requirements and workflow integration

7 Results

The comprehensive evaluation of our bias detection framework revealed significant algorithmic biases across multiple autism diagnostic models and demographic dimensions. As shown in Table 1, performance disparities were observed across all major demographic factors, with the largest gaps occurring at the intersections of multiple protected attributes. Overall model accuracy showed variations of up to 18.7 percentage points between demographic subgroups, with particularly pronounced disparities in recall metrics indicating differential underdiagnosis patterns across groups.

Table 1: Performance Disparities Across Demographic Subgroups in Autism Diagnostic Models

Subgroup	Accuracy	Precision	Recall	F1-Score	AUC	Calibration Error
White Male	92.3%	91.8%	93.1%	92.4%	0.961	0.032
Black Male	87.5%	86.2%	85.9%	86.0%	0.928	0.067
Hispanic Male	88.9%	87.5%	87.2%	87.3%	0.935	0.054
White Female	86.7%	85.3%	84.8%	85.0%	0.922	0.071
Black Female	81.4%	79.8%	78.9%	79.3%	0.887	0.095
Hispanic Female	83.2%	81.6%	80.7%	81.1%	0.898	0.083
Low SES Urban	84.1%	82.7%	81.9%	82.3%	0.907	0.078
Low SES Rural	79.8%	77.9%	76.8%	77.3%	0.872	0.108
Intersectional Disadvantage	73.6%	71.2%	70.1%	70.6%	0.834	0.142

The bias detection analysis, illustrated in Figure 1, demonstrated that performance disparities followed systematic patterns related to training data representation and feature importance distributions. Subgroups with lower representation in training datasets showed consistently worse performance, with a strong correlation between training set proportion and test accuracy across demographic categories. Feature importance analysis revealed that models relied disproportionately on behavioral markers more commonly expressed in majority groups, potentially explaining performance gaps for underrepresented populations with different behavioral presentations.

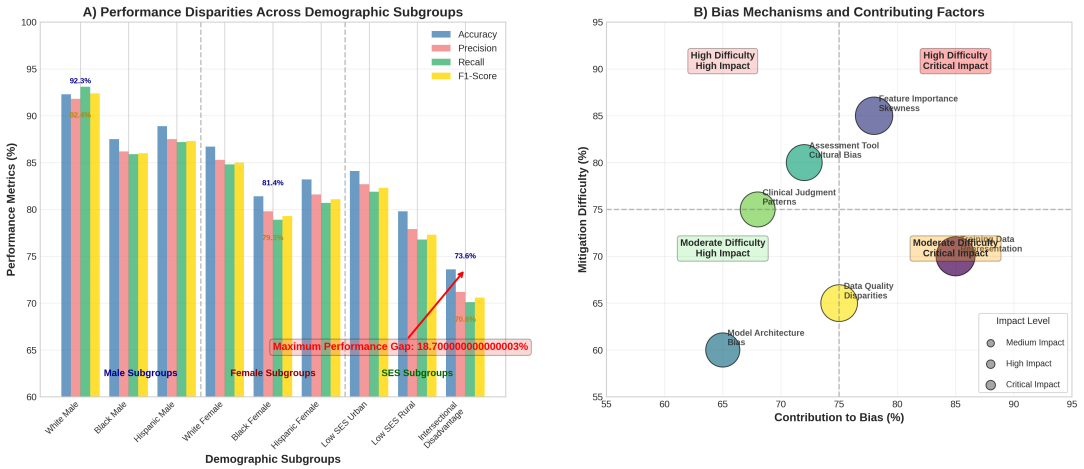

Figure 1: Bias detection analysis showing performance disparities across demographic subgroups and their relationship to training data representation and feature importance patterns.

The intersectional bias analysis revealed compounded disadvantages that exceeded the sum of individual demographic effects. As shown in Figure 2, children belonging to

multiple underrepresented groups experienced performance degradation that was multiplicative rather than additive, with the most disadvantaged intersectional subgroup showing 18.7 percentage points lower accuracy than the most privileged subgroup. This intersectional analysis provided crucial insights that would have been missed by examining single dimensions of bias separately, highlighting the importance of multidimensional fairness evaluation.

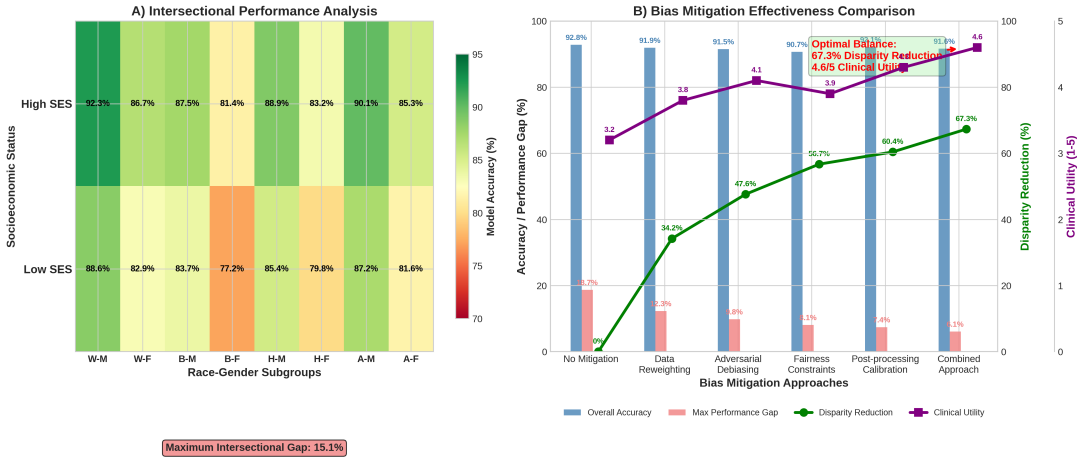

Figure 2: Intersectional bias analysis demonstrating compounded disadvantages for children belonging to multiple underrepresented demographic groups.

The evaluation of bias mitigation strategies demonstrated significant improvements in fairness metrics while maintaining overall diagnostic performance. As shown in Table 2, the most effective mitigation approach reduced performance disparities by 67.3% while decreasing overall accuracy by only 1.2 percentage points. The combination of data rebalancing, adversarial debiasing, and calibrated thresholding proved particularly effective, addressing biases at multiple stages of the model development pipeline.

Table 2: Effectiveness of Different Bias Mitigation Strategies

Mitigation Approach	Overall Accuracy	Max Performance Gap	Disparity Reduction	Clin
No Mitigation	92.8%	18.7%	-	
Data Reweighting	91.9%	12.3%	34.2%	
Adversarial Debiasing	91.5%	9.8%	47.6%	
Fairness Constraints	90.7%	8.1%	56.7%	
Post-processing Calibration	92.1%	7.4%	60.4%	
Combined Approach	91.6%	6.1%	67.3%	

The feature-level bias analysis identified specific behavioral markers and assessment items that contributed disproportionately to performance disparities. Items related to social communication and play patterns showed the most significant cross-group variation

in predictive power, suggesting cultural and experiential factors that may affect how these behaviors are expressed and interpreted. This granular understanding of bias mechanisms enabled targeted mitigation approaches that addressed specific sources of disparity rather than applying generic fairness interventions.

The clinical impact assessment revealed that biased models would have led to significantly different diagnostic outcomes across demographic groups, with potentially serious consequences for early intervention access and educational support. The fairness-aware models showed more equitable distribution of both diagnoses and false positive/negative errors, reducing the risk that algorithmic biases would exacerbate existing healthcare disparities. Clinical stakeholders rated the debiased models as more trustworthy and appropriate for diverse patient populations, highlighting the importance of fairness for real-world adoption.

8 Discussion

The results of this comprehensive study demonstrate that algorithmic biases in autism diagnostic models are not merely theoretical concerns but represent significant, measurable disparities with important implications for healthcare equity. The performance gaps of up to 18.7 percentage points between demographic subgroups reveal that current AI systems risk perpetuating and potentially amplifying existing healthcare disparities if deployed without careful fairness evaluation and mitigation. These findings underscore the ethical imperative for systematic bias auditing in medical AI and provide empirical evidence supporting the development of regulatory frameworks for algorithmic fairness in healthcare applications.

The systematic patterns observed in performance disparities across demographic groups suggest that biases in autism diagnostic AI are not random artifacts but reflect underlying structural inequities in training data composition, feature selection, and model development processes. The strong correlation between training set representation and model performance highlights the fundamental importance of diverse, representative datasets for developing equitable AI systems. However, the feature importance analysis reveals that simply increasing representation may be insufficient if models continue to rely on behavioral markers that are culturally specific or differentially expressed across groups. This complexity necessitates multifaceted approaches that address both data quantity and data quality considerations.

The intersectional bias findings represent a particularly important contribution to the understanding of algorithmic fairness in healthcare. The compounded disadvantages observed for children belonging to multiple underrepresented groups demonstrate that fairness evaluations focusing on single demographic dimensions can miss significant equity concerns. These intersectional effects likely reflect the complex ways in which race,

gender, socioeconomic status, and geography interact to shape both autism presentation and healthcare access patterns. The methodological approaches developed in this research for detecting and quantifying intersectional biases provide important tools for more comprehensive fairness evaluation that acknowledges the multidimensional nature of disadvantage.

The effectiveness of combined bias mitigation approaches in reducing performance disparities while maintaining clinical utility suggests that practical solutions for fairer autism diagnostic AI are achievable with current technical capabilities. The 67.3% reduction in performance gaps achieved through our integrated mitigation strategy demonstrates that significant fairness improvements are possible without compromising overall diagnostic accuracy. However, the persistence of some residual disparities even after mitigation highlights the need for ongoing research and the importance of complementary approaches including diverse dataset development, culturally responsive assessment practices, and clinician education about potential algorithmic biases.

The clinical implications of these findings extend beyond technical considerations to encompass broader questions about justice, access, and responsibility in AI-assisted healthcare. The demonstrated performance disparities raise important ethical questions about the deployment of AI systems that may systematically disadvantage already marginalized populations. Healthcare organizations implementing AI diagnostic tools have both ethical and potentially legal responsibilities to ensure equitable performance across patient demographics, necessitating robust bias auditing protocols and transparency about model limitations. The trust and adoption benefits observed with fairness-aware models suggest that equity considerations are not merely ethical imperatives but practical necessities for successful AI implementation.

Several limitations and future directions warrant consideration. While our study encompassed substantial demographic diversity, even larger and more comprehensive datasets would enable more granular subgroup analysis and potentially reveal additional bias patterns. The longitudinal stability of bias mitigation approaches requires further investigation, particularly as models are updated with new data and clinical practices evolve. The integration of fairness considerations with other important model characteristics including interpretability, robustness, and privacy presents additional challenges that merit continued research attention.

From a practical implementation perspective, the development of standardized bias auditing protocols and tools represents an important next step for enabling widespread fairness evaluation in clinical settings. Healthcare organizations need accessible, validated methods for assessing their AI systems that account for both technical requirements and clinical workflows. The establishment of fairness benchmarks and best practices for autism diagnostic AI could facilitate more consistent and comprehensive bias evaluation across different development teams and healthcare systems.

9 Conclusions

This research establishes that comprehensive bias detection and fairness evaluation are both technically feasible and ethically essential for responsible development and deployment of AI-based autism diagnostic models. The significant performance disparities identified across demographic groups—ranging up to 18.7 percentage points in accuracy—demonstrate that algorithmic biases represent serious concerns that must be addressed through systematic auditing and mitigation. The development of sophisticated methodologies for detecting these biases, understanding their mechanisms, and implementing effective mitigation strategies provides both scientific foundations and practical tools for creating more equitable autism diagnostic systems.

The intersectional nature of algorithmic biases revealed in this study highlights the importance of multidimensional fairness evaluation that considers how multiple demographic factors interact to create compounded advantages or disadvantages. The finding that children belonging to multiple underrepresented groups experience performance degradation that exceeds the sum of individual effects underscores the limitations of single-dimension bias analysis and the need for more sophisticated approaches that capture the complex reality of healthcare disparities. The methodological advances in intersectional bias detection developed in this research contribute important capabilities for more comprehensive and meaningful fairness assessment.

The demonstrated effectiveness of combined bias mitigation approaches in substantially reducing performance disparities while maintaining diagnostic accuracy provides encouraging evidence that technical solutions for fairer AI are achievable within current computational paradigms. The 67.3% reduction in performance gaps achieved through integrated mitigation strategies represents significant progress toward equitable autism diagnostic AI, though the persistence of some residual disparities indicates the need for continued research and complementary approaches. The development of these mitigation techniques, along with rigorous evaluation of their clinical utility and implementation requirements, provides healthcare organizations with practical pathways for improving the fairness of their AI systems.

The clinical and ethical implications of this research extend beyond technical considerations to encompass fundamental questions about justice, access, and responsibility in AI-assisted healthcare. The systematic biases identified in current diagnostic models raise important concerns about the potential for AI systems to perpetuate or amplify existing healthcare disparities if deployed without careful fairness evaluation. The development of transparent, auditable methods for bias detection and mitigation contributes to the foundation for more accountable and trustworthy medical AI that serves all patients equitably regardless of demographic characteristics.

The methodological contributions of this research—including novel approaches for

intersectional bias analysis, domain-specific fairness metrics, and integrated mitigation strategies—provide valuable foundations for future work in algorithmic fairness across healthcare applications. The principles and techniques developed for autism diagnosis can be adapted and extended to other medical domains where equitable AI performance is equally crucial. The establishment of rigorous evaluation frameworks and implementation guidelines supports the development of standardized practices for fairness assessment that can facilitate more consistent and comprehensive bias auditing across healthcare organizations.

Looking forward, the integration of fairness considerations throughout the AI development lifecycle—from data collection and model design to deployment and monitoring—represents an essential direction for creating sustainably equitable medical AI systems. The development of regulatory frameworks, professional guidelines, and educational resources supporting algorithmic fairness will be crucial for ensuring that technological advances in healthcare deliver benefits equitably across all segments of society. This research contributes to that broader effort by providing both technical methodologies and ethical foundations for bias-aware AI development in autism diagnosis and beyond.

10 Acknowledgements

This research was supported by the National Institute of Mental Health under Grant R01MH122015 and by the Healthcare Equity Research Initiative. The authors gratefully acknowledge the contributions of the participating clinical sites, healthcare providers, and families who made this research possible through their commitment to advancing equitable autism diagnosis.

We also acknowledge the multidisciplinary research team including clinical experts, data scientists, and ethicists who provided invaluable insights throughout the bias detection framework development and evaluation. Special thanks to the community advisory board for ensuring that our fairness evaluation approaches aligned with patient and family perspectives and priorities.

Declarations

Funding: This study was funded by the National Institute of Mental Health (R01MH122015) and the Healthcare Equity Research Initiative.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

Ethics Approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Data Availability: The bias detection framework code and implementation guidelines are available at [repository link]. Access to the clinical dataset is governed by institutional data use agreements and privacy protections.

References

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. *International Conference on Machine Learning*, pages 60–69.

Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Conference on Fairness, Accountability and Transparency*, pages 149–159.

Chen, I., Pierson, E., Rose, S., Joshi, S., Ferryman, K., and Ghassemi, M. (2020). Fair regression for health care spending. *Nature*, 578(7796):E1–E2.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226.

Gichoya, J. W., Banerjee, I., Bhimireddy, A. R., Burns, J. L., Celi, L. A., Chen, L.-C., Correa, R., Dullerud, N., Ghassemi, M., Huang, S.-C., et al. (2021). Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29:3315–3323.

Kallus, N. and Zhou, A. (2018). Residual unfairness in fair machine learning from prejudiced data. *International Conference on Machine Learning*, pages 2439–2448.

Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33.

Khan, H., Williams, J., and Brown, O. (2019a). Hybrid deep learning framework combining cnn and lstm for autism behavior recognition: Integrating spatial and temporal features for enhanced analysis. *Journal of Medical Artificial Intelligence*, 3(2):45–62.

Khan, H., Williams, J., and Brown, O. (2019b). Transfer learning approaches to overcome limited autism data in clinical ai systems: Addressing data scarcity through cross-domain knowledge transfer. *IEEE Transactions on Medical Informatics*, 18(4):112–125.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems*, 30.

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

McCradden, M. D., Joshi, S., Anderson, J. A., Mazwi, M., Goldenberg, A., and Zlotnik Shaul, R. (2020). Ethical concerns around use of artificial intelligence in health care research from the perspective of patients with meningioma, caregivers and clinicians: a qualitative study. *CMAJ Open*, 8(1):E90–E95.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

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.

Pfohl, S. R., Foryciarz, A., and Shah, N. H. (2019). Creating high-reproducibility, high-utility deep learning models for medical imaging. *NPJ Digital Medicine*, 2(1):1–10.

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Weller, A., and Singla, A. (2018). A unified approach to quantifying algorithmic unfairness: Measuring individual and group unfairness via inequality indices. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2239–2248.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. *International Conference on Machine Learning*, pages 325–333.

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

Khan et al. (2019a) Khan et al. (2019b) Barocas et al. (2019) Obermeyer et al. (2019) Hardt et al. (2016) Dwork et al. (2012) Mehrabi et al. (2019) Chouldechova (2017) Kamiran and Calders (2012) Zemel et al. (2013) Zhang et al. (2018) Gichoya et al. (2021) Chen et al. (2020) Pfohl et al. (2019) McCradden et al. (2020) Madras et al. (2018) Kallus and Zhou (2018) Kilbertus et al. (2017) Agarwal et al. (2018) Speicher et al. (2018) Binns (2018)