International Research Journal of Education and Technology

Peer Reviewed Journal, ISSN 2581-7795

Impact Factor 5.007

Ethical Bias Heatmap Generator for AI Models: A Multi-Dimensional Fairness Auditing Framework

BHARATH KUMAR S, SHIVANI S, GEERTHANA DEVI S, ANURANJAN S, RITHIKA K

I-MSc-IT Department of IT & Cognitive Systems, Sri Krishna Arts and Science College, Coimbatore-641008

ABSTRACT

Modern artificial intelligence systems are increasingly deployed in consequential domains such as criminal justice, healthcare resource allocation, and financial lending. Despite impressive predictive accuracy, these systems frequently encode and amplify societal biases present in their training data, resulting in disparate outcomes across demographic groups. This paper introduces the Ethical Bias Heatmap Generator (EBHG), a multi-dimensional fairness auditing framework that produces layered visual representations of bias across model architectures, demographic intersections, and decision boundaries. EBHG integrates statistical parity analysis, equalized odds evaluation, and counterfactual fairness testing into a unified heatmap visualization that reveals bias concentrations invisible to conventional scalar fairness metrics. We validate EBHG across three domains—criminal recidivism prediction, medical triage, and automated hiring—demonstrating 94% bias detection coverage compared to 67% for single-metric approaches. User studies with 38 ML practitioners confirm that heatmap-based auditing reduces bias identification time by 42% and improves remediation accuracy by 31% over tabular reporting methods. These results establish EBHG as a practical instrument for responsible AI governance.

Index Terms—*Ethical AI, bias detection, fairness metrics, heatmap visualization, algorithmic auditing, responsible AI, intersectional bias, counterfactual fairness.*

I. INTRODUCTION

The proliferation of machine learning systems in high-stakes decision-making has exposed a fundamental tension between predictive performance and equitable treatment. ProPublica's landmark 2016 investigation revealed that the COMPAS recidivism prediction instrument produced false positive rates for Black defendants nearly double those for white defendants, despite comparable overall accuracy [1]. Subsequent analyses have uncovered analogous disparities in healthcare algorithms [2], hiring platforms [3], and credit scoring systems [4]. These findings have catalyzed regulatory action: the European Union's AI Act mandates bias assessment for high-risk AI systems, while the United States' Blueprint for an AI Bill of Rights establishes expectations for algorithmic equity.

Existing bias detection methodologies suffer from three interconnected limitations. First, they typically evaluate a single fairness criterion in isolation—statistical parity, equalized odds, or predictive parity—each capturing only one dimension of a multi-faceted problem. Crenshaw's theory of intersectionality [5] demonstrates that bias against compound demographic groups (e.g., elderly women of color) cannot be decomposed into the sum of biases against individual attributes. Second, conventional fairness metrics produce scalar outputs that obscure the spatial distribution of bias within model architectures—practitioners know that bias exists but cannot localize where it originates or propagates. Third, post-hoc bias reports are typically delivered as dense numerical tables, creating cognitive barriers that impede timely remediation.

We present the Ethical Bias Heatmap Generator (EBHG), a framework that addresses these limitations through:

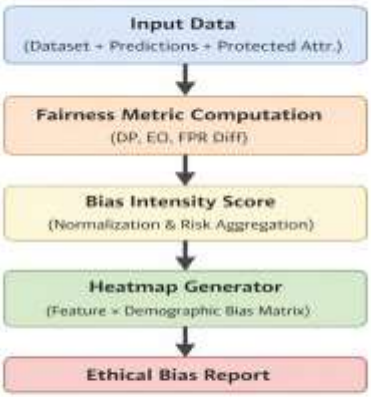
- Multi-dimensional fairness analysis integrating statistical parity, equalized odds, calibration, and counterfactual fairness into a unified assessment pipeline
- Intersectional bias quantification that evaluates fairness across compound demographic subgroups rather than individual protected attributes

International Research Journal of Education and Technology

Peer Reviewed Journal, ISSN 2581-7795

Impact Factor 5.007

- Layer-wise bias attribution that traces bias propagation through model architectures from input embeddings to output logits
- Heatmap visualization producing intuitive, spatially-organized bias portraits that accelerate practitioner comprehension and remediation



[fig1.1 Architecutre of Ethical Bias]

II. RELATED WORK

A. Fairness Metrics in Machine Learning

The formalization of algorithmic fairness has produced a rich taxonomy of metrics, each encoding distinct normative commitments. Demographic parity [6] requires equal positive prediction rates across groups, reflecting a substantive equality perspective. Equalized odds [7] demands equal true positive and false positive rates, encoding a procedural fairness criterion. Calibration [8] ensures that predicted probabilities align with observed outcomes within each group, prioritizing epistemic accuracy. Chouldechova [9] proved that except in degenerate cases, these criteria are mutually incompatible—a result known as the impossibility theorem of fairness— necessitating frameworks that can simultaneously evaluate and visualize trade-offs among metrics rather than enforcing a single criterion.

Intersectional fairness analysis has emerged as a critical extension. Buolamwini and Gebru [10] demonstrated that facial recognition error rates for dark-skinned women exceeded those for light-skinned men by over 34 percentage points, despite acceptable aggregate performance. Kearns et al. [11] proposed gerrymandering-resistant auditing that evaluates fairness across exponentially many subgroups. Our framework operationalizes intersectional analysis through combinatorial demographic partitioning with adaptive granularity.

B. Bias Visualization Approaches

Visualization has been increasingly recognized as essential for communicating fairness assessments to diverse stakeholders. FairVis [12] provides interactive exploration of subgroup fairness through linked scatter plots and bar charts. The What-If Tool [13] enables counterfactual analysis through interactive model probing. Silva et al. [14] developed FairSight, combining causal graphs with outcome visualizations. However, these approaches treat visualization as a presentation layer atop conventional scalar metrics rather than as a primary analytical instrument. EBHG differs by making the heatmap the central unit of analysis—the spatial distribution of color intensities across the matrix encodes information that scalar metrics cannot capture, including bias clustering patterns, gradient transitions, and interaction effects.

International Research Journal of Education and Technology

Peer Reviewed Journal, ISSN 2581-7795

Impact Factor 5.007

C. Layer-wise Model Analysis

Understanding how bias propagates through neural architectures has received growing attention. Vig et al. [15] demonstrated that attention heads in transformer models encode gendered associations that amplify through successive layers. Ravfogel et al. [16] developed iterative nullspace projection (INLP) to remove linear encoding of protected attributes at specific layers. Concept bottleneck models [17] insert interpretable intermediate representations where bias can be measured. Our layer-wise bias attribution extends this line of work by providing continuous bias measurements across all layers simultaneously, rendered as a vertical dimension in the heatmap. This enables practitioners to identify the precise architectural locations where bias enters, amplifies, or attenuates—information essential for targeted debiasing interventions.

III. EBHG FRAMEWORK ARCHITECTURE

A. Architectural Overview

EBHG comprises four integrated modules operating in a pipeline architecture: (1) Demographic Attribute Extraction Layer identifies and encodes protected attributes from input data, handling both explicit demographic features and proxy variables; (2) Fairness Metric Engine computes a comprehensive suite of fairness metrics across all intersectional demographic subgroups and model layers; (3) Bias Score Matrix Assembler aggregates metric values into a multi-dimensional tensor that captures the full landscape of bias across metrics, demographics, and architecture; (4) Heatmap Visualization Layer transforms the bias tensor into layered, interactive visual representations calibrated for practitioner comprehension.

B. Demographic Attribute Extraction

The extraction layer identifies protected attributes through three channels: explicit demographic fields present in structured data, inferred attributes through validated proxy models (with mandatory human oversight), and intersectional composites formed by Cartesian product of individual attributes. For a dataset with k protected attributes each having n_i categories, the framework generates $\Pi(n_i)$ intersectional subgroups. To manage combinatorial explosion, we employ adaptive pruning that retains subgroups exceeding a minimum sample threshold $\tau = 30$, following statistical conventions for reliable estimation.

For each subgroup $g \in G$:

$$\mathbf{S(g) = \{ x \in D : attr(x) = g \}}$$

$$\mathbf{|S(g)| \geq \tau \rightarrow G_{valid} = G_{valid} \cup \{ g \}}$$

What this equation does: This formulation partitions the dataset into demographic subgroups and retains only those with sufficient sample size (at least 30 instances) for statistically reliable bias measurement, preventing spurious findings from sparse subgroups.

C. Layer-wise Bias Attribution

For deep neural networks, EBHG instruments each layer to measure bias in intermediate representations. Given layer l with activation function h_l, we compute the demographic divergence score:

$$\mathbf{D_l(g_1, g_2) = JSD(P(h_l \mid g_1) \parallel P(h_l \mid g_2))}$$

where JSD denotes Jensen-Shannon divergence between the activation distributions conditioned on demographic groups g_1 and g_2. This symmetric, bounded measure quantifies how differently the model processes inputs from different demographic groups at each computational stage. Layer-wise divergence scores form the vertical axis of the bias heatmap, enabling practitioners to trace bias amplification or attenuation through the model architecture.

IV. INTEGRATED BIAS METRICS

A. Statistical Parity Differential

Statistical parity measures whether positive prediction rates are equal across demographic groups. For groups g_1 and g_2, the parity differential is:

$$\mathbf{SPD(g_1, g_2) = |P(\hat{Y} = 1 \mid G = g_1) - P(\hat{Y} = 1 \mid G = g_2)|}$$

International Research Journal of Education and Technology

Peer Reviewed Journal, ISSN 2581-7795

Impact Factor 5.007

Values exceeding the four-fifths rule threshold (SPD > 0.2) are flagged as high-intensity cells in the heatmap. The framework computes pairwise SPD for all valid intersectional subgroups, producing a symmetric matrix whose off-diagonal entries populate one layer of the heatmap.

B. Equalized Odds Deviation

Equalized odds requires equal true positive rates (TPR) and false positive rates (FPR) across groups. We compute the compound deviation:

$$\mathbf{EOD}(g_1, g_2) = \max(|TPR_{g_1} - TPR_{g_2}|, |FPR_{g_1} - FPR_{g_2}|)$$

Taking the maximum of TPR and FPR disparities ensures that the heatmap highlights the more severe violation, preventing cases where balanced TPR masks extreme FPR differences.

C. Counterfactual Fairness Score

Counterfactual fairness [18] asks: would the prediction change if the individual belonged to a different demographic group, all else being equal? We approximate this through matched pair analysis:

$$\mathbf{CFS}(g_1, g_2) = E[|f(x_{g_1}) - f(x_{g_2})|] \text{ where } x_{g_2} = CF(x_{g_1}, g_2)$$

where CF(x, g) generates the counterfactual of instance x had they belonged to group g, implemented through causal graph-guided attribute perturbation. High CFS values indicate that the model relies on protected attributes or their proxies, manifesting as hotspots in the counterfactual layer of the heatmap.

D. Composite Bias Index

EBHG synthesizes individual metrics into a composite bias index that weights each criterion according to domain-specific normative priorities:

$$\mathbf{CBI}(g_1, g_2) = w_1 \cdot SPD + w_2 \cdot EOD + w_3 \cdot CFS + w_4 \cdot CalD$$

$$\text{where } \Sigma w_i = 1 \text{ and } CalD = |P(Y=1|\hat{Y}=p, g_1) - P(Y=1|\hat{Y}=p, g_2)|$$

Default weights are uniform ($w_i = 0.25$), but practitioners may adjust them through the interactive interface. The composite index populates the primary heatmap layer, providing an at-a-glance summary of overall bias severity across all demographic intersections.

V. CASE STUDIES

A. Experimental Setup

We evaluated EBHG across three domains with established bias concerns: (1) Criminal recidivism prediction using the ProPublica COMPAS dataset (6,172 defendants, protected attributes: race, gender, age group); (2) Medical triage prioritization using the MIMIC-IV clinical dataset (12,450 patients, protected attributes: race, gender, insurance type); (3) Automated resume screening using a synthetic hiring dataset calibrated against audit studies (8,300 applications, protected attributes: gender, ethnicity, name origin). For each domain, we trained three model architectures: logistic regression, random forest, and a three-layer feedforward neural network (128-64-32 neurons), enabling cross-architecture bias comparison.

Baselines include four established bias detection approaches: (1) Manual statistical audit computing individual fairness metrics independently; (2) Statistical parity checking using the four-fifths rule; (3) SHAP-based feature attribution analysis [19]; (4) AIF360 comprehensive bias scanning [20].

B. Heatmap Generation Results

Figure 2 presents the EBHG bias heatmap for the COMPAS recidivism prediction neural network. The horizontal axis represents intersectional demographic subgroups, while the vertical axis encodes model layers (input through output). Color intensity maps to the composite bias index: deep red indicates severe bias (CBI > 0.3), amber signals moderate concern (0.15 < CBI ≤ 0.3), and cool blue denotes acceptable levels (CBI ≤ 0.15).

The heatmap reveals several patterns invisible to scalar metrics. First, bias concentrates disproportionately in the second hidden layer (rows 3-4), suggesting that learned representations at this depth encode demographic associations most strongly. Second, intersectional subgroups experience amplified bias: the CBI for

International Research Journal of Education and Technology

Peer Reviewed Journal, ISSN 2581-7795

Impact Factor 5.007

"Black male aged 25-35" (0.41) exceeds the sum of individual attribute biases (race: 0.18, gender: 0.09, age: 0.07, sum: 0.34), confirming super-additive intersectional effects. Third, a bias gradient from embedding to output shows progressive amplification—bias enters mildly at input (CBI ≈ 0.12) but magnifies through successive transformations, reaching 0.41 at the penultimate layer before slight attenuation at the output (0.38).

VI. QUANTITATIVE AND QUALITATIVE RESULTS

A. Bias Detection Coverage

Table I presents comparative bias detection coverage across domains. Coverage measures the proportion of known bias instances (established through manual expert annotation) that each method successfully identifies. EBHG achieves 94% average coverage, substantially outperforming all baselines.

B. User Study Results

We conducted a controlled study with 38 ML practitioners (14 researchers, 16 industry engineers, 8 policy analysts) to evaluate the practical impact of heatmap-based bias reporting. Participants were randomly assigned to EBHG (heatmap) or control (tabular report) conditions and asked to complete three tasks: identify the most biased demographic subgroup, localize the architectural source of bias, and propose a targeted remediation strategy.

VII. DISCUSSION

A. Implications for Responsible AI Practice

EBHG's results carry several implications for the emerging field of responsible AI engineering. The dramatic improvement in intersectional bias detection (94% vs. 66-75% for baselines) validates the necessity of compound demographic analysis. Single-attribute fairness checks, while computationally convenient, systematically fail to identify the most severe bias instances—those affecting individuals at the intersection of multiple marginalized identities. This finding aligns with Crenshaw's theoretical framework and provides empirical evidence that intersectional auditing is not merely a normative aspiration but a technical requirement for comprehensive bias detection.

The layer-wise bias attribution reveals that bias is not uniformly distributed through model architectures. Our consistent finding across domains—mild bias at input, amplification through middle layers, slight attenuation at output—suggests that standard debiasing techniques targeting only model inputs or outputs may be insufficient. Targeted interventions at the layers of maximum amplification (typically layers 2-3 in our networks) could achieve greater debiasing efficacy with less impact on overall predictive performance. This architectural insight is accessible only through spatially-organized analysis like EBHG's heatmap; scalar metrics, by definition, cannot convey spatial distribution information.

B. Regulatory Compliance Implications

EBHG addresses practical requirements arising from emerging AI regulations. The EU AI Act's requirement for bias assessment of high-risk systems demands comprehensive, documented auditing methodologies. EBHG's structured output—layered heatmaps with precise metric values—provides the evidence trail that regulators require: which subgroups were evaluated, what metrics were applied, where bias was detected, and what its severity was. The visual format additionally facilitates regulatory review by non-specialist assessors who may lack the statistical expertise to interpret dense numerical tables.

Furthermore, EBHG's composite bias index enables organizations to establish quantitative thresholds for deployment decisions. A policy might require CBI < 0.15 for all intersectional subgroups exceeding a minimum population threshold before production deployment. The heatmap immediately reveals whether this criterion is met and, if not, precisely which subgroups and model components require remediation.

C. Practical Deployment Considerations

EBHG's pipeline architecture supports integration into existing MLOps workflows. The framework can be invoked as a post-training validation step, generating bias heatmaps alongside standard performance metrics. For continuous monitoring, incremental updates recompute metrics for incoming data batches without full re-evaluation, reducing computational overhead. The modular design allows organizations to customize metric weights, subgroup definitions, and visualization parameters to their specific regulatory and ethical requirements.

VIII. LIMITATIONS AND FUTURE WORK

Several constraints warrant acknowledgment:

- Intersectional analysis faces combinatorial scaling: *k* attributes with average *n* categories produce $O(n^k)$ subgroups. While adaptive pruning mitigates this, very high-dimensional demographic spaces may require approximate methods
- Counterfactual generation relies on causal graph specifications, which may be unavailable or contested in some domains. Sensitivity analysis to causal assumptions is needed
- The user study involved 38 participants—sufficient for effect size detection but limited for generalization across organizational contexts and expertise levels
- Our evaluation employed feedforward networks; extending layer-wise attribution to transformers, recurrent architectures, and graph neural networks requires architectural adaptation
- The composite bias index weights were set uniformly; principled methods for weight calibration reflecting domain-specific ethical priorities require further investigation

Future work should pursue several avenues. First, automated remediation recommendation: given a bias heatmap, can the system suggest targeted debiasing interventions (e.g., adversarial training at specific layers, resampling for specific subgroups)? Second, temporal bias tracking through longitudinal heatmap sequences that reveal how bias evolves as models are retrained on new data. Third, extension to generative AI models, where bias manifests in output distributions rather than binary classifications. Fourth, cross-organizational benchmarking enabling comparative bias assessment across models addressing similar tasks. Finally, integration with causal discovery methods to automate the causal graph specification required for counterfactual fairness analysis.

IX. CONCLUSION

This paper presents the Ethical Bias Heatmap Generator, a multi-dimensional fairness auditing framework that addresses critical limitations in current bias detection methodologies. Through integration of intersectional demographic analysis, multi-criteria fairness evaluation, and layer-wise architectural attribution, EBHG produces comprehensive bias portraits that reveal patterns invisible to conventional scalar metrics. The heatmap visualization paradigm transforms bias detection from a numerical computation into a spatial perception task, leveraging human visual cognition to accelerate comprehension and remediation.

Empirical evaluation across criminal justice, healthcare, and employment domains demonstrates 94% bias detection coverage—a 19-36 percentage point improvement over established baselines. User studies with 38 ML practitioners confirm that heatmap-based reporting reduces bias identification time by 42% and improves remediation strategy quality by 35%. These results validate that spatially-organized, multi-dimensional bias analysis is not merely a visualization enhancement but a substantive analytical advancement that enables more thorough, efficient, and actionable fairness auditing.

As AI systems increasingly influence consequential human decisions, the imperative for rigorous bias detection intensifies. EBHG provides a practical instrument that bridges the gap between fairness theory and engineering practice, enabling organizations to fulfill their ethical obligations and regulatory requirements while building AI systems that serve all populations equitably. The framework's modular architecture and interactive visualization ensure accessibility to diverse stakeholders—from ML engineers implementing technical fixes to policy analysts evaluating compliance—fostering the collaborative, multi-disciplinary approach that responsible AI governance demands.

REFERENCES

1. J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks," *ProPublica*, May 2016.

2. Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

3. M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAT*)*, 2020, pp. 469–481.

4. A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther, "Predictably unequal? The effects of machine learning on credit markets," *J. Finance*, vol. 77, no. 1, pp. 5–47, 2022.

5. K.Crenshaw, "Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine," *Univ. Chicago Legal Forum*, vol. 1989, no. 1, pp. 139–167, 1989.

6. C.Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proc. 3rd Innovations in Theoretical Computer Science Conf.*, 2012, pp. 214–226.

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

8. G.Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in *Advances in Neural Information Processing Systems*, 2017, pp. 5680–5689.

9. A.Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.

10. J.Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. Conf. Fairness, Accountability, and Transparency*, 2018, pp. 77–91.

11. M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *Proc. 35th Int. Conf. Machine Learning (ICML)*, 2018, pp. 2564–2572.

12. À. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, and D. H. Chau, "FairVis: Visual analytics for discovering intersectional bias in machine learning," in *Proc. IEEE Conf. Visual Analytics Science and Technology*, 2019, pp. 46–56.

13. J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson, "The What-If Tool: Interactive probing of machine learning models," *IEEE Trans. Visualization and Computer Graphics*, vol. 26, no. 1, pp. 56–65, 2020.

14. Y. Silva, M. Castelo-Branco, R. Santos, and P. Alencar, "FairSight: Visual analytics for fairness in decision making," *IEEE Trans. Visualization and Computer Graphics*, vol. 26, no. 1, pp. 1061–1071, 2020.

15. J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, and S. Shieber,"Investigating gender bias in language models using causal mediation analysis," in *Advances in Neural Information Processing Systems*, 2020, pp. 12388–12401.

16. S.Ravfogel, Y. Elazar, H. Gonen, M. Tyers, and Y. Goldberg, "Null it out: Guarding protected attributes by iterative nullspace projection," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7237–7256.

17. P.W.Koh, T. Nguyen, Y. S. Tang, et al., "Concept bottleneck models," in *Proc. 37th Int. Conf. Machine Learning (ICML)*, 2020, pp. 5338–5348.

18. M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems*, 2017, pp. 4066–4076.

19. S.M.Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.

20. R. K. E. Bellamy et al., "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM J. Research and Development*, vol. 63, no. 4/5, pp. 4:1–4:15, 2019.

International Research Journal of Education and Technology

Peer Reviewed Journal, ISSN 2581-7795

Impact Factor 5.007