Data-Centric Algorithmic Fairness: A Systematic Review of Data-Level Interventions for Algorithmic Fairness

LUCA DECK*, AI Responsibility Centre, University of Bayreuth & Fraunhofer FIT, Germany
DOMENIQUE ZIPPERLING*, AI Responsibility Centre, University of Bayreuth & Fraunhofer FIT, Germany
LEO JESSAT, University of Bayreuth, Germany
NIKLAS KÜHL, AI Responsibility Centre, University of Bayreuth & Fraunhofer FIT, Germany

Algorithmic fairness has historically largely focused on model constraints and post-hoc adjustments to satisfy fairness metrics. However, such interventions often impose top-down fairness constraints that remain politically and legally contested and frequently encounter resistance from practitioners, e.g., due to feared accuracy-fairness trade-offs. In this paper, we investigate how the "Data-Centric AI" paradigm—i.e., improving performance through data-level interventions—can be applied toward algorithmic fairness metrics by systematically reviewing 79 papers and mapping the identified data-level fairness interventions to the two pillars of data-centric AI: data refinement (e.g., data cleaning and feature engineering) and data extension (e.g., targeted acquisition of instances and features). Our findings reveal three patterns present in current research: First, extension strategies remain under-explored despite their potential to address systemic imbalances. Second, most studies treat fairness metrics as a secondary or hindsight consideration with limited depth of empirical evaluation, which calls for more efforts focused on specific fairness objectives along the entire lifecycle. Third, we observe a lack of consistent terminology and guidance for the effective selection of techniques for practitioners. We propose the data-centric algorithmic fairness paradigm to establish a unified language for data-level fairness interventions and highlight its potential to augment model-centric approaches for future applications.

Keywords: Data-Centric AI, Algorithmic Fairness, Bias Mitigation, Pre-Processing, Data Acquisition

Reference Format:
Luca Deck, Domenique Zipperling, Leo Jessat, and Niklas Kühl. 2026. Data-Centric Algorithmic Fairness: A Systematic Review of Data-Level Interventions for Algorithmic Fairness. In *Proceedings of Fifth European Conference on Algorithmic Fairness (ECAF'26).* Proceedings of Machine Learning Research, 23 pages.

1 Introduction

The rapid integration of automated decision-making systems into critical domains—ranging from hiring and credit scoring to judicial risk assessment—has catalyzed a rigorous academic debate

*Both authors contributed equally to this research.

Authors' Contact Information: Luca Deck, AI Responsibility Centre, University of Bayreuth & Fraunhofer FIT, Germany; Domenique Zipperling, AI Responsibility Centre, University of Bayreuth & Fraunhofer FIT, Germany; Leo Jessat, University of Bayreuth, Germany; Niklas Kühl, AI Responsibility Centre, University of Bayreuth & Fraunhofer FIT, Germany.

This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

ECAF'26, September 02–September 04, 2026, Ghent, BE

© 2026 Copyright held by the owner/author(s).

regarding algorithmic fairness [7, 78]. Historically, this debate has largely centered on "model-centric" solutions, i.e., mathematical constraints applied during or after model training to satisfy formal metrics such as demographic parity or equalized odds [48, 114]. However, these top-down interventions are increasingly meeting resistance. From a legal perspective, "fixing" outcomes creates tensions with non-discrimination law [28, 85]; from a practitioner's view, these constraints often introduce perceived trade-offs between predictive accuracy and socially or ethically desirable outcomes [65, 105] that hinder adoption in practice [5, 32, 87, 97].

Amidst these tensions, a paradigm shift is occurring in the broader machine learning (ML) community toward *Data-Centric AI (DCAI)* [56, 109]. While model-centric AI treats data as a static, often flawed input to be rectified at an algorithmic level, DCAI aims to improve system performance through systematic engineering of the dataset itself. In theory, this presents a powerful complement to model-centric fairness approaches: rather than policing a model trained on biased data, data-level interventions could address bias at its source [30, 71]. This shift implies a different understanding of algorithmic fairness. Rather than treating fairness only as a post-hoc constraint imposed on trained models, a data-centric perspective highlights the possibility of embedding fairness considerations into the data on which AI systems are built. Despite the intuitive appeal of a data-centric approach to fairness, research remains fragmented. Although many pre-processing techniques exist, they are rarely framed within the broader DCAI discourse, leading to a lack of conceptual clarity and shared terminology.

In this paper, we bridge this gap by conducting a systematic literature review (SLR) of 79 papers to investigate how DCAI principles can be operationalized for algorithmic fairness. The corpus includes 41 primary studies on data-level fairness interventions and 38 surveys that contextualize how prior research frames the relationship between data and fairness. We map the primary interventions onto the two core pillars of the DCAI paradigm: *refine* and *extend*. *Refine* techniques enhance the quality of existing records, such as label sanitization, re-weighting, and fairness-aware feature engineering, while *extend* techniques alter the dataset's composition through targeted acquisition of underrepresented instances, features or labels. Our synthesis reveals three main findings. First, the reviewed primary literature is strongly oriented towards *refine* techniques: 35 of the 41 papers address at least one *refine* subdimension, whereas only six focus on *extend* approaches. Second, *refine* techniques are dominated by instance-level interventions, especially the removal of low-quality or bias-inducing instances and the increased weighting, sampling, or generation of fairness-relevant observations. By contrast, label quality, proxy-like feature information, and *extend* approaches receive limited systematic attention. Third, no reviewed primary study combines the two dimensions, indicating a lack of integrated strategies across data acquisition, coverage, refinement, and evaluation.

Based on these findings, we make two contributions. First, we conceptualize *Data-Centric Algorithmic Fairness (DCAF)* as a distinct perspective that shifts attention from model-centric correction to the fairness-sensitive curation, enhancement, and extension of datasets. By translating the DCAI paradigm to algorithmic fairness, we provide a unified terminology for data-level fairness interventions. Second, we provide a systematic overview of existing DCAF techniques. Based on our SLR, we map data-level fairness interventions onto the *refine* and *extend* subdimensions, identifying dominant methodological patterns, including the strong concentration on refinement-oriented and instance-level techniques. Our findings suggest that DCAF should be understood

as more than repairing datasets in hindsight through simple pre-processing techniques. It rather treats fairness as a data quality requirement from the outset and connects targeted data acquisition, coverage, enhancement, and evaluation across the AI lifecycle. This shifts attention from individual pre-processing techniques toward broader data strategies that clarify when fairness problems require cleaning, reweighting, augmentation, synthetic generation, or additional data collection.

After introducing the conceptual foundations in Section 2 and the review method in Section 3, Section 4 first synthesizes the survey literature, while Section 5 presents the classification of primary studies along the DCAI pillars and subdimensions. Section 6 discusses the results and highlights three key implications for practice and future research. In Section 7, we conclude.

2 Background

In this paper, we introduce data-centric algorithmic fairness (DCAF) as a new paradigm that places focus on data-level interventions to improve algorithmic fairness. Therefore, we start by covering the two cornerstones of this paradigm: algorithmic fairness and DCAI.

2.1 Algorithmic Fairness

Recent failures of automated systems in high-risk domains have debunked the initial premise that algorithms act as inherently neutral decision-makers [2, 6, 26, 39]. Early attempts to mitigate unfairness by simply excluding sensitive attributes ("fairness through unawareness") have proven ineffective due to redundant encoding, i.e., the capacity of ML models to reconstruct protected characteristics through correlated proxy variables [35]. Consequently, the research community has shifted toward the proactive enforcement of algorithmic fairness constraints, typically implemented via in-processing or post-processing interventions. The emerging metrics can be generally categorized into group fairness and individual fairness [6, 13, 104]. Group fairness focuses on distributive parity across demographic groups, e.g., defined by race or gender [22]. Popular group fairness metrics include demographic parity [36] and equality of opportunity [49]. In contrast, individual fairness demands that similar individuals should receive similar outcomes, necessitating the definition of a domain-specific similarity metrics [9, 35]. A critical challenge in the deployment of these systems is the "impossibility theorem" of fairness: in most real-world scenarios characterized by disparate base rates in the training data, multiple fairness criteria cannot be satisfied concurrently [31, 65]. Thus, (formally) fair AI systems require crucial design choices about which formal measures to apply in which context.

2.2 Data-Centric AI

Recent work on DCAI has shifted attention from exclusively model-centric improvement strategies toward the systematic design, refinement, and extension of data throughout the ML lifecycle [56, 76, 94, 109, 115]. Within this perspective, the relevant forms of data intervention can be broadly organized into *refine* and *extend* techniques [56]. *Refine* refers to interventions that improve the quality or composition of already available data, for example, through cleaning, relabeling, reweighting, or resampling. *Extend*, by contrast, refers to interventions that add new information to the dataset, such as additional instances, features, or labels. Applied to fairness metrics, *refine* techniques may correct, rebalance, or restructure existing data, whereas *extend* techniques may

enlarge the representation of relevant groups, patterns, or regions of the data space through data acquisition. A complete overview of the dimensions proposed by Jakubik et al. [56] can be found in Table 1.

Table 1. Structured overview of *refine* and *extend* techniques in data-centric AI, adapted from Jakubik et al. [56].

Dimension	Goal	Subdimension	Description
Refine	Improve data quality on an instance level	**R1**	Improve feature quality
		R2	Improve label quality
	Improve data quality on a dataset level	**R3**	Increase volume of high-relevance instances
		R4	Reduce volume of low-quality instances
		R5	Increase volume of relevant features
		R6	Reduce volume of unmeaningful features
Extend	Acquire new data for blind spots on a dataset level	**E1**	Acquire new instances
		E2	Acquire new features
		E3	Acquire new labels

Conceptually, this understanding aligns closely with the common distinction between *pre-processing* versus *in-processing* and *post-processing* stages in algorithmic fairness [109], where pre-processing intervenes in the data before model training, e.g., by repairing, reweighting, resampling, or extending the dataset. In-processing and post-processing, in contrast, primarily operate at the level of the model, learning objective, or prediction outputs and can therefore be understood as predominantly model-centric or output-centric interventions [14, 109].

2.3 Data-Centric Algorithmic Fairness

While the DCAI paradigm was initially motivated primarily by predictive performance, robustness, and generalization, the same logic can be applied to algorithmic fairness: If model outcomes are strongly shaped by the quality, composition, and coverage of the training data, then fairness metrics should likewise be understood as something that can be optimized at data-level rather than only through modifications of the ML algorithm [21, 30, 68, 71, 99, 109, 115]. This application of the DCAI paradigm toward algorithmic fairness motivates what we refer to as the *data-centric algorithmic fairness (DCAF)* paradigm. More specifically, we use DCAF to describe fairness-oriented interventions that target the data basis from which group disparities emerge, rather than primarily modifying the learning objective, decision rule, or outputs of an ML model. This understanding acknowledges that fairness may involve broader goals, such as perceived fairness or longer-term societal effects, while focusing on *formal fairness* as one of its central direct and operationalizable objectives in ML system development [30]. While DCAF is located mainly in the typical pre-processing stages, the sources for unfairness should also be addressed even more upstream in the data-generation process (e.g., by changing recruitment policies instead of merely resampling training data).

Definition: Data-Centric Algorithmic Fairness

Data-centric algorithmic fairness is a paradigm that locates the primary intervention point for algorithmic fairness in the data (rather than the model), using systematic curation, enhancement, documentation, and targeted acquisition to address biases at their source before they propagate into model behavior.

3 Method

We conduct an SLR to identify recent work on data-level interventions for algorithmic fairness. The review follows established SLR guidelines [64, 110] and uses PRISMA for documenting the screening and selection process [81]. To capture recent developments as comprehensively as possible, we considered both peer-reviewed academic literature and grey literature [40]. The academic search covered Wiley, ScienceDirect, IEEE, EBSCO, and JSTOR, while arXiv.org was used as an additional preprint source. Table 2 summarizes the Boolean search strings applied to titles, abstracts, and keywords.

Across all sources, we identified 1,766 records. After removing duplicates, non-English papers, and non-scientific records, 695 publications remained for detailed screening. Titles and abstracts were manually screened for relevance, resulting in an initial seed set of 59 studies. Studies were excluded if they focused solely on model-centric interventions, did not involve data preprocessing, or lacked a fairness-oriented objective. Three guiding questions informed the selection: (1) Does the study apply the data-centric AI paradigm to fairness? (2) Does it explicitly link data-centric techniques to fairness goals? (3) Does it evaluate the fairness performance of such techniques?

Additionally, we conducted backward and forward snowballing over three iterations [111], supported by Citationchaser [46] and the Lens.org citation database. The final corpus comprises 79 papers, of which 41 are primary research articles, and 38 are surveys. Since this review focuses on studies that apply data-level fairness interventions, the 41 primary research articles form the empirical basis of our paper, for the classification along the DCAI pillars and subsequent analysis. We classify these papers using the *refine–extend* taxonomy proposed by Jakubik et al. [56], allowing us to identify dominant research trajectories and systematic blind spots in DCAF. Because surveys made up a substantial share of the corpus, we examined them separately to contextualize the primary studies and assess where prior research has concentrated. Since surveys synthesize established research streams, they reveal dominant framings of data and fairness. We

Search Field	Boolean Search String
Abstract	`ai AND data cent* AND fair*`
Title	`ai AND data AND fair*`
Abstract	`ai AND data AND fair* AND (pre-process* OR preprocess* OR feature engineering)`
Keywords	`ai AND data AND fairness`
Title + Abstract	`data AND ai AND fair*`
Abstract	`data-centric AND fairness`

Table 2. Overview of Boolean search strings applied in literature selection.

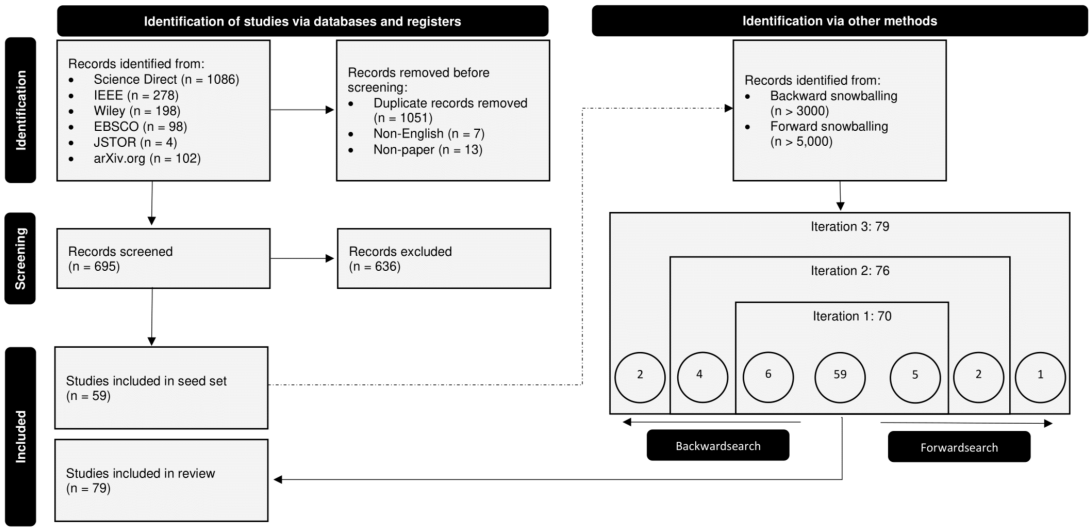

Fig. 1. PRISMA flowchart describing the article selection procedure, adapted from Page et al. [81].

therefore grouped them by main emphasis, such as fairness metrics, bias mitigation, data quality, documentation, and DCAI. The following sections first summarize this survey literature before analyzing primary data-level fairness interventions through the *refine–extend* taxonomy.

4 Overview of Survey Literature

The 38 surveys in our corpus provide a broader view of how existing research has organized questions of fairness, bias, and data-related ML practice. They are dominated by systematic and scoping reviews, while taxonomy-oriented and explicitly conceptual contributions are less frequent. Thematically, most surveys focus on mitigation techniques, data infrastructure, and bias in data. Fairness definitions, metrics, and application domains are discussed as well, but less often serve as the primary organizing perspective. The full list of surveys can be found in Table 5.

4.1 Mitigation-Oriented Fairness Surveys

A substantial part of the surveys addresses algorithmic fairness and bias mitigation. Several surveys organize the field around mitigation techniques, commonly distinguishing between pre-processing, in-processing, and post-processing approaches [79, 100, 101, 108]. Others combine mitigation-oriented discussions with broader coverage of fairness metrics, bias sources, and methodological challenges [14, 38, 80, 94]. Within this strand, data-level interventions appear primarily as one family of mitigation approaches. Surveys on data preparation, preprocessing, data augmentation, and imbalanced learning show that data-oriented techniques are already recognized within fairness research [53, 60, 62, 75, 100, 108]. However, they are usually embedded in broader mitigation taxonomies rather than treated as a distinct data-centric perspective.

4.2 Data Quality, Datasets, and Data Infrastructure

A second strand focuses on data quality, datasets, and data infrastructure. Some surveys are closely related to the DCAI paradigm, especially where they emphasize systematic data improvement, refinement, extension, data quality enhancement, or lifecycle management [56, 76, 96]. Related

reviews address data collection, dataset quality, data preparation, and tools or systems for ML [88, 109, 116]. Several contributions are particularly relevant for fairness-oriented data-centric work. Le Quy et al. [69] survey datasets used in fairness-aware ML, highlighting the role of benchmark datasets in shaping fairness research. Lu et al. [73] review synthetic data generation and discuss data availability, privacy, and fairness. Overall, this strand foregrounds the importance of data-related foundations for ML, although fairness is not always the central organizing concern.

4.3 Dataset Practices, Documentation, and Data Bias

A third strand examines datasets as socio-technical artifacts whose construction, documentation, reuse, and preprocessing can shape fairness-related outcomes. Paullada et al. [83] discuss dataset development and use in ML, including representation, documentation, and reuse. Gebru et al. [41] introduce datasheets for datasets as a framework for improving transparency and accountability. Lucchesi et al. [74] propose Smallset Timelines to make preprocessing decisions more visible and reproducible. Other surveys examine how dataset reuse, benchmark concentration, preprocessing decisions, missingness, feedback effects, and broader data practices affect the validity and social implications of ML [3, 37, 66, 92, 95, 102]. This literature emphasizes that fairness-related problems often emerge before model training, through the ways data are collected, documented, transformed, and reused.

4.4 Fairness Definitions, Metrics, and Conceptual Perspectives

A smaller group of surveys and conceptual contributions focuses primarily on fairness definitions, metrics, and conceptual framing. Hutchinson and Mitchell [54] provide a historical account of fairness concepts in educational and employment testing and connect these debates to algorithmic fairness. Corbett-Davies et al. [24] critically examine formal fairness definitions and argue that fairness criteria should be evaluated in relation to their context-specific consequences. Other surveys, including Chen et al. [21], Pagano et al. [80], and Caton and Haas [14], also discuss fairness definitions and metrics, but usually as part of broader reviews of fairness methods, mitigation techniques, or application areas. Across the survey corpus, fairness metrics are therefore present as a recurring topic, but less often constitute the main organizing focus.

4.5 Synthesis

Taken together, the survey literature shows that data-related fairness concerns are already present, but scattered across separate research streams. Mitigation surveys treat data-level interventions as one technique family among many, DCAI and data-quality surveys emphasize data improvement, and dataset-practice surveys focus on documentation, reuse, and bias in data construction. This fragmentation makes a data-centric classification of algorithmic fairness research valuable as it brings these dispersed perspectives into a unified terminology and clarifies which data-level interventions are actually used for fairness. The following analysis, therefore, applies the DCAI paradigm to primary studies that operationalize data-level fairness interventions.

5 Current State of Research on Data-Level Fairness Interventions

This section reports the descriptive results of the classification and examines how the identified papers are distributed across the refinement and extension subdimensions of DCAI. It first outlines

the overall coverage of the pillars and their subdimensions and recurring combination patterns, before turning to the individual subdimensions, their associated method families, and exemplary papers. The goal is to identify dominant research patterns as well as neglected areas in the literature on data-centric interventions for algorithmic fairness.

5.1 Overview

Table 3 summarizes the classification of the 41 included papers. The distribution shows a clear imbalance between the two pillars: 35 papers address at least one *refine* subdimension, whereas only six papers address any *extend* subdimension. No paper is classified as addressing both dimensions simultaneously, indicating that the literature treats DCAF mainly as a problem of modifying already available datasets, while the acquisition or expansion of data is rarely integrated.

Within the *refine* dimension, the strongest emphasis lies on instance-level curation. R4, which captures the reduction of low-quality or bias-inducing instances, is the most frequent subdimension with 23 papers. R3 (increase high-relevant instances) and R5 (increase relevant features) follow with 20 papers each, indicating a strong focus on increasing the relevance or coverage of training instances through sampling, weighting, or synthetic augmentation. R1 (increase feature quality) is covered by 16 papers, R2 (increase label quality) by 13 papers, and R6 (reduce meaningless features) by only 7 papers. This distribution suggests that the literature concentrates most strongly on changing the composition of training instances, while label quality and especially the reduction of meaningless or proxy-like features receive less systematic attention.

The breadth of refinement coverage also shows a mixed picture. Of the 35 papers, ten address only one subdimension, three address two, nine address three, nine address four, and four address five. No paper covers all six subdimensions. Thus, over 70 % of studies go beyond a single intervention type, but even broader papers usually remain selective. The literature is therefore not purely narrow, but it also does not yet offer comprehensive refinement frameworks that jointly address feature quality, label quality, instance selection, augmentation, and feature reduction.

The *extend* pillar is much narrower. Of the six papers, five address only E1 (acquire instances), while one paper addresses E2 (acquire features) and E3 (acquire labels). No paper covers all three subdimensions, and no paper combines E1 with E2 or E3. This pattern shows that extension-oriented work is not only rare but also fragmented. Existing studies mostly focus on data coverage or acquisition of instances, while broader analytical or conceptual extensions of DCAF are represented by only one paper.

The co-occurrence analysis shows that refinement subdimensions cluster around a few combinations. The strongest pair is R3+R4, which appears in 17 of the 35 papers. This suggests that the dominant refinement logic is instance-level curation: improving fairness by increasing the relevance of useful observations while reducing the influence of harmful or low-quality ones. This pattern is reinforced by the most frequent higher-order combinations, especially R1+R3+R4 and R3+R4+R5, which each appear in 9 papers, and R2+R3+R4, which appears in 7 papers. These combinations indicate that instance-level interventions often serve as the core around which feature repair, label correction, or synthetic augmentation are added.

At the same time, our results reveal several neglected areas. R6 is the weakest subdimension of the *refine* dimension, suggesting that fewer studies explicitly target the removal of feature information. R2 is also less frequently addressed, indicating that label quality receives less attention

Source	Refine						Extend		
	R1	R2	R3	R4	R5	R6	E1	E2	E3
***Refine* only (35 papers)**									
Biswas and Rajan [10]	✓	✓		✓	✓	✓			
Bödi and Grabner [11]	✓								
Celis et al. [15]	✓		✓	✓					
Chakraborty et al. [17]	✓	✓	✓						
Chaudhari et al. [18]	✓		✓	✓		✓			
Chaudhari et al. [19]					✓				
Clemente et al. [23]	✓	✓		✓		✓			
Dang et al. [25]	✓		✓	✓		✓			
de Oliveira and Berton [27]		✓		✓	✓				
Duong and Conrad [33]	✓	✓	✓	✓					
Duong and Conrad [34]			✓	✓	✓				
González-Sendino et al. [42]					✓				
Gonzalez Zelaya [43]			✓	✓	✓				
Gray et al. [44]	✓								
Gujar et al. [45]					✓				
Iosifidis and Ntoutsi [55]			✓	✓					
Jiang et al. [57]					✓				
Kamiran and Calders [58]	✓	✓	✓	✓	✓				
Kim et al. [63]			✓	✓	✓				
Krasanakis et al. [67]	✓	✓	✓			✓			
Langenberg et al. [68]			✓	✓	✓				
Lee et al. [70]					✓				
Lin et al. [72]		✓	✓	✓					
Mazzoni et al. [77]					✓				
Pascual-Triana et al. [82]				✓					
Peng et al. [84]		✓	✓	✓	✓	✓			
Popoola and Sheppard [86]	✓		✓	✓	✓				
Rosado Gómez et al. [89]	✓		✓	✓	✓				
Sattigeri et al. [90]					✓				
Sha et al. [91]		✓	✓		✓				
Shahrezaei et al. [93]		✓	✓	✓	✓				
Verma et al. [103]	✓	✓	✓	✓		✓			
Wang and Singh [107]	✓			✓					
Yan et al. [112]				✓	✓				
Yu et al. [113]	✓	✓	✓	✓					
***Extend* only (6 papers)**									
Asudeh et al. [4]							✓		
Chen et al. [20]							✓		
Hasan and Pradhan [50]							✓		
Kärkkäinen and Joo [59]							✓		
Li et al. [71]								✓	✓
Tae and Whang [98]							✓		
Total mentions	16	13	20	23	20	7	5	1	1

Table 3. Overview of papers classified by *refine* and *extend* dimensions. A checkmark [✓] indicates the topic is a central focus of the article.

than sampling, pruning, or augmentation. The most pronounced gap, however, concerns the absence of papers that connect the two pillars towards a holistic framework. Current work does not link preprocessing interventions with questions of what additional data should be collected, for whom, and under which fairness objective. This absence points to a missing lifecycle perspective

in DCAF: fairness is mostly studied after the dataset already exists, rather than as a combined problem of data acquisition, coverage, refinement, and evaluation.

5.2 Dimension-Specific Results

After identifying aggregated patterns of focus and neglect, this subsection discusses the individual subdimensions in more detail. For each subdimension, we summarize the dominant method families, list representative techniques, and highlight exemplary papers that illustrate how the dimension is addressed in the literature. Table 4 provides an overview of the different method families and concrete techniques mentioned in the papers per DCAI pillar and subdimension.

5.2.1 Refine Subdimension

R1: Improve feature quality. This subdimension covers work that improves the quality, neutrality, or reliability of existing input data before training. Key approaches include feature repair, preprocessing improvement, distribution correction, and representation-level debiasing. For example, Wang and Singh [107] link missing values and selection bias to fairness, while Celis et al. [15] adjust group representation rates through maximum-entropy preprocessing. Other work studies fairness in preprocessing pipelines [10] or constructs invariant representations that ignore sensitive variation [11]. Overall, feature quality is treated as a fairness-relevant property of how input information is encoded and transformed.

R2: Improve label quality. This category includes studies that address biased or insufficient supervision signals before training, including relabeling, label-side correction, group-label rebalancing, and detecting contradictory supervision. Kamiran and Calders [58] provide a foundational preprocessing framework with label-side interventions such as massaging, while de Oliveira and Berton [27] examine fair preprocessing under label scarcity. Other work rebalances protected attributes [84] or identifies cases where similar feature profiles receive different labels due to protected-attribute variation [18]. These studies treat labels as fairness-critical data components.

R3: Increase volume of high-relevance instances. Work belonging to this subdimension increases the influence of fairness-relevant observations, especially minority-group, protected-category, or underrepresented cases. Methods include targeted sampling, fairness-aware batches, adaptive weighting, representation adjustment, and synthetic addition. Examples include FABBOO for fairness-aware online learning under class imbalance [55], fairness-aware batch sampling [63], protected-category sampling [86], adaptive sensitive reweighting [67], and maximum-entropy representation adjustment [15]. These methods selectively increase the training relevance of observations important for fair decision boundaries.

R4: Reduce volume of low-quality instances. This subdimension covers methods that reduce the influence of harmful, biased, noisy, redundant, or fairness-detrimental observations. Approaches include bias-aware pruning, fairness-aware undersampling, biased-data removal, subset selection, and filtering under missingness or selection bias. Examples include Fair-ONB, which removes overlap-region instances near decision boundaries [82], debiased subset construction [103], bias-inducing instance detection [18], fairness-driven data removal [34], and subset selection for discrimination reduction [33]. These studies treat data reduction as an active fairness intervention.

R5: Increase volume of relevant features. This subdimension expands coverage of fairness-relevant feature regions through synthetic data generation, fairness-aware augmentation, and

diversity-enhancing generation. Rather than adding raw columns, these methods expose the model to sparse feature combinations, subgroup patterns, or minority cases. Examples include GAN-based anomaly generation [70], Fairness GAN [90], FairGen [19], GenEthos [45], GenFair [77], and fair causal data generation [42]. They create additional informative data configurations.

R6: Reduce volume of unmeaningful features. This subdimension reduces the influence of proxy-like, bias-inducing, or low-value feature information. Main approaches include proxy suppression, protected-attribute masking or rebalancing, transformer auditing, and removal of discriminatory feature–label configurations. Examples include FairMask for reducing protected-attribute

Subdimension	Method families	Exemplary techniques and papers
R1: Improve feature quality	Feature-quality repair and preprocessing-quality improvement [10, 23, 107]; distribution correction [15]; representation-level debiasing [11]; fairness-aware preprocessing pipelines [58, 89, 113].	Maximum-entropy preprocessing [15]; adaptive sensitive reweighting [67]; invariant representation learning [11]; missingness and selection-bias analysis [107]; data profiling for high-quality data [23]; compositional fairness analysis of data transformers [10].
R2: Improve label quality	Relabeling and label-side correction [17, 58, 113]; protected-attribute or group-label rebalancing [33, 84]; correction of contradictory supervision [18]; fair preprocessing under limited labels [27].	Massaging / relabeling-style preprocessing [58]; fair preprocessing in few-labeled settings [27]; FairMask model-based rebalancing of protected attributes [84]; FairBalance [113]; Fairway [17]; protected-attribute-induced label-conflict detection [18].
R3: Increase volume of high-relevance instances	Targeted oversampling and fairness-aware sampling [55, 63, 86]; adaptive weighting of fairness-critical observations [67]; protected-group representation adjustment [15, 84]; fairness-aware synthetic addition [33].	FABBOO [55]; fairness-aware batch sampling [63]; protected-category sampling [86]; adaptive sensitive reweighting [67]; maximum-entropy representation-rate adjustment [15]; FairMask protected-attribute rebalancing [84]; FairBalance [113]; synthetic-data addition in an optimization framework [33].
R4: Reduce volume of low-quality instances	Bias-aware pruning and biased-data removal [18, 19, 103]; fairness-aware undersampling [82]; fairness-driven subset selection [33, 34]; filtering or downweighting under missingness and selection bias [107].	Fair-ONB guided undersampling [82]; biased training-data removal [103]; bias-inducing sample removal [18, 19]; fairness-driven data removal [34]; genetic-algorithm data subset optimization [33]; Fair Class Balancing [112]; missingness and selection-bias filtering [107].
R5: Increase volume of relevant features	Synthetic data generation [19, 45, 57, 70, 90]; fairness-aware augmentation [33, 63, 112]; diversity-enhancing data generation [42, 77]; fair causal data generation [42].	Advanced R-GAN / regularized GAN [70]; Fairness GAN [90]; FairGen [19]; GenEthos [45]; GenFair [77]; fair causal data generation [42]; synthetic dataset generation [57]; synthetic-data addition or exclusive synthetic-data use [33]; Fair Class Balancing [112].
R6: Reduce volume of unmeaningful features	Proxy suppression and protected-attribute masking or rebalancing [67, 84]; preprocessing-transformer auditing [10]; removal of bias-inducing transformations or protected-attribute-induced conflicts [18, 103]; data profiling for feature-quality control [23].	Compositional fairness analysis of data transformers [10]; data profiling for high-quality data [23]; FairMask protected-attribute rebalancing [84]; adaptive sensitive reweighting [67]; biased training-data removal [103]; removal of protected-attribute-induced contradictory instances [18].
E1: Acquire new instances	Coverage-driven instance acquisition and repair [4, 20]; selective acquisition for fairness-relevant slices [50, 98]; balanced dataset construction for underrepresented demographic groups [59].	Coverage assessment and repair of rare attribute combinations [4]; bias–variance–noise decomposition to identify when additional samples can reduce discrimination [20]; *DataSift* with data valuation, partitioning, multi-armed bandits, and influence functions [50]; *FairFace* as balanced race, gender, and age dataset construction [59]; *Slice Tuner* for budget allocation across data slices [98].
E2: Acquire new features	Analysis of how feature availability, data-generation choices, and dataset-level factors shape fairness outcomes [71].	Data-centric factor analysis showing that fairness outcomes depend on the information represented in the dataset; however, the work remains primarily diagnostic and does not provide an operational feature-acquisition strategy [71].
E3: Acquire new labels	Analysis of how label-related data-generation choices and dataset-level assumptions affect downstream fairness [71].	Data-centric fairness-factor analysis discussing how target definitions, labels, and dataset construction choices influence fairness outcomes; however, explicit fairness-aware label acquisition remains largely underdeveloped [71].

Table 4. Representative method families and concrete techniques for the *refine* and *extend* dimensions.

influence [84], adaptive sensitive reweighting [67], auditing preprocessing transformers for unfair dependencies [10], and identifying feature–label conflicts linked to protected attributes [18]. Overall, these methods aim to limit unfair correlations encoded in the feature representation.

5.2.2 Extend Subimension

Most studies in the DCAF literature focus on *refine* techniques—such as cleaning, reweighting, or synthetic resampling—assuming the dataset is fixed. In contrast, only a few studies explore *extend* techniques, where fairness is addressed by acquiring additional data. The following publications fall into the *extend* dimension and offer relevant contributions in this area.

E1: Acquiring new instances. This subdimension acquires or constructs additional instances for underrepresented groups, rare attribute combinations, or fairness-relevant slices. This work links insufficient data coverage to discriminatory model behavior, reduced generalization, or incomplete representation [4, 20, 59]. Some studies remain diagnostic, while others construct balanced datasets or optimize acquisition budgets. For example, *DATASIFT* acquires labeled data under multiple objectives, including fairness [50], and *Slice Tuner* allocates acquisition budgets across slices to improve accuracy and fairness [98]. Overall, E1 is the most prominent type, although existing work remains heterogeneous and often dataset-specific or technically complex.

E2: Acquiring new features. Evidence for feature acquisition is much more limited. Conceptually, E2 is relevant when fairness depends on which contextual variables, proxies, or attributes are available. Existing work shows that data-generation choices and feature availability can affect fairness outcomes [71]. However, it does not provide a general strategy for identifying, collecting, or validating new features for bias reduction. Therefore, this subdimension should be treated as a conceptually relevant but only weakly operationalized type of the *extend* dimension.

E3: Acquiring new labels. Label acquisition is marginally represented. It is relevant when unfairness is linked to missing, noisy, selectively observed, or poorly defined target labels. Existing work discusses how label-related data-generation shapes fairness outcomes [71], but does not develop an operational fairness-aware label acquisition strategy. The literature provides little guidance on when additional labels should be collected, which labels should be prioritized, or how fairness objectives should shape label acquisition budgets. Therefore, this subdimension remains unexplored.

Across these dimensions, DCAF work frames bias as a problem of data quality and data collection, not only model optimization [109]. However, this framing often remains conceptual. Overall, *extend* techniques are recognized as important but unevenly developed. The acquisition of instances is comparatively established, whereas the acquisition of features and labels is better understood as an underexplored pathway.

6 Discussion

In this paper, we shift the focus from model-centric fairness interventions toward a data-centric perspective on algorithmic fairness. Building on the data-centric AI (DCAI) paradigm [56]—which emphasizes the systematic refinement and extension of data as a means of improving ML systems—we examine how algorithmic fairness can be addressed at the level of the dataset rather than primarily through downstream model adjustments. Using this lens, we analyzed 41 peer-reviewed articles identified through a structured literature review. Our results show that

current research is strongly concentrated on *data refinement*: 35 papers address at least one *refine* subdimension, while only 6 papers focus on *data extension*. Moreover, no paper combines both pillars, indicating that data-centric algorithmic fairness (DCAF) is rarely treated as an integrated lifecycle issue spanning acquisition, coverage, refinement, and evaluation. More specifically, the reviewed literature is dominated by instance-level refinement techniques, particularly the removal of low-quality or bias-inducing instances and the reweighting, resampling, or generation of fairness-relevant observations. Recurring combinations of refinement techniques indicate a dominant logic of amplifying useful instances while reducing harmful instances. In contrast, label quality, proxy-like feature information, and especially *extend* techniques remain underexplored. From our observations, we draw three key implications for practice and research.

6.1 From Dataset Repair to a Fair Data Strategy

Overall, our findings show that DCAF research has largely created a *refine* landscape. Most reviewed papers focus on modifying already available datasets through cleaning, pruning, reweighting, resampling, or synthetic augmentation, whereas comparatively few studies address the question of which data should be collected, constructed, or expanded in the first place. This imbalance indicates that data-level fairness is rather approached as a matter of *dataset repair* rather than *data strategy*. We expect that dataset repair is the dominant approach because strategic data collection bears more resource investment and risks due to hardly predictable cost-benefit tradeoffs. Case studies investigating fairness considerations usually use all of the data available for training and testing models. However, we can only understand whether more (diverse) data would improve fairness by systematically withholding or acquiring (parts of the) data. Evaluating how effective and generalizable such strategies are will require empirical research that we hope to spark with this work. Moving toward a more strategic understanding of DCAF requires treating datasets not as fixed artifacts but as objects that can be proactively planned, acquired, extended, and governed with fairness goals in mind. This does not diminish the importance of *refine* techniques but rather highlights that such techniques may be insufficient when fairness problems originate from missing coverage, structural underrepresentation, or inadequate data collection processes.

6.2 From Fairness as a Hindsight Consideration Toward a Lifecycle Perspective

The shift towards holistic data strategies also motivates a broader lifecycle perspective, including questions of data acquisition and dataset construction alongside pre-processing and all downstream fairness interventions. If fairness-relevant data deficits emerge already in the data generation or data sampling process, they are difficult to resolve after the data has already been collected, e.g., when they show imbalances or blind spots. This lifecycle view does not necessarily include only developers but also decision-makers and managers responsible for the data generation processes and even affected parties who can, e.g., influence the data basis via collective action [8, 47]. The reviewed literature rarely connects data acquisition, coverage assessment, preprocessing, augmentation, refinement, and evaluation. In particular, no reviewed paper combines the *refine* and *extend* pillars, suggesting that these strands of research remain largely separate. Building on lifecycle-oriented perspectives on fairness and data governance [1, 30], DCAF should be conceptualized as an iterative process in which fairness considerations inform what data is collected, how

it is represented, how it is refined, how models are evaluated, and whether further data collection is necessary. Such a lifecycle perspective is particularly relevant for intersectional fairness [61]. Much fairness research focuses on disparities between broad demographic categories, such as men and women. While these categories can reveal important inequities, they may also obscure harms experienced by marginalized subgroups, such as elderly minorities or black women [12]. These intersectional disadvantages can persist even when aggregate group-level metrics appear balanced [72]. From a DCAF perspective, this means that subgroup visibility must be considered during data collection, representation, and sampling, rather than only during model evaluation. While data-level interventions are primarily designed to improve accuracy, robustness, or generalization, fairness is often treated as an additional evaluation criterion rather than as the central design objective. For DCAF to develop as a more coherent research agenda, fairness should be treated as a data quality requirement from the outset.

6.3 From Academic Methods to Adoption in Practice

From interview studies, we know that practitioners are struggling to put popular fairness toolkits—which primarily focus on in-processing and post-processing techniques—to effective use [5, 32, 87, 97]. The reasons for that are manifold, but a consistent pattern shows tensions with external business incentives and self-enforced "codes of conduct" that lead to hesitation and uncertainty about using these tools to adjust model outputs. We posit that, in contrast to model-centric fairness interventions, DCAF offers a promising paradigm that might be more aligned with these practical realities. Some practitioners have even explicitly called for methodological guidance during data collection and pre-processing [51, 87]. While many fairness interventions are perceived to distort an assumed "ground truth", DCAF directly questions and shapes this "ground truth"—sometimes in more explicit ways (e.g., R2), sometimes in more subtle ways (e.g., E3). This is a crucial consideration for the debate on fairness-accuracy tradeoffs, in which "bias transforming" [106] techniques are often understood as merely shifting resources or opportunities from privileged toward unprivileged subgroups in a zero-sum-game [85]. In particular, *extend* techniques may be a fruitful perspective towards "pragmatic fairness" [97], e.g., through a less biased ground truth to begin with or through non-zero-sum improvements of (sub-group) accuracy.

Our work organizes a broad range of data-level interventions into a consistent terminology that provides clarity to academia and practice. Based on empirical and qualitative work, future work should establish clearer guidance on when different strategies should be used, how they can be combined, and which fairness objectives they are best suited to address. Such guidelines should help practitioners diagnose whether a fairness problem is primarily caused by biased labels, proxy-like features, low-quality instances, insufficient subgroup coverage, or missing data altogether. They should also support decisions about whether cleaning, reweighting, augmentation, synthetic generation, or additional data collection is the most appropriate response.

6.4 Limitations

Our work is not without limitations. First, the review is limited to 41 peer-reviewed papers identified through a tailored search and inclusion criteria. As a result, relevant work from adjacent areas such as data quality, data governance, or dataset documentation may remain outside the sample. Second, the distinction between *refine* and *extend* is analytically useful, but not always clear-cut.

Methods like synthetic data generation, augmentation, and resampling may blur the boundary between *refining* existing data and *extending* the dataset. The criterion we use for classification is whether added data contains newly collected (as opposed to reconfigured) information. Third, our analysis focuses on mapping and classifying existing interventions rather than assessing their comparative empirical effectiveness. While we identify dominant patterns and blind spots, future work should explore which interventions work best under specific fairness metrics, datasets, or organizational conditions. Fourth, our proposed DCAF paradigm only tackles a very narrow account of fairness. Our review only synthesizes data-centric interventions towards highly formalized metrics of algorithmic fairness. Aspects of data, such as data collection processes or representativeness, can be intrusive, discriminatory, or unfair in many ways and still comply with popular fairness metrics (e.g., when helpful sensitive data is obtained without consent). Further, the DCAF paradigm should not be pursued in isolation or applied uncritically. Careless use of some subdimensions, such as removing non-informative or proxy-like features (R6), may reflect legitimate concerns about pitfalls such as "fairness through unawareness" [29, 35]. For example, recent theoretical and empirical work suggests that deterministic, unquantized fair representation learning approaches often fail to fully remove sensitive information, even when they appear fair on the surface [16]. In practice, DCAF should therefore always be embedded in broader domain-specific legal, ethical, and organizational discourses and invite stakeholder deliberation about what constitutes "fair" data.

7 Conclusion

This work introduces the *Data-Centric Algorithmic Fairness (DCAF)* paradigm, which aims to improve algorithmic fairness bottom-up through purely data-centric interventions [56]—instead of imposing top-down fairness constraints. Our systematic literature review reveals a broad but scattered coverage of data-centric techniques. At the same time, we observe a lack of unified terminology and a crucial blind spot on data acquisition strategies. While, most works focus on *refine* techniques that polish inherently flawed datasets, little is known about the costs and benefits of *extend* strategies where new data is required to improve representativeness and algorithmic fairness. The DCAF paradigm is intended to help researchers and practitioners to strategically dedicate efforts towards targeted aspects of their data and derive generalizable guidelines from future empirical work. It advocates for a transition from reactive algorithmic or post-hoc patching to proactive data curation. Curating datasets that reflect the world in a more equitable way may bypass the inescapable frictions of politically fraught fairness constraints. This may allow algorithmic fairness to emerge as a fundamental property of the data rather than a forced, post-hoc constraint on the model. A promising field of reseach lies ahead.

Appendix

Source	Type				Focus				
	Taxonomy/Framework	Systematic/ Scoping Review	Narrative Review	Historical Overview	Fairness Definitions	Mitigation Techniques	Bias in Data	Application Domains	Data Infrastructure
Ashurst and Weller [3]			✓				✓		
Caton and Haas [14]		✓			✓	✓			
Chen et al. [21]		✓			✓	(✓)	(✓)	(✓)	
Corbett-Davies et al. [24]			✓		✓				
Ferrara [37]			✓				✓		
Ferrara [38]			✓		(✓)	✓	✓		
Gebru et al. [41]	✓								✓
Hort et al. [52]		✓				✓			
Huang et al. [53]		✓				✓			
Hutchinson and Mitchell [54]				✓	✓				
Jakubik et al. [56]	✓							(✓)	✓
Kaur et al. [60]		✓				✓		(✓)	
Kidwai-Khan et al. [62]			✓			✓			
Koch et al. [66]			✓				✓		
Le Quy et al. [69]		✓							✓
Lu et al. [73]		✓							✓
Lucchesi et al. [74]	✓								✓
Maharana et al. [75]		✓				✓			
Majeed and Hwang [76]	✓								✓
Ntoutsi et al. [79]		✓				✓			
Pagano et al. [80]		✓			✓	✓	(✓)		
Paullada et al. [83]			✓				✓		
Roh et al. [88]		✓							✓
Shahbazi et al. [92]		✓					✓		
Siddique et al. [94]		✓				✓	✓		
Simson et al. [95]			✓				✓		
Singh [96]		✓							✓
Tawakuli and Engel [100]			✓			✓			
Trigo et al. [101]		✓				✓			
Varsha [102]		✓					✓		
Werner de Vargas et al. [108]		✓				✓			
Whang et al. [109]		✓							✓
Zhou et al. [116]		✓							✓
Total mentions	4	19	8	1	5 (1)	13 (1)	9 (2)	(3)	10

Table 5. Overview of surveys classified by type and focus. A checkmark [✓] indicates the topic is a central focus. A parenthesized checkmark [(✓)] denotes secondary coverage.

References

[1] Avinash Agarwal and Harsh Agarwal. [n. d.]. A Seven-Layer Model with Checklists for Standardising Fairness Assessment throughout the AI Lifecycle. 4 ([n. d.]), 299–314. https://doi.org/10.1007/s43681-023-00266-9

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. [n. d.]. Machine Bias: There's Software Used across the Country to Predict Future Criminals. {And} It's Biased against Blacks. In *Ethics of Data and Analytics: Concepts and Cases* (first edition ed.), Kirsten Martin (Ed.). CRC Press Taylor & Francis Group, 254–264. https://doi.org/10.1201/9781003278290-37

[3] Carolyn Ashurst and Adrian Weller. 2023. Fairness Without Demographic Data: A Survey of Approaches. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3617694.3623234

[4] Abolfazl Asudeh, Zhongjun Jin, and H. V. Jagadish. 2019. Assessing and Remedying Coverage for a Given Dataset. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 554–565. https://doi.org/10.1109/ICDE.2019.00056

[5] Agathe Balayn, Mireia Yurrita, Jie Yang, and Ujwal Gadiraju. [n. d.]. "Fairness Toolkits, A Checkbox Culture?" On the Factors That Fragment Developer Practices in Handling Algorithmic Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '23)*. Association for Computing Machinery, 482–495. https://doi.org/10.1145/3600211.3604674

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. [n. d.]. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org.

[7] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *SSRN Electronic Journal* (2016).

[8] Omri Ben-Dov, Samira Samadi, Amartya Sanyal, and Alexandru Ţifrea. 2025. Fairness for the People, by the People: Minority Collective Action. https://doi.org/10.48550/arXiv.2508.15374

[9] Reuben Binns. 2020. On the Apparent Conflict between Individual and Group Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3351095.3372864

[10] Sumon Biswas and Hridesh Rajan. 2021. Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, Diomidis Spinellis, Georgios Gousios, Marsha Chechik, and Massimiliano Di Penta (Eds.). ACM, New York, NY, USA, 981–993. https://doi.org/10.1145/3468264.3468536

[11] Linda Helen Bödi and Helmut Grabner. 2020. Learning to ignore : fair and task independent representations. https://doi.org/10.21256/zhaw-21602

[12] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Conference on Fairness, Accountability and Transparency* (2018), 77–91.

[13] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. A Clarification of the Nuances in the Fairness Metrics Landscape. *Scientific Reports* 12, 1 (2022), 4209. https://doi.org/10.1038/s41598-022-07939-1

[14] Simon Caton and Christian Haas. 2024. Fairness in Machine Learning: A Survey. *Comput. Surveys* 56, 7 (2024), 1–38. https://doi.org/10.1145/3616865

[15] L. Elisa Celis, Vijay Keswani, and Nisheeth K. Vishnoi. 2019. Data preprocessing to mitigate bias: A maximum entropy based approach. http://arxiv.org/pdf/1906.02164v2

[16] Mattia Cerrato, Marius Köppel, Philipp Wolf, and Stefan Kramer. 2024. 10 Years of Fair Representations: Challenges and Opportunities. https://doi.org/10.48550/arXiv.2407.03834

[17] Joymallya Chakraborty, Suvodeep Majumder, Zhe Yu, and Tim Menzies. 2020. Fairway: a way to build fair ML software. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, Prem Devanbu, Myra Cohen, and Thomas Zimmermann (Eds.). ACM, New York, NY, USA, 654–665. https://doi.org/10.1145/3368089.3409697

[18] Bhushan Chaudhari, Akash Agarwal, and Tanmoy Bhowmik. 2022. Simultaneous Improvement of ML Model Fairness and Performance by Identifying Bias in Data. http://arxiv.org/pdf/2210.13182v1

[19] Bhushan Chaudhari, Himanshu Chaudhary, Aakash Agarwal, Kamna Meena, and Tanmoy Bhowmik. 2022. FairGen: Fair Synthetic Data Generation. http://arxiv.org/pdf/2210.13023v2

[20] Irene Chen, Fredrik D. Johansson, and David Sontag. 2018. Why Is My Classifier Discriminatory? https://doi.org/10.48550/arXiv.1805.12002

[21] Pu Chen, Linna Wu, and Lei Wang. 2023. AI Fairness in Data Management and Analytics: A Review on Challenges, Methodologies and Applications. *Applied Sciences* 13, 18 (2023), 10258. https://doi.org/10.3390/app131810258

[22] Alexandra Chouldechova. [n. d.]. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. 5, 2 ([n. d.]), 153–163. https://doi.org/10.1089/big.2016.0047

[23] Fabiana Clemente, Gonçalo Martins Ribeiro, Alexandre Quemy, Miriam Seoane Santos, Ricardo Cardoso Pereira, and Alex Barros. 2023. ydata-profiling: Accelerating data-centric AI with high-quality data. *Neurocomputing* 554 (2023), 126585. https://doi.org/10.1016/j.neucom.2023.126585

[24] Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. 2018. The Measure and Mismeasure of Fairness. *Journal of Machine Learning Research* (2018). http://arxiv.org/pdf/1808.00023v3

[25] Vien Ngoc Dang, Anna Cascarano, Rosa H. Mulder, Charlotte Cecil, Maria A. Zuluaga, Jerónimo Hernández-González, and Karim Lekadir. 2024. Fairness and bias correction in machine learning for depression prediction across four study populations. *Scientific reports* 14, 1 (2024), 7848. https://doi.org/10.1038/s41598-024-58427-7

[26] Jeffrey Dastin. [n. d.]. Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women *. In *Ethics of Data and Analytics: Concepts and Cases* (first edition ed.), Kirsten Martin (Ed.). CRC Press Taylor & Francis Group, 296–299. https://doi.org/10.1201/9781003278290-44

[27] Willian Dihanster Gomes de Oliveira and Lilian Berton. 2024. A comparative study of pre-processing algorithms for fair classification in few labeled data context. *AI and Ethics* (2024). https://doi.org/10.1007/s43681-024-00601-8

[28] Luca Deck, Jan-Laurin Müller, Conradin Braun, Domenique Zipperling, and Niklas Kühl. 2024. Implications of the {AI} Act for Non-Discrimination Law and Algorithmic Fairness. In *European Workshop on Algorithmic Fairness (EWAF'24)*. https://ceur-ws.org/Vol-3908/paper_39.pdf

[29] Luca Deck, Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. 2024. A Critical Survey on Fairness Benefits of {Explainable AI}. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://arxiv.org/abs/2310.13007

[30] Luca Deck, Astrid Schomäcker, Timo Speith, Jakob Schöffer, Lena Kästner, and Niklas Kühl. 2024. Mapping the Potential of Explainable AI for Fairness Along the AI Lifecycle. In *European Workshop on Algorithmic Fairness (EWAF'24)*. https://ceur-ws.org/Vol-3908/paper_38.pdf

[31] Marybeth Defrance and Tijl Bie. [n. d.]. Maximal Fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (New York,NY,United States) *(ACM Digital Library)*. Association for Computing Machinery, 851–880. https://doi.org/10.1145/3593013.3594048

[32] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. [n. d.]. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA) *(FAccT '22)*. Association for Computing Machinery, 473–484. https://doi.org/10.1145/3531146.3533113

[33] Manh Khoi Duong and Stefan Conrad. 2024. Towards Fairness and Privacy: A Novel Data Pre-processing Optimization Framework for Non-binary Protected Attributes. In *Data Science and Machine Learning*, Diana Benavides-Prado, Sarah Erfani, Philippe Fournier-Viger, Yee Ling Boo, and Yun Sing Koh (Eds.). Springer Nature Singapore, Singapore, 105–120.

[34] Manh Khoi Duong and Stefan Conrad. 2024. Trusting Fair Data: Leveraging Quality in Fairness-Driven Data Removal Techniques. In *Big Data Analytics and Knowledge Discovery*, Robert Wrembel, Silvia Chiusano, Gabriele Kotsis, A. Min Tjoa, and Ismail Khalil (Eds.). Lecture Notes in Computer Science, Vol. 14912. Springer Nature Switzerland, Cham, 375–380. https://doi.org/10.1007/978-3-031-68323-7{_}33

[35] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. [n. d.]. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12* (New York, New York, USA). ACM Press. https://doi.org/10.1145/2090236.2090255

[36] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 259–268.

[37] Emilio Ferrara. 2024. The Butterfly Effect in artificial intelligence systems: Implications for AI bias and fairness. *Machine Learning with Applications* 15 (2024), 100525. https://doi.org/10.1016/j.mlwa.2024.100525

[38] Emilio Ferrara. 2024. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci* 6, 1 (2024), 3. https://doi.org/10.3390/sci6010003

[39] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadoral, and Ansgar Walther. 2022. Predictably Unequal? The Effects of Machine Learning on Credit Markets. *The Journal of Finance* 77, 1 (2022), 5–47. https://doi.org/10.1111/jofi.13090

[40] Vahid Garousi, Michael Felderer, and Mika V. Mäntylä. 2019. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and Software Technology* 106 (2019), 101–121. https://doi.org/10.1016/j.infsof.2018.09.006

[41] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92. https://doi.org/10.1145/3458723

[42] Rubén González-Sendino, Emilio Serrano, and Javier Bajo. 2024. Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making. *Future Generation Computer Systems* 155 (2024), 384–401. https://doi.org/10.1016/j.future.2024.02.023

[43] Carlos Vladimiro Gonzalez Zelaya. 2019. Towards Explaining the Effects of Data Preprocessing on Machine Learning. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 2086–2090. https://doi.org/10.1109/icde.2019.00245

[44] D. Gray, D. Bowes, N. Davey, Yi Sun, and B. Christianson. 2011. The misuse of the NASA Metrics Data Program data sets for automated software defect prediction. In *15th Annual Conference on Evaluation & Assessment in Software Engineering (EASE 2011)*. IET, 96–103. https://doi.org/10.1049/ic.2011.0012

[45] Shubham Gujar, Tanishka Shah, Dewen Honawale, Vedant Bhosale, Faizan Khan, Devika Verma, and Rakesh Ranjan. 2022. GenEthos: A Synthetic Data Generation System With Bias Detection And Mitigation. In *2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*. IEEE, 1–6. https://doi.org/10.1109/ic3sis54991.2022.9885653

[46] N. R. Haddaway, M. J. Grainger, and C. T. Gray. 2021. citationchaser: An R package and Shiny app for forward and backward citation chasing in academic searching. https://doi.org/10.5281/zenodo.4543513. https://doi.org/10.5281/zenodo.4543513 Version 0.0.3.

[47] Moritz Hardt, Eric Mazumdar, Celestine Mendler-Dünner, and Tijana Zrnic. 2023. Algorithmic Collective Action in Machine Learning. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 12570–12586. https://proceedings.mlr.press/v202/hardt23a.html

[48] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *30th Conference on Neural Information Processing Systems (NIPS 2016)*.

[49] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *30th Conference on Neural Information Processing Systems (NIPS 2016)*. https://arxiv.org/pdf/1610.02413

[50] Jahid Hasan and Romila Pradhan. 2024. Data Acquisition for Improving Model Fairness using Reinforcement Learning. https://doi.org/10.48550/arXiv.2412.03009

[51] Kenneth Holstein, Jennifer Wortmann Vaughan, Hal Daumé, Miroslav Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

[52] Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. 2024. Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. *ACM Journal on Responsible Computing* 1, 2 (2024), 1–52. https://doi.org/10.1145/3631326

[53] Yu Huang, Jingchuan Guo, Wei-Han Chen, Hsin-Yueh Lin, Huilin Tang, Fei Wang, Hua Xu, and Jiang Bian. 2024. A scoping review of fair machine learning techniques when using real-world data. *Journal of biomedical informatics* 151 (2024), 104622. https://doi.org/10.1016/j.jbi.2024.104622

[54] Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of Test (Un)fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 49–58. https://doi.org/10.1145/3287560.3287600

[55] Vasileios Iosifidis and Eirini Ntoutsi. 2020. $$\mathsf FABBOO$$ - Online Fairness-Aware Learning Under Class Imbalance. In *Discovery Science*, Annalisa Appice, Grigorios Tsoumakas, Yannis Manolopoulos, and Stan Matwin (Eds.). Lecture Notes in Computer Science, Vol. 12323. Springer International Publishing, Cham, 159–174. https://doi.org/10.1007/978-3-030-61527-7{_}11

[56] Johannes Jakubik, Michael Vössing, Niklas Kühl, Jannis Walk, and Gerhard Satzger. 2024. Data-Centric Artificial Intelligence. *Business & Information Systems Engineering* 66, 4 (March 2024), 507–515. https://doi.org/10.1007/s12599-024-00857-8

[57] Lan Jiang, Clara Belitz, and Nigel Bosch. 2024. Synthetic Dataset Generation for Fairer Unfairness Research. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. ACM, New York, NY, USA, 200–209. https://doi.org/10.1145/3636555.3636868

[58] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33. https://doi.org/10.1007/s10115-011-0463-8

[59] Kimmo Kärkkäinen and Jungseock Joo. 2019. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. https://doi.org/10.48550/arXiv.1908.04913

[60] Harsurinder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. 2020. A Systematic Review on Imbalanced Data Challenges in Machine Learning. *Comput. Surveys* 52, 4 (2020), 1–36. https://doi.org/10.1145/3343440

[61] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35 International Conference on Machine Learning*.

[62] Farah Kidwai-Khan, Rixin Wang, Melissa Skanderson, Cynthia A. Brandt, Samah Fodeh, and Julie A. Womack. 2024. A roadmap to artificial intelligence (AI): Methods for designing and building AI ready data to promote fairness. *Journal of biomedical informatics* 154 (2024), 104654. https://doi.org/10.1016/j.jbi.2024.104654

[63] Dohyung Kim, Sungho Park, Sunhee Hwang, and Hyeran Byun. 2023. Fair classification by loss balancing via fairness-aware batch sampling. *Neurocomputing* 518 (2023), 231–241. https://doi.org/10.1016/j.neucom.2022.11.018

[64] Barbara Kitchenham and Stuart Charters. 2007. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. Technical Report EBSE-2007-01. EBSE Technical Report. https://www.cs.auckland.ac.nz/~norsaremah/2007%20SLR%20Guidelines.pdf

[65] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores: 67(43): 1–23. *Leibniz International Proceedings in Informatics* 67, 43 (2017), 1–23.

[66] Bernard Koch, Emily Denton, Alex Hanna, and Jacob G. Foster. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. http://arxiv.org/pdf/2112.01716v1

[67] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-aware Classification. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.). ACM Press, New York, New York, USA, 853–862. https://doi.org/10.1145/3178876.3186133

[68] Anna Langenberg, Shih-Chi Ma, Tatiana Ermakova, and Benjamin Fabian. 2023. Formal Group Fairness and Accuracy in Automated Decision Making. *Mathematics* 11, 8 (2023), 1771. https://doi.org/10.3390/math11081771

[69] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. 2022. A survey on datasets for fairness–aware machine learning. *WIREs Data Mining and Knowledge Discovery* 12, 3 (2022). https://doi.org/10.1002/widm.1452

[70] Junhak Lee, Dayeon Jung, Jihoon Moon, and Seungmin Rho. 2025. Advanced R-GAN: Generating anomaly data for improved detection in imbalanced datasets using regularized generative adversarial networks. *Alexandria Engineering Journal* 111 (2025), 491–510. https://doi.org/10.1016/j.aej.2024.10.084

[71] Nianyun Li, Naman Goel, and Elliott Ash. 2022. Data-Centric Factors in Algorithmic Fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, Vincent Conitzer, John Tasioulas, Matthias Scheutz, Ryan Calo, Martina Mara, and Annette Zimmermann (Eds.). ACM, New York, NY, USA, 396–410. https://doi.org/10.1145/3514094.3534147

[72] Yin Lin, Samika Gupta, and H. V. Jagadish. 2024. Mitigating Subgroup Unfairness in Machine Learning Classifiers: A Data-Driven Approach. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 2151–2163. https://doi.org/10.1109/icde60146.2024.00171

[73] Yingzhou Lu, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. 2023. Machine Learning for Synthetic Data Generation: A Review. http://arxiv.org/pdf/2302.04062v9

[74] Lydia R. Lucchesi, Petra M. Kuhnert, Jenny L. Davis, and Lexing Xie. 2022. Smallset Timelines: A Visual Representation of Data Preprocessing Decisions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 1136–1153. https://doi.org/10.1145/3531146.3533175

[75] Kiran Maharana, Surajit Mondal, and Bhushankumar Nemade. 2022. A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings* 3, 1 (2022), 91–99. https://doi.org/10.1016/j.gltp.2022.04.020

[76] Abdul Majeed and Seong Oun Hwang. 2024. Towards Unlocking the Hidden Potentials of the Data-Centric AI Paradigm in the Modern Era. *Applied System Innovation* 7, 4 (2024), 54. https://doi.org/10.3390/asi7040054

[77] Federico Mazzoni, Marta Marchiori Manerba, Martina Cinquini, Riccardo Guidotti, and Salvatore Ruggieri. 2023. GenFair: A Genetic Fairness-Enhancing Data Generation Framework. In *Discovery Science*, Albert Bifet, Ana Carolina Lorena, Rita P. Ribeiro, João Gama, and Pedro H. Abreu (Eds.). Lecture Notes in Computer Science, Vol. 14276. Springer Nature Switzerland, Cham, 356–371. https://doi.org/10.1007/978-3-031-45275-8{_}24

[78] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computer Surveys* 54, 6 (2021).

[79] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernandez, Harith Alani, Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. 2020. Bias in data–driven artificial intelligence systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery* 10, 3 (2020). https://doi.org/10.1002/widm.1356

[80] Tiago P. Pagano, Rafael B. Loureiro, Fernanda V. N. Lisboa, Rodrigo M. Peixoto, Guilherme A. S. Guimarães, Gustavo O. R. Cruz, Maira M. Araujo, Lucas L. Santos, Marco A. S. Cruz, Ewerton L. S. Oliveira, Ingrid Winkler, and Erick G. S. Nascimento. 2023. 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* 7, 1 (2023), 15. https://doi.org/10.3390/bdcc7010015

[81] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Journal of Clinical Epidemiology* 134 (2021), 178–189. https://doi.org/10.1016/j.jclinepi.2021.03.001

[82] José Daniel Pascual-Triana, Alberto Fernández, Paulo Novais, and Francisco Herrera. 2025. Fair Overlap Number of Balls (Fair-ONB): A Data-Morphology-based Undersampling Method for Bias Reduction. *Statistics* 54, 6 (2025), 1–19. https://doi.org/10.1080/02331888.2025.2476029

[83] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns (New York, N.Y.)* 2, 11 (2021), 100336. https://doi.org/10.1016/j.patter.2021.100336

[84] Kewen Peng, Joymallya Chakraborty, and Tim Menzies. 2023. FairMask: Better Fairness via Model-Based Rebalancing of Protected Attributes. *IEEE Transactions on Software Engineering* 49, 4 (2023), 2426–2439. https://doi.org/10.1109/TSE.2022.3220713

[85] Robert Lee Poe and Soumia Zohra El Mestari. 2024. The Conflict Between Algorithmic Fairness and Non-Discrimination: An Analysis of Fair Automated Hiring. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 1907–1916. https://doi.org/10.1145/3630106.3659015

[86] Gideon Popoola and John Sheppard. 2024. Investigating and Mitigating the Performance–Fairness Tradeoff via Protected-Category Sampling. *Electronics* 13, 15 (2024), 3024. https://doi.org/10.3390/electronics13153024

[87] Brianna Richardson, Jean Garcia-Gathright, Samuel F. Way, Jennifer Thom, and Henriette Cramer. [n. d.]. Towards Fairness in Practice: A Practitioner-Oriented Rubric for Evaluating Fair ML Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA) *(CHI '21)*. Association for Computing Machinery, 1–13. https://doi.org/10.1145/3411764.3445604

[88] Yuji Roh, Geon Heo, and Steven Euijong Whang. 2021. A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective. *IEEE Transactions on Knowledge and Data Engineering* 33, 4 (2021), 1328–1347. https://doi.org/10.1109/TKDE.2019.2946162

[89] Alveiro Alonso Rosado Gómez, Maritza Liliana Calderón Benavides, and Oscar Espinosa. 2024. Data preprocessing to improve fairness in machine learning models: An application to the reintegration process of demobilized members of armed groups in Colombia. *Applied Soft Computing* 152 (2024), 111193. https://doi.org/10.1016/j.asoc.2023.111193

[90] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney. 2019. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development* 63, 4/5 (2019), 3:1–3:9. https://doi.org/10.1147/jrd.2019.2945519

[91] Lele Sha, Dragan Gašević, and Guanliang Chen. 2023. Lessons from debiasing data for fair and accurate predictive modeling in education. *Expert Systems with Applications* 228 (2023), 120323. https://doi.org/10.1016/j.eswa.2023.120323

[92] Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and H. V. Jagadish. 2023. Representation Bias in Data: A Survey on Identification and Resolution Techniques. *Comput. Surveys* 55, 13s (2023), 1–39. https://doi.org/10.1145/3588433

[93] Maliheh Heidarpour Shahrezaei, Róisín Loughran, and Kevin Mc Daid. 2023. Pre-processing Techniques to Mitigate Against Algorithmic Bias. In *2023 31st Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*. IEEE, 1–4. https://doi.org/10.1109/AICS60730.2023.10470759

[94] Sunzida Siddique, Mohd Ariful Haque, Roy George, Kishor Datta Gupta, Debashis Gupta, and Md Jobair Hossain Faruk. 2024. Survey on Machine Learning Biases and Mitigation Techniques. *Digital* 4, 1 (2024), 1–68. https://doi.org/10.3390/digital4010001

[95] Jan Simson, Alessandro Fabris, and Christoph Kern. 2024. Lazy Data Practices Harm Fairness Research. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 642–659. https://doi.org/10.1145/3630106.3658931

[96] Prerna Singh. 2023. Systematic review of data-centric approaches in artificial intelligence and machine learning. *Data Science and Management* 6, 3 (2023), 144–157. https://doi.org/10.1016/j.dsm.2023.06.001

[97] Jessie J. Smith, Michael Madaio, Robin Burke, and Casey Fiesler. 2025. Pragmatic Fairness: Evaluating ML Fairness Within the Constraints of Industry. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, 628–638. https://doi.org/10.1145/3715275.3732040

[98] Ki Hyun Tae and Steven Euijong Whang. 2021. Slice Tuner: A Selective Data Acquisition Framework for Accurate and Fair Machine Learning Models. In *Proceedings of the 2021 International Conference on Management of Data* (Virtual Event, China) *(SIGMOD '21)*. Association for Computing Machinery, New York, NY, USA, 1771–1783. https://doi.org/10.1145/3448016.3452792

[99] Savaş Takan, Duygu Ergün, Sinem Getir Yaman, and Onur Kılınççeker. 2023. Bias in human data: A feedback from social sciences. *WIREs Data Mining and Knowledge Discovery* 13, 4 (2023). https://doi.org/10.1002/widm.1498

[100] Amal Tawakuli and Thomas Engel. 2024. Make your data fair: A survey of data preprocessing techniques that address biases in data towards fair AI. *Journal of Engineering Research* (2024). https://doi.org/10.1016/j.jer.2024.06.016

[101] António Trigo, Nubia Stein, and Fernando Paulo Belfo. 2024. Strategies to improve fairness in artificial intelligence:A systematic literature review. *Education for Information* 40, 3 (2024), 323–346. https://doi.org/10.3233/efi-240045

[102] P.S. Varsha. 2023. How can we manage biases in artificial intelligence systems – A systematic literature review. *International Journal of Information Management Data Insights* 3, 1 (2023), 100165. https://doi.org/10.1016/j.jjimei.2023.100165

[103] Sahil Verma, Michael Ernst, and Rene Just. 2021. Removing biased data to improve fairness and accuracy. https://doi.org/10.48550/arXiv.2102.03054

[104] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness*, Yuriy Brun (Ed.). ACM, 1–7. https://doi.org/10.1145/3194770.3194776

[105] Moritz von Zahn, Stefan Feuerriegel, and Niklas Kuehl. 2022. The Cost of Fairness in AI: Evidence from E-Commerce. *Business & Information Systems Engineering* 64, 3 (2022), 335–348. https://doi.org/10.1007/s12599-021-00716-w

[106] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2020. Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law. *West Virginia Law Review* 123 (2020), 735. https://heinonline.org/HOL/Page?handle=hein.journals/wvb123&id=765&div=&collection=

[107] Yanchen Wang and Lisa Singh. 2021. Analyzing the impact of missing values and selection bias on fairness. *International Journal of Data Science and Analytics* 12, 2 (2021), 101–119. https://doi.org/10.1007/s41060-021-00259-z

[108] Vitor Werner de Vargas, Jorge Arthur Schneider Aranda, Ricardo Dos Santos Costa, Paulo Ricardo Da Silva Pereira, and Jorge Luis Victória Barbosa. 2023. Imbalanced data preprocessing techniques for machine learning: a systematic mapping study. *Knowledge and Information Systems* 65, 1 (2023), 31–57. https://doi.org/10.1007/s10115-022-01772-8

[109] Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. 2023. Data collection and quality challenges in deep learning: a data-centric AI perspective. *The VLDB Journal* 32, 4 (2023), 791–813. https://doi.org/10.1007/s00778-

022-00775-9

[110] Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering* (London, England, United Kingdom) *(EASE '14)*. Association for Computing Machinery, New York, NY, USA, Article 38, 10 pages. https://doi.org/10.1145/2601248.2601268

[111] Claes Wohlin, Marcos Kalinowski, Katia Romero Felizardo, and Emilia Mendes. 2022. Successful combination of database search and snowballing for identification of primary studies in systematic literature studies. *Information and Software Technology* 147 (2022), 106908. https://doi.org/10.1016/j.infsof.2022.106908

[112] Shen Yan, Hsien-te Kao, and Emilio Ferrara. 2020. Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudre Mauroux (Eds.). ACM, New York, NY, USA, 1715–1724. https://doi.org/10.1145/3340531.3411980

[113] Zhe Yu, Joymallya Chakraborty, and Tim Menzies. 2021. FairBalance: How to Achieve Equalized Odds With Data Pre-processing. https://arxiv.org/abs/2107.08310. Preprint on arXiv:2107.08310.

[114] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact. In *Proceedings of the 26th International Conference on World Wide Web (ACM Digital Library)*. International World Wide Web Conferences Steering Committee, 1171–1180.

[115] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2025. Data-centric Artificial Intelligence: A Survey. *Comput. Surveys* 57, 5 (2025). https://doi.org/10.1145/3711118

[116] Yuhan Zhou, Fengjiao Tu, Kewei Sha, Junhua Ding, and Haihua Chen. 2024. A Survey on Data Quality Dimensions and Tools for Machine Learning Invited Paper. In *2024 IEEE International Conference on Artificial Intelligence Testing (AITest)*. IEEE, 120–131. https://doi.org/10.1109/AITest62860.2024.00023