

A Data-Centric Framework for Detecting and Mitigating Bias in Machine Learning Datasets: Metrics, Auditing, Rebalancing, and Validation

Dr. Nitin Mishra¹, Ms. Swati Atre²

¹Professor, Department of Computer Science & Engineering, Parul University, Vadodara, India

Email: drnitinmishra10@gmail.com

²Assistant Professor, Department of Computer Science, Dr. C.V. Raman University, Khandwa (M.P.) and PhD Scholar, SAM Global University, Bhopal (M.P.)

Email: atre.swati@gmail.com

Abstract— Machine learning systems can inherit, reproduce, and amplify systematic errors already present in their training data. Although algorithmic fairness has produced a large body of mitigation methods, dataset-level bias is often treated as a preprocessing detail rather than as a measurable quality problem. This paper develops a data-centric framework for detecting and mitigating machine learning dataset bias before model deployment. The framework integrates five stages: (i) dataset profiling and provenance inspection, (ii) representation-bias measurement, (iii) label, sampling, and distribution-shift diagnostics, (iv) mitigation through reweighting, resampling, relabeling, augmentation, or targeted collection, and (v) stratified validation using fairness, utility, robustness, and subgroup generalization metrics. A taxonomy of bias sources is mapped to appropriate detection signals, including prevalence ratios, entropy and divergence measures, conditional error gaps, missingness differentials, label-quality audits, and train–test distribution comparisons. The proposed framework also emphasizes dataset documentation, human review, sensitivity analysis, and external validation because statistical parity alone cannot establish that a dataset is socially or scientifically appropriate. The contribution is a practical synthesis rather than a claim of a new empirically validated algorithm; it is intended to guide researchers and practitioners in building reproducible bias-auditing pipelines across tabular, image, text, and multimodal datasets.

Keywords— dataset bias, representation bias, machine learning fairness, data-centric AI, bias detection, data auditing, dataset shift, reweighting, subgroup evaluation, responsible AI

1. Introduction

Machine learning (ML) systems are commonly evaluated by predictive accuracy, but accuracy can conceal systematic differences in performance or representation across demographic, geographic, linguistic, socioeconomic, or clinical groups. A model can therefore appear strong on an aggregate test set while being unreliable or unfair for subpopulations that are rare, absent, systematically mislabeled, or sampled through a different process [1], [2]. This problem is not confined to the learning algorithm. Bias may enter during problem formulation, data collection, sampling, annotation, feature construction, preprocessing, train–test splitting, and downstream deployment [1], [3].

Recent literature distinguishes between fairness of the learned model and bias embedded in the data itself. Representation bias, for example, arises when the data-generating process inadequately represents relevant groups; it can result from selection effects, historical inequity, or sampling procedures [4]. A systematic survey of representation bias argues that dataset-level identification and resolution deserve treatment as a problem

independent of the final model [4]. Similarly, work on fairness datasets shows that widely reused benchmarks may have documentation, inclusivity, consent, labeling, and external-validity limitations [5]. These concerns are especially important when benchmark results are interpreted as evidence of general fairness.

This paper asks a practical question: *How can researchers detect and correct bias in a dataset before it becomes a model problem?* The answer proposed here is a structured data-centric workflow that links bias type, measurable signal, intervention, and validation criterion. Unlike a single fairness score, the framework treats bias as multidimensional. It also recognizes that fairness metrics can conflict and that the appropriate criterion depends on the use case, target population, causal context, and harm model [2], [6].

The principal contributions are: (1) a unified taxonomy of dataset bias relevant to contemporary ML pipelines; (2) a measurable audit matrix covering representation, label, sampling, feature, and distribution biases; (3) a mitigation decision framework spanning data-level and model-level interventions; and (4) a validation protocol that combines fairness, predictive utility, robustness, documentation, and external generalization.

2. Background and Related Work

A substantial literature has formalized fairness in terms such as demographic parity, equal opportunity, equalized odds, calibration, and individual fairness [6–10]. These definitions are useful, but they do not by themselves reveal why the observed disparity occurs. For example, a gap in false-negative rates may reflect class imbalance, a measurement problem, under-representation, label bias, or genuine differences in conditional risk. Consequently, auditing should precede or accompany model-level fairness optimization.

Preprocessing methods have long included reweighting, resampling, massaging, and selective suppression of attributes [11]. Representation-learning approaches attempt to remove sensitive information while preserving task utility [12], while constrained optimization can incorporate fairness requirements during model training [13]. Post-processing techniques can equalize selected error rates after training [14]. Surveys of bias mitigation show a broad ecosystem of pre-, in-, and post-processing interventions, with substantial variation in datasets, metrics, and evaluation practices [15].

A data-centric perspective is particularly important because training and evaluation populations may differ. Dataset shift can occur because of changing prevalence, covariate distributions, sampling frames, collection devices, geography, or recruitment practices, and this shift can undermine generalization even when conventional validation is strong [16]. Recent work also demonstrates that common benchmark datasets can have limited temporal and geographic external validity [17]. These findings support targeted subgroup validation and repeated monitoring rather than one-time fairness certification.

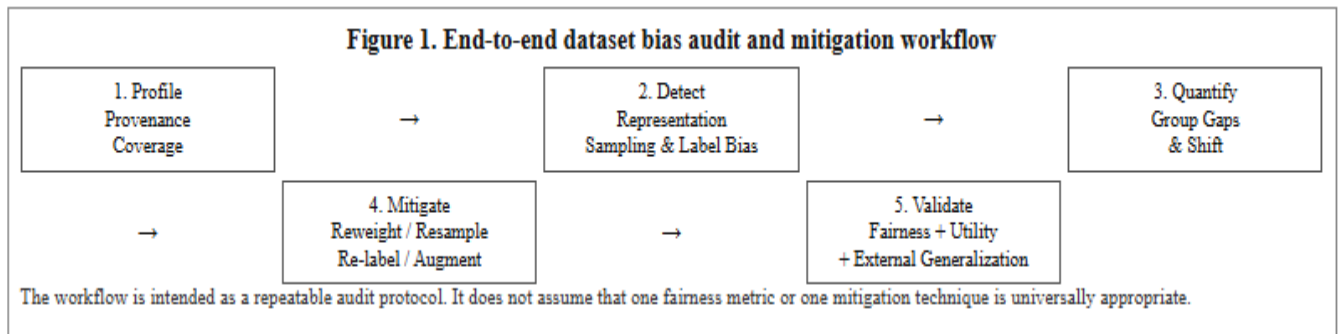
3. Taxonomy of Machine Learning Dataset Bias

Bias source	Typical mechanism	Observable signal	Priority response
Representation bias	Groups absent or under-represented	Prevalence ratios, coverage gaps, entropy	Targeted sampling/collection
Selection/sampling bias	Non-random inclusion or recruitment	Sampling weights, selection odds, subgroup shift	Reweighting, stratification, redesign
Label/annotation bias	Systematic differences in labels	Inter-rater disagreement, label parity, error rates	Re-annotation, adjudication
Measurement bias	Features measured differently by group	Missingness/scale/device differences	Calibration, feature redesign
Historical/social bias	Past inequities encoded in outcomes	Conditional disparity with contextual evidence	Causal review + domain intervention
Aggregation bias	One model/data representation assumed adequate for all groups	Subgroup-specific residual patterns	Stratified modeling/representation
Distribution shift	Training and target populations differ	KL/JS divergence, MMD, PSI, classifier two-sample tests	External data, adaptation, monitoring
Preprocessing bias	Cleaning, imputation, augmentation or filtering affects groups unequally	Before/after subgroup statistics	Pipeline audit + sensitivity analysis

This taxonomy follows the data-centered emphasis in the bias literature [1], [4], [15], [18]. Importantly, the categories can co-occur: under-representation may produce unstable labels; selection effects may create apparent class imbalance; and a measurement problem can look like a label problem if the downstream label is defined from biased measurements.

4. Proposed Data-Centric Bias Detection Framework

Figure 1. End-to-end dataset bias audit and mitigation workflow



The workflow is intended as a repeatable audit protocol. It does not assume that one fairness metric or one mitigation technique is universally appropriate.

4.1 Stage 1: Dataset profiling and provenance

The first stage creates a machine-readable dataset profile containing source, collection window, geographical coverage, population definition, inclusion and exclusion rules, sensitive attributes, missingness, label-generation process, and known transformations. Dataset documentation reduces the risk of treating a benchmark as context-free evidence [5], while “datasheet”-style documentation is useful for communicating intended use and known limitations [19]. Sensitive attributes should be handled according to the legal, ethical, and scientific context; hiding a sensitive attribute does not remove the underlying bias and can prevent measurement.

4.2 Stage 2: Representation and composition diagnostics

For a protected or relevant group g , let n_g be its count and N the total sample size. A simple representation proportion is:

$$R_g = n_g / N$$

For a reference population proportion p_g^* , the representation ratio is:

$$RR_g = R_g / p_g^*$$

Values substantially below 1 indicate under-representation relative to the target population, although an “ideal” proportion cannot be inferred without a defensible target population. The review of demographic dataset metrics recommends selecting a compact set of non-redundant measures rather than indiscriminately reporting every available index [20]. Recommended diagnostics include subgroup prevalence, minimum cell counts, coefficient of variation, entropy of the group distribution, representation ratio, and a diversity index.

4.3 Stage 3: Conditional label and sampling bias

Representation alone is insufficient. The distribution of labels must also be inspected by group. For binary classification, compare positive-label rates:

$$P(Y=1 | A=g)$$

and stratify these rates by clinically, operationally, or scientifically relevant covariates X . Large unadjusted differences can be legitimate or can reflect biased selection; therefore, the audit should compare both marginal and conditional patterns. Label quality should be assessed through multi-annotator agreement, disagreement

concentration by group, adjudication rates, and re-labeling samples. The recent literature on label and selection bias demonstrates that the effectiveness of mitigation depends on the underlying type of bias, rather than on a single universal correction [21].

4.4 Stage 4: Distribution-shift detection

To compare a training distribution P and a target or validation distribution Q , researchers can use distribution distances such as KL divergence, Jensen–Shannon divergence, population stability index, maximum mean discrepancy, or a classifier-based two-sample test. For continuous features, standardized mean differences and quantile comparisons are useful; for categorical variables, chi-square or exact tests can identify compositional changes. Shift should be measured globally and within subgroups. Dataset-shift research shows why ordinary random train–test splits may be insufficient when the deployment population differs from the training cohort [16].

4.5 Stage 5: Cross-check with model-level fairness

After dataset diagnostics, a baseline model should be evaluated by subgroup. At minimum, report accuracy or task-appropriate utility together with false-positive rate (FPR), false-negative rate (FNR), true-positive rate (TPR), positive predictive value (PPV), and calibration where relevant. Common fairness criteria can be expressed as:

Demographic Parity: $P(\hat{Y}=1|A=g) \approx P(\hat{Y}=1|A=h)$

Equal Opportunity: $P(\hat{Y}=1|Y=1,A=g) \approx P(\hat{Y}=1|Y=1,A=h)$

Equalized Odds: $TPR_g \approx TPR_h$ and $FPR_g \approx FPR_h$

These criteria should not be treated as interchangeable. The fairness literature documents structural trade-offs among competing criteria, particularly when base rates differ [7], [8], [22].

5. Bias Detection Metrics and Audit Rules

Audit family	Example metric	What it diagnoses	Interpretation
Representation	Representation ratio	Under/over coverage	Compare with justified target population
Composition	Entropy / Gini / concentration	Dominance by a few groups	Context-dependent; use alongside counts
Association	Chi-square, Cramér’s V	Group–label dependence	Signals require substantive interpretation
Distribution shift	JS divergence / MMD / PSI	Train–target mismatch	Large values indicate review or adaptation
Label quality	Agreement, adjudication gap	Group-specific annotation quality	Low agreement may imply ambiguous constructs
Model error	FPR/FNR/TPR gap	Outcome disparity after learning	Interpret with base rates and use case
Calibration	Calibration error by group	Probability reliability	Important for risk scores and ranking

Robustness	Worst-group performance	Tail-group vulnerability	Prefer reporting minimum/percentile performance
------------	-------------------------	--------------------------	---

A useful audit rule is to avoid a single-threshold “bias/no-bias” decision. Instead, create a severity matrix using (a) statistical magnitude, (b) uncertainty, (c) practical effect size, (d) sample sufficiency, (e) harm severity, and (f) external-validity implications. Bootstrap confidence intervals are recommended for subgroup metrics when sample sizes are limited. Multiple testing corrections should be considered when dozens of features and groups are audited.

6. Mitigation Strategies

6.1 Reweighting and resampling

Reweighting assigns higher importance to under-represented observations and lower weight to over-represented observations. A common population-adjustment form is:

$$w_g = P^*(A=g) / P(A=g)$$

where P^* is the target distribution. Kamiran and Calders demonstrated preprocessing techniques including reweighting, resampling, massaging, and selective suppression [11]. Reweighting preserves observations but may increase variance if weights become extreme; therefore, weight clipping and effective sample-size reporting are advisable.

6.2 Targeted data collection and augmentation

When a subgroup is genuinely absent or severely under-represented, synthetic balancing alone is not a complete solution. Targeted collection from the intended deployment population is usually more defensible. For images, augmentation should preserve clinically or semantically valid attributes; for text, paraphrasing can inadvertently change demographic cues or label semantics. Synthetic data should therefore be audited for memorization, artifacts, and subgroup fidelity before being counted as evidence of diversity.

6.3 Label correction and adjudication

Where label bias is detected, the preferred intervention is to revisit the label-generation process. A blind relabeling sample, independent adjudication, and group-balanced quality-control protocol can estimate whether annotator error differs by subgroup. This is especially important when the target label is a historical decision rather than a direct ground-truth measurement. Correcting labels without understanding the generating process may simply move bias from one variable to another.

6.4 Feature and pipeline correction

Measurement bias can be addressed through standardized instrumentation, device normalization, missingness-aware modeling, or replacement of proxies whose measurement process differs by group. Pipeline auditing should compare subgroup statistics before and after imputation, normalization, outlier removal, feature selection, and augmentation. An intervention is acceptable only if it improves the intended fairness/utility objective without introducing unacceptable new disparities.

6.5 In-processing and post-processing as secondary controls

Data-level mitigation should be complemented—not replaced—by model-level controls when residual disparities remain. Fairness-aware learning, adversarial representation learning, constrained optimization, and threshold

adjustment are established approaches [12–15]. However, model-level correction cannot repair an absent population, a systematically invalid label, or a deployment shift that was never represented in the training data.

7. Proposed Bias Audit Algorithm

The following pseudo-procedure operationalizes the framework. It is deliberately model-agnostic and can be implemented with Python, pandas, scikit-learn, Fairlearn, AIF360, or equivalent tools.

1. Define the target population, intended use, relevant groups, and harm scenarios.
2. Profile provenance, missingness, duplicates, label sources, collection period, and subgroup coverage.
3. Compute representation, class balance, group–label association, and feature-distribution diagnostics.
4. Run train–validation and train–target distribution-shift tests globally and by subgroup.
5. Audit label quality with a stratified human-review sample and inter-rater agreement.
6. Train a transparent baseline model and report subgroup utility, FPR/FNR/TPR, calibration, and worst-group performance.
7. Select mitigation according to the diagnosed bias mechanism; do not choose a method solely because it improves one fairness score.
8. Re-run the complete audit after mitigation and compare utility, fairness, robustness, and uncertainty.
9. Validate on an independent or temporally/geographically shifted cohort whenever feasible.
10. Document unresolved limitations and establish post-deployment monitoring indicators.

8. Tools and Reproducibility

Open-source fairness libraries can reduce implementation errors and make audits easier to reproduce. AI Fairness 360 provides dataset and model fairness metrics and mitigation algorithms in a common toolkit [23]. Dataset documentation research also suggests that practitioners should record task definition, source, demographics, consent, labeling, preprocessing, and known limitations [5], [19]. Reproducibility should therefore include the exact dataset version, preprocessing code, random seeds, sampling strategy, metric definitions, subgroup definitions, and audit outputs.

For practical implementation, a bias-audit report should contain: dataset fingerprint; subgroup counts; target-population assumptions; label-quality results; shift statistics; baseline and mitigated model metrics; confidence intervals; mitigation parameters; and a decision log. This creates an auditable link between the observed bias, chosen intervention, and reported conclusion.

9. Research Gaps and Future Directions

First, current literature still relies heavily on a small set of benchmark fairness datasets, creating a risk of methodological overfitting to familiar tabular tasks. The “Retiring Adult” study illustrates how dataset construction and external validity can limit conclusions drawn from long-standing benchmarks [17]. Second, bias detection

remains fragmented across representation, labels, sampling, and model evaluation; integrated pipelines are comparatively less standardized [4], [15]. Third, fairness metrics are often binary and single-metric, even though real systems may involve multiclass outcomes, intersectional groups, missing sensitive attributes, and continuous risk scores [3].

Fourth, dataset bias can change over time. Monitoring should therefore include temporal drift and subgroup-specific drift, rather than assuming a static training distribution [16], [17]. Fifth, synthetic and generative data raise a new question: can augmentation increase representation while reproducing the same latent stereotypes and measurement artifacts? Finally, there is a need for stronger causal analysis. A statistical disparity can identify a problem, but deciding whether to correct, preserve, or explain the disparity requires domain knowledge, causal assumptions, and stakeholder participation.

10. Conclusion

Detecting ML dataset bias is fundamentally a data-quality, measurement, and population-validity problem as well as a fairness problem. The framework proposed in this paper connects dataset profiling, representation analysis, label-quality auditing, distribution-shift detection, subgroup model evaluation, targeted mitigation, and independent validation in one repeatable workflow. The central principle is that bias should be *measured before it is mitigated*, and mitigation should be chosen according to the mechanism that generated the disparity. A dataset cannot be certified as unbiased by a single fairness metric, nor can a model be considered fair merely because its overall accuracy is high. Robust practice requires documented target populations, transparent assumptions, multiple complementary metrics, uncertainty estimates, external validation, and continuous monitoring. The proposed framework provides a practical foundation for building such evidence-driven ML pipelines across domains.

References

- [1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys*, vol. 54, pp. 1–35, 2019.
- [2] S. Caton and C. Haas, "Fairness in Machine Learning: A Survey," *ACM Computing Surveys*, vol. 56, pp. 1–38, 2020.
- [3] T. P. Pagano, R. B. Loureiro, F. V. Lisboa, R. M. Peixoto, G. A. S. Guimarães, G. O. R. Cruz, M. M. Araujo, L. L. Santos, M. A. S. Cruz, E. L. Oliveira, I. Winkler, and E. S. Nascimento, "Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods," *Big Data and Cognitive Computing*, vol. 7, no. 15, 2023.
- [4] N. Shahbazi, Y. Lin, A. Asudeh, and H. V. Jagadish, "Representation Bias in Data: A Survey on Identification and Resolution Techniques," *ACM Computing Surveys*, vol. 55, pp. 1–39, 2022.
- [5] A. Fabris, S. Messina, G. Silvello, and G. A. Susto, "Algorithmic fairness datasets: the story so far," *Data Mining and Knowledge Discovery*, vol. 36, pp. 2074–2152, 2022.
- [6] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012.
- [7] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," in *Advances in Neural Information Processing Systems*, vol. 29, 2016.

- [8] 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.
- [9] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *8th Innovations in Theoretical Computer Science Conference*, 2017.
- [10] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, pp. 325–333, 2013.
- [11] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, pp. 1–33, 2011.
- [12] F. Kamishima, T. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50, 2012.
- [13] M. Zafar, I. Valera, M. Gomez-Rodriguez, and K. Gummadi, "Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171–1180, 2017.
- [14] M. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On Fairness and Calibration," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [15] M. Hort, Z. Chen, J. Zhang, F. Sarro, and M. Harman, "Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey," *ACM Journal on Responsible Computing*, vol. 1, pp. 1–52, 2022.
- [16] J. Dockès, G. Varoquaux, and J. B. Poline, "Preventing dataset shift from breaking machine-learning biomarkers," *GigaScience*, vol. 10, 2021.
- [17] F. Ding, M. Hardt, J. Miller, and L. Schmidt, "Retiring Adult: New Datasets for Fair Machine Learning," *Advances in Neural Information Processing Systems*, pp. 6478–6490, 2021.
- [18] A. Balayn, C. Lofi, and G. Houben, "Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches," *The VLDB Journal*, vol. 30, pp. 739–768, 2021.
- [19] M. Mitchell et al., "Model Cards for Model Reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, 2019.
- [20] I. Dominguez-Catena, D. Paternain, and M. Galar, "Metrics for Dataset Demographic Bias: A Case Study on Facial Expression Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, pp. 5209–5226, 2023.
- [21] M. Favier, T. Calders, S. Pinxteren, and J. Meyer, "How to be fair? A study of label and selection bias," *Machine Learning*, vol. 112, pp. 5081–5104, 2023.
- [22] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2018.
- [23] R. K. E. Bellamy et al., "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, 2019.
- [24] T. Calders, E. Ntoutsis, M. Pechenizkiy, B. Rosenhahn, and S. Ruggieri, "Introduction to The Special Section on Bias and Fairness in AI," *ACM SIGKDD Explorations Newsletter*, vol. 23, pp. 1–3, 2021.

- [25] D. Sculley et al., “Hidden Technical Debt in Machine Learning Systems,” in *Advances in Neural Information Processing Systems*, vol. 28, 2015.
- [26] T. Gebru et al., “Datasheets for Datasets,” *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.
- [27] S. Goethals, D. Martens, and T. Calders, “PreCoF: counterfactual explanations for fairness,” *Machine Learning*, 2023.
- [28] M. Miceli, J. Posada, and T. Yang, “Studying Up Machine Learning Data,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, pp. 1–14, 2021.