

SL	Year	Title	Publication Venue	Dataset	Dataset Type	Descriptions	Classifier Used	Sensitive Attribute	Taxonomy	Domain	Evaluation Metrics	Library	Year	Summary
1	2018	FairGAN: Fairness-aware Generative Adversarial Networks	CIKM 2018 (published version), also available on arXiv	1. Toy dataset, 2. The UCI Adult	Tabular	Benchmarked on the UCI Adult Income dataset, compared with GAN, NaïveFairGAN-I, and NaïveFairGAN-II.	SVM(Linear) SVM(RBF) Decision Tree	Gender (e.g., male/female)	GAN	Finance	Uses Risk Difference and Balanced Error Rate (BER) to evaluate fairness of both synthetic data and trained classifiers.	Web Of Science	2018	Introduces FairGAN , a dual-discriminator GAN designed to generate fair tabular data. It ensures fairness in both data and classifiers by minimizing disparate treatment and disparate impact, and is extensively evaluated on UCI Adult data.
2	2018	Path-Specific Counterfactual Fairness	AAAI Conference on Artificial Intelligence, 2023	UCI Adult, German Credit, and Berkeley Admissions datasets,	Tabular		deep neural networks	Race or Gender	VAE-based	Healthcare, Finance, Education	MMD (Maximum Mean Discrepancy)	Web Of Science	2018	The paper proposes a novel fairness framework called path-specific counterfactual fairness, which selectively removes the unfair effects of sensitive attributes while preserving fair pathways. It introduces a latent inference-projection method that corrects descendants of sensitive attributes along unfair paths without discarding predictive information. It is validated through experiments on synthetic, Berkeley Admission, UCI Adult, and German Credit datasets.
3	2018	FairGAN: GANs-based Fairness-aware Learning for Recommendations with Implicit Feedback	ACM Web Conference 2022	UCI Adult dataset	Tabular	Tested on the UCI Adult dataset; compared against Naïve FairGAN, Original GAN, and real-data baselines	Random Forest, Logistic Regression, SVM	Gender, Race	GANs	Recommender Systems	– Statistical Parity – Disparate Impact (DI) – Accuracy of downstream classifiers trained on synthetic data	Scopus	2018	Introduces FairGAN, a generative adversarial model that synthesizes fair tabular data by removing disparate treatment and disparate impact. Includes two discriminators: one for data realism and one to enforce independence from sensitive attributes. Tested on the UCI Adult dataset, it shows improved fairness (Statistical Parity and Disparate Impact) while preserving classification accuracy.
4	2019	FairGAN+: Achieving Fair Data Generation and Classification through Generative Adversarial Nets	IEEE CIBCB (2019)	The UCI adult income dataset.	Tabular	Benchmarked on the UCI Adult dataset, compared against FairGAN, ACGAN, and Adversarial Debiasing methods	SVM (for BER)	Gender	GAN	Fair ML (census data)		Web Of Science	2019	Extends FairGAN by introducing FairGAN+ , which simultaneously trains a generator and classifier to achieve fairness in both data and predictions using three discriminators. Evaluated on the UCI Adult dataset with excellent fairness-utility trade-offs.
5	2019	It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution	EMNLP-IJCNLP 2019	Wikipedia and Annotated Gigaword corpora, with comparisons to WED (Bolukbasi et al.) and CDA (Lu et al., 2018) variants.	Text	Benchmarked on English Gigaword and Wikipedia. Compared with Word Embedding Debiasing (WED) and previous Counterfactual Data Augmentation (CDA) approaches.	SVM (RBF kernel)	Gender	None	NLP	Evaluates with WEAT (Word Embedding Association Test) for direct bias, cluster purity and gender reclassification accuracy for indirect bias, and analogy test accuracy for representation bias.	Web Of Science	2019	Proposes Counterfactual Data Substitution (CDS) and Names Intervention to debias word embeddings. Achieves a 49% reduction in gender-cluster purity and improved non-biased analogies while maintaining performance on core NLP tasks.
6	2019	<i>Fairness GAN: Generating Datasets with Fairness Properties Using a Generative Adversarial Network</i>	IBM Journal of Research and Development	CelebA, Soccer, and Quick, Draw	Image		Logistic Regression	Gender, Skin tone	GANs (AC-GAN variant)	Vision	Demographic Parity (DP) and Equality of Opportunity (EO)	Scopus	2019	Proposes a novel GAN-based approach for generating fair synthetic datasets from high-dimensional multimedia data. The model enforces demographic parity and equality of opportunity through adversarial loss components. Evaluations on CelebA, Soccer, and Quick, Draw! datasets show significant improvements in fairness while maintaining image quality. Compared with the Reweighting baseline, Fairness GAN demonstrates superior fairness outcomes.
7	2020	Queens are Powerful too: Mitigating Gender	EMNLP 2020	LIGHT dialogue dataset Uses a proprietary telecom	Text	Tabular (Telecom records: demographic, billing, call usage, service info)	Dialogue safety classifier	Gender	Generative Models: GPT-2 (language model, autoregressive); fine-tuned	NLP	Evaluates fairness using Statistical Parity Difference (SPD) and Disparate Impact (DI) across gender groups..	Web Of Science	2020	This study investigates how popular data sampling techniques (DSTs)—namely Random Over Sampler (ROS), Random Under Sampler (RUS), SMOTE, and ADASYN—affect algorithmic fairness in customer churn prediction (CCP) tasks. Using a real-world proprietary telecom dataset of 100,000 customer records, the authors simulate class

SL	Year	Title	Publication Venue	Dataset	Dataset Type	Descriptions	Classifier Used	Sensitive Attribute	Taxonomy	Domain	Evaluation Metrics	Library	Year	Summary
		Bias in Dialogue Generation		dataset (100K samples) from Telekom Malaysia.					Transformers					imbalance scenarios (5%, 15%, 30%) and apply DSTs before training six common classifiers. Fairness is evaluated using Statistical Parity Difference and Disparate Impact, with gender as the protected attribute. While oversampling techniques improved classification performance, they often led to higher fairness violations, especially in extremely imbalanced cases. The study concludes that although DSTs improve model accuracy, they can inadvertently exacerbate gender bias—underscoring the need for a careful trade-off between accuracy and fairness in pre-processing.
8	2020	Towards Accuracy-Fairness Paradox: Adversarial Example-based Data Augmentation for Visual Debiasing	28th ACM International Conference on Multimedia (ACM MM 2020)	C-MNIST and CelebA	Image	Evaluated on C-MNIST and CelebA, compared against Down-sampling, Reweighting, Adv Debiasing, CycleGAN	VGG-16, Bias Classifier (CNN-based), CycleGAN, Adversarial Classifiers (AEDA variants)	Gender	GANs (CycleGAN), Adversarial Example Generators (AEDA),	Vision	Equality of Opportunity and group fairness: compares true positive rates between groups.	Web Of Science	2020	The paper proposes AEDA, a method that generates adversarial examples to augment data for visual debiasing. The approach improves both fairness and accuracy by balancing biased data distributions in real-world (CelebA) and simulated (C-MNIST) datasets. It introduces three variants: AEDA_pre, AEDA_online, and AEDA_robust, evaluated using model bias and balanced accuracy.
9	2021	Conditionally Independent Data Generation	UAI 2021 (Uncertainty in Artificial Intelligence)	Adult Census Income dataset	Tabular		Logistic Regression, Decision Tree, Naive Bayes, SVM	Gender	GANs	Finance, Criminal Justice	Evaluates fairness using metrics like Equalized Odds Difference (EOD) and Maximum Conditional Statistical Dispersion (MCSD) across protected groups and admissible variables.	Web Of Science	2021	Proposes a GAN-based approach to generate synthetic data that satisfies conditional independence constraints, enabling fairness-aware generation aligned with criteria like equalized odds and conditional statistical parity. Offers strong theoretical basis and empirical validation.
10	2021	Constructing a Fair Classifier with Generated Fair Data	AAAI 2021	Adult, COMPAS, German Credit, and MEPS datasets.	Tabular	Benchmarked against 6 fairness-aware models (e.g., FairGAN, AdvDeb, MFC, LAFTR) using Adult, COMPAS, German Credit, and MEPS datasets.	Logistic Regression, Decision Tree, SVM, XGBoost	Gender, Race	VAEs, GANs (VAE-GAN)	Fair ML	Absolute Equal Opportunity Difference, Absolute Average Odds Difference, Statistical Parity Difference, and Balanced Accuracy Difference.	Web Of Science	2021	Proposes a VAE-GAN-based model to generate perfectly balanced synthetic datasets across sensitive groups and labels. Trains classifiers on synthetic data and uses transfer learning on real data to ensure both fairness and accuracy. Validated on four benchmark datasets with strong empirical results.
11	2021	Counterfactual Fairness with Disentangled Causal Effect VAE	AAAI 2021	Tabular (Adult Income), Image (CelebA)	Tabular, Image		Logistic Regression, SVM	Gender, Race	VAEs (DCEVAE)	Healthcare, Finance	Evaluates fairness through Total Effect (TE), Counterfactual Effect (CE), CE Error, and also classification accuracy on generated fair data.	Web Of Science	2021	Proposes DCEVAE, a VAE-based model that disentangles causal and correlated latent variables to generate counterfactual fair data. Evaluated on UCI Adult and CelebA datasets using counterfactual and fairness metrics.
12	2021	Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3D deep generative models	Medical Image Analysis, Elsevier	UK Biobank MRI dataset	Image	UK Biobank MRI dataset. Compares against methods such as DRO, ERM, and Confounder-Free Networks (CF-Net).	CNN-based	Age, Sex	Age, sex (sensitive); WMH volume (targeted)	Vision Healthcare	DSCΔ (subgroup disparity), DSCSTD (standard deviation), DSCSKEW (skewness), and Jensen-Shannon divergence on radiomics parameter	Web Of Science	2021	Introduces CounterSynth , a 3D generative model that synthesizes anatomically plausible counterfactual brain images conditioned on demographic attributes, improving equity and robustness in MRI-based predictive tasks.
13	2021	Fairness without the Sensitive Attribute via Causal Variational Autoencoder	AAAI 2021	UCI adult dataset	Tabular	Benchmarked on Adult UCI and Default datasets. Compared against FairRF, ProxyFairness, ARL, and models trained with true sensitive attributes.	Neural network-based	Gender, race (sensitive); Income (targeted)	VAEs (Causal VAE)	Finance	evaluates fairness using Demographic Parity (P-rule) and Equalized Odds (ΔFPR , ΔFNR) across multiple configurations.	Web Of Science	2021	Proposes SRCAE , a causal VAE that reconstructs sensitive proxies from observed data using expert-informed causal graphs. These proxies are then used for adversarial fairness training without requiring real sensitive attributes, outperforming prior methods in both fairness and accuracy.
14	2021	Learning Disentangl	AAAI 2021	(Facial attribute	Image	Benchmarked on CelebA and UTK	Neural network-	Gender, age, ethnicity	VAEs (FD-	Vision	Equal Opportunity, Equalized Odds, and introduces a new	Web Of Science	2021	Proposes FD-VAE , which disentangles representations into target, protected, and mutual

SL	Year	Title	Publication Venue	Dataset	Dataset Type	Descriptions	Classifier Used	Sensitive Attribute	Taxonomy	Domain	Evaluation Metrics	Library	Year	Summary
		ed Representation for Fair Facial Attribute Classification via Fairness-aware Information Alignment		datasets: CelebA, UTK Face)		Face datasets; compared with VAE, β -VAE, FactorVAE, FFVAE, adversarial training, and more.	based classifiers	(sensitive); Attractiveness, ethnicity, age (targeted)	VAE)		metric: Equalized Accuracy.			latent spaces. Includes a novel decorrelation loss and a fairness-aware downstream classifier. Evaluated on CelebA and UTK Face , achieving strong fairness improvements using standard and proposed fairness metrics.
15	2021	<i>Towards Fair Federated Learning with Zero-Shot Data Augmentation</i>	CVPRW 2021	MNIST, FMNIST, CIFAR-10	image		Neural network-based	Client subgroups (e.g., gender, age, ethnicity)	Zero-shot Data Augmentation	Vision	Uses variance of client accuracies (client-level fairness) and variance of class-wise accuracy (class-level fairness). Lower variance = better fairness.	Web Of Science	2021	Proposes a zero-shot data generation (ZSDG) framework to generate synthetic image data in federated learning settings where client data is non-IID or scarce. The method improves both client-level and class-level fairness using synthetic data. Fairness is evaluated using variance in accuracy across clients and classes on MNIST, FMNIST, and CIFAR-10.
16	2021	<i>On the Fairness of Generative Adversarial Networks (GANs)</i>	NIR 2021	MNIST (digit images), SVHN (house number images), and CelebA (facial images)	Image	Experiments on MNIST, SVHN, and CelebA datasets. GAN variants like SGAN and PGAN are compared.	Neural network-based discriminators in GANs	Gender, Skin tone	GANs (Stacked GAN)	Vision	Fairness is evaluated via distributional analysis across demographic groups (e.g., ratio of black vs white digits, dark vs light skin). While not standard metrics like SPD or EOD, it uses meaningful representation balance	IEEE	2021	This paper identifies bias in GAN-generated image data when group distributions (e.g., skin tone, background color) are imbalanced. It proposes a GAN ensemble (boosting-style) to generate synthetic data more fairly across groups. Experiments on MNIST, SVHN, and CelebA show that even balanced training data can lead to biased generation, which the proposed method helps to mitigate.
17	2021	<i>Accuracy and Fairness in a Conditional Generative Adversarial Model of Crime Prediction</i>	BESC 2020	SIEDCO crime dataset	Image	Uses SIEDCO crime dataset (official Colombian crime data) and compares with existing spatio-temporal models like self-exciting Poisson point processes. Though not globally standard, it is real, domain-relevant data.	GAN-based classifiers	Income Level (Stratum)	GANs	Criminal Justice	Uses calibration-based fairness: compares predicted probabilities with observed frequencies and assesses bias across income groups	IEEE	2021	Proposes a ConvLSTM-based cGAN for spatio-temporal crime prediction in Bogotá. The paper addresses fairness by introducing a calibration penalty (MMCE) during training to reduce prediction bias across income groups. Evaluation is done via calibration tests, and AUC performance reaches 0.86. Data is from the official Colombian crime registry (SIEDCO).
18	2021	<i>COUNTERGAN: Generating Realistic Counterfactuals with RGANs</i>	38th Conference on Uncertainty in Artificial Intelligence (UAI 2022)	Uses MNIST (image), Pima Diabetes & COMPAS (tabular).	Tabular, Image		Neural network-based (target classifier, discriminator)	Race, Gender	GANs (Residual GAN)	Healthcare, Finance, Criminal Justice	Fairness and utility assessed using Prediction Gain, Realism (AE Error), Actionability (L1 norm), and Latency. Bias exposure is shown through counterfactual examples (e.g., changing race).	Scopus	2021	Proposes CounteRGAN, a Residual GAN that generates realistic and actionable counterfactuals by learning perturbations. Evaluates fairness using prediction gain, realism (autoencoder error), actionability (L1 norm), and latency. Applied to MNIST, Pima Diabetes, and COMPAS datasets. Highlights how counterfactuals expose model biases and improve interpretability. Outperforms baselines in fairness, speed, and realism.
19	2021	<i>Fairness for Image Generation with Uncertain Sensitive Attributes</i>	International Conference on Machine Learning (ICML) 2021	RDP, PR, CPR, SPE	Image	MNIST, FFHQ, AFHQ	StyleGAN2	Ethnicity, Species (Cats/Dogs)	GANs (StyleGAN2), Diffusion Models (Posterior Sampling via Langevin Dynamics)	Vision		Scopus	2021	The paper introduces a sampling-based generative framework using Langevin dynamics that ensures fairness when sensitive attributes (e.g., race, gender) are uncertain or not explicitly available. It proposes fairness metrics tailored to generative models (RDP, PR, CPR, SPE) and evaluates five configurations of image upsampling models on datasets like MNIST, FFHQ, and AFHQ. The study demonstrates that conventional methods often produce biased outputs, while their method achieves better proportional representation and symmetry in the generated samples.
20	2021	DECAF: Generating Fair Synthetic Data Using Causally-Aware	(NeurIPS 2021)	Adult, COMPAS, and Law	Tabular	Benchmarks conducted on Adult, COMPAS, and Law School	MLP (Multi-Layer Perceptron)	Gender, Race	GANs (DECAF GAN)	Finance	Evaluates with Counterfactual Fairness (CF), Demographic Parity (DP), and Equalized Odds (EO).	Web of Science	2021	The paper proposes DECAF, a generative model that uses structural causal models (SCMs) to produce counterfactually fair synthetic tabular data. It intervenes on the causal graph to remove unfair influences from sensitive attributes. Evaluations on datasets like Adult, COMPAS, and Law School show that DECAF improves demographic parity and

SL	Year	Title	Publication Venue	Dataset	Dataset Type	Descriptions	Classifier Used	Sensitive Attribute	Taxonomy	Domain	Evaluation Metrics	Library	Year	Summary
		Generative Networks				School datasets. DECAF is compared against FairGAN, Naive GANs, and other causal models like CF-GAN.								equalized odds compared to GAN-based and naive baselines.
21	2022	Investigating Bias with a Synthetic Data Generator	IJCAI-23	Synthetic	Tabular	All experiments are conducted on synthetic datasets only; no real-world dataset benchmarks or comparisons to prior methods.	Random Forest (RF, BRF, ERF)	Sensitive: A (ethnicity, gender); Target: Y (debt repayment, work performance)	Fully synthetic data (parametric: Bernoulli, Gamma, Binomial, Normal)	Finance	Evaluates fairness using Demographic Parity difference (ΔDP) and ΔAcc (Accuracy gap between sensitive groups).	Web Of Science	2022	Proposes a flexible framework to generate synthetic datasets with various combinations of biases. Combines empirical experiments and philosophical discussion to evaluate fairness strategies. Evaluates 25 configurations and provides a public toolkit for controlled bias modeling and fairness testing.
22	2022	Augmentations in Hypergraph Contrastive Learning: Fabricated and Generative	NeurIPS 2022	Evaluated across 13 benchmark datasets (e.g., Cora, Pubmed, Yelp),	Tabular		HyperGNNs (HGNN, HyperGCN, AllSet)	Node labels	VAEs (VHGAE)	Finance	Evaluates fairness using ΔSP (Statistical Parity) and ΔEO (Equalized Odds) across three curated datasets (German, Recidivism, Credit Defaulter).	Web Of Science	2022	Proposes HyperGCL, a framework for contrastive learning on hypergraphs. Introduces both handcrafted and generative augmentations, including the first variational hypergraph autoencoder. Evaluates extensively for fairness, robustness, and utility on 13 datasets.
23	2022	ImpartialGAN: Fair and Unbiased Classification	IEEE IRI 2022 (International Conference on Information Reuse and Integration for Data Science)	The UCI Adult Dataset	Tabular	Benchmarked on the UCI Adult dataset and compared with FairGAN. Multiple configurations with classifiers (SVM, RBF, Decision Tree) are tested.	GAN-based (ImpartialGAN)	Sensitive: s (gender, race); Target: y (income, loan approval)	GANs (FairGAN, medGAN)	Finance, Employment, Healthcare	Uses Risk Difference between groups (protected vs. unprotected) on synthetic data and classifier outputs.	Web Of Science	2022	Proposes ImpartialGAN , an enhanced GAN with three discriminators to generate fair synthetic data by eliminating correlations between protected and unprotected attributes. Outperforms FairGAN in risk difference while maintaining good accuracy.
24	2022	OpenXAI: Towards a Transparent Evaluation of Post-hoc Model Explanations	Neural Information Processing Systems (NeurIPS), 2022	Adult, COMPAS, German Credit.	Tabular		Neural networks, Gradient-based (Vanilla Gradients, Gradient x Input, SmoothGrad, Integrated Gradients)	Sensitive: Protected attributes (e.g., race, gender); Target: Model predictions (e.g., loan approval, recidivism)	Synthetic data generators	Healthcare, Finance, Science Criminal justice	using 11 group-based disparity metrics across subgroups	Web Of Science	2022	The paper introduces OpenXAI, an open-source framework for systematically benchmarking post hoc explanation methods. It includes a novel synthetic data generator, multiple real-world datasets, and 22 evaluation metrics covering faithfulness, stability, and fairness. The framework highlights disparities in explanation quality across demographic groups and provides leaderboards for transparent comparison of explanation methods. OpenXAI supports continuous community-driven updates and is highly relevant to fair data generation and fairness assessment.
25	2022	TabFairGAN: Fair Tabular Data Generation with Generative Adversarial Networks	Machine Learning and Knowledge Extraction (MAKE), MDPI, 2022	Adult, COMPAS, Bank Marketing, and Law School datasets.	Tabular	Benchmarked against CRDI, CTGAN, and TGAN using Adult, COMPAS, Bank Marketing, and Law School datasets.	GAN-based (TabFairGAN, WGAN)	Sensitive: S (protected, e.g., race, gender); Target: Y (decision, e.g., recidivism, income)	GANs (WGAN, TGAN, CTGAN)	Criminal Justice, Finance, Healthcare, Advertising	Discrimination Score (DS)	Others	2022	The paper introduces TabFairGAN, a Wasserstein GAN with two-phase training: one for learning the data distribution and the other to enforce demographic fairness using a modified loss function. It demonstrates superior fairness and competitive utility across Adult, COMPAS, Bank Marketing, and Law School datasets, outperforming CRDI and CTGAN in reducing Discrimination Score.
26	2023	A Fair Generative Model Using LeCam Divergence	AAAI 2023	CelebA, UTKFace, FairFace	Image	Real facial image datasets (CelebA, UTKFace, FairFace) with sensitive attributes like gender and race.	GAN-based (GAN, Conditional GAN, StyleGAN)	Sensitive: z (demographics, e.g., gender, race); Target: Generated samples (images)	GANs (LC-GAN)	Computer Vision, Healthcare	Fairness Discrepancy (L2 norm of demographic group imbalance) Intra-FID (Frechet Inception Distance for individual groups)	Web Of Science	2023	The paper proposes a GAN-based framework for fair data generation without relying on demographic labels. It introduces LeCam divergence as a fairness regularizer, which remains effective even with a small reference dataset. The method shows improved fairness discrepancy and sample quality across benchmark datasets like CelebA, UTKFace, and FairFace, outperforming prior state-of-the-art approaches.
27	2023	A Novel Fairness-Aware Ensemble Model Based on	Finance, Criminal Justice	UCI German UCI Adult UCI Bank Compas	Tabular		Ensemble (LightGBM, XGBoost, RF, GBDT, AdaBoost, Stacking)	Sensitive: Sensitive attributes (e.g., race, gender); Target: Label (e.g., credit)	Oversampling (SMOTE, ADASYN)	Finance, Healthcare, Criminal Justice	Benchmarked on German, Adult, Bank, and Compas datasets from UCI and ProPublica. Compared with models by Kearns, Kamiran, Kamishima, and Pleis	Web Of Science	2023	Proposes FAEM , an ensemble model that balances sensitive attributes via hybrid sampling (cross-validated undersampling + ADASYN oversampling) and adjusts predictions in stacking for fairness. Validated on 4 datasets, showing high fairness gains with minimal accuracy loss.

SL	Year	Title	Publication Venue	Dataset	Dataset Type	Descriptions	Classifier Used	Sensitive Attribute	Taxonomy	Domain	Evaluation Metrics	Library	Year	Summary
		Hybrid Sampling and Modified Stacking for fair Classification						approval, recidivism)						
28	2023	Fair Generative Models via Transfer Learning	AAAI 2023	CelebA, UTKFace, and FFHQ	Tabular	Benchmarked on CelebA, UTKFace, and FFHQ, comparing with SOTA baseline (Choi et al., 2020).	GAN-based (fairTL, fairTL++)	Sensitive: Gender, BlackHair; Target: Generated samples (images)	GANs	Computer Vision, healthcare)	Fairness Discrepancy (FD), a numerical metric based on deviation from uniform SA distributions	Web Of Science	2023	Proposes fairTL and fairTL++, transfer learning approaches to train fair GANs using large biased and small fair datasets. The models adapt pre-trained generators to align with fair distributions of sensitive attributes without requiring labeled data. Achieves SOTA results in both sample quality (FID) and fairness (FD), even under resource-constrained setups.
29	2023	Bias On Demand: Investigating Bias with a Synthetic Data Generator	IJCAI 2023 (Demonstrations Track)	Synthetic	Tabular	Focuses entirely on synthetic datasets generated by the framework; no use of standard datasets or direct model baselines.	Random Forest	Sensitive: A (e.g., gender, ethnicity); Target: Y (e.g., college admission, loan approval)	GANs, Diffusion models, Autoencoders, GMMs, HMMs	Education, Financial	Evaluates bias and fairness using metrics like Demographic Parity, Accuracy Difference, and fairness-aware performance assessments.	Web Of Science	2023	Proposes BiasOnDemand, a flexible toolkit for generating synthetic datasets embedded with controlled levels of historical, measurement, and representation bias. Offers formal bias modeling and public code for fairness evaluation. Enables controlled experiments in fairness research.
30	2023	Fair-CDA: Continuous and Directional Augmentation for Group Fairness	AAAI 2023	Tabular (Adult), Image (CelebA), Recommendation (MovieLens)	Tabular, Image		Neural networks (MLP, ResNet-18, LightGCN)	Sensitive: Gender, Supplier (minority/majority); Target: Salary >50K, Smiling/Wavy Hair/Attractive, Recommendation	GANs (FairGAN); VAEs (β -VAE, FFVAE)	Employment, Computer Vision, Recommendation System, Education	Evaluates using Demographic Parity (ΔDP) and Equalized Odds (ΔEO), both well-established fairness metrics.	Web Of Science	2023	Proposes a two-stage method that decomposes features into sensitive and task-relevant parts, then applies directional semantic augmentation to enforce fairness. Demonstrates state-of-the-art fairness and utility trade-off on diverse tasks. Deployable at scale.
31	2023	Contrastive Mixture of Posteriors for Counterfactual Inference, Data Integration and Fairness	NeurIPS 2022	Adult, CelebA, and Colored MNIST,	Tabular, Image	Statistical Parity, Demographic Parity, and Equalized Odds as downstream metrics, alongside performance and representation disentanglement.	Random Forest	Sensitive: Batch, Gender, Stimulation; Target: Cancer type, Income, Gene expression	VAEs (CVAE, VFAE, trVAE, CoMP)	Computational Biology, Fairness		Web Of Science	2023	Proposes CoMP, a generative VAE-based model that uses contrastive learning to enforce counterfactual fairness and disentangle causal representations. Generates data representations that support fairness and robustness across tasks such as classification and domain integration.
32	2023	CREST: A Joint Framework for Rationalization and Counterfactual Text Generation	ACL 2023	IMDB and SNLI datasets	Text		Transformer-based (RoBERTa, T5, GPT-2)	Sensitive: N/A; Target: Labels (sentiment, NLI)	GANs, VAEs, Masked LMs (T5)	NLP		Web Of Science	2023	Introduces CREST, which unifies counterfactual text generation with selective rationalization. Generates high-quality counterfactuals, then uses them for data augmentation and agreement-based rationale training , leading to improved robustness and interpretability across diverse NLP tasks.
33	2023	Fairness in Face Presentation Attack Detection	Pattern Recognition, 2023	SiW-M face anti-spoofing dataset and FairFace dataset,	Image	using demographic subgroup accuracy gaps (e.g., False Acceptance Rate and False Rejection Rate across gender and skin tone groups), and shows	CNN-based (ResNet50, VGG16)	Sensitive: Gender, Race, Appearance (eyeglasses, beard); Target: PAD (bona fide vs. attack)	Data augmentation (FairSWAP)	Biometrics (Face Recognition, PAD),, Vision		Web Of Science	2023	Introduces the CAAD-PAD dataset , a diverse annotated benchmark for studying bias in PAD systems. Proposes FairSWAP , a cross-attribute patch-based data augmentation method that boosts PAD fairness and performance across demographic and non-demographic groups

SL	Year	Title	Publication Venue	Dataset	Dataset Type	Descriptions	Classifier Used	Sensitive Attribute	Taxonomy	Domain	Evaluation Matrics	Library	Year	Summary
						reduction in bias after augmentation.								
34	2023	Fairness-and Uncertainty-Aware Data Generation for Data-Driven Design	Journal of Computing and Information Science in Engineering, May 2024		Tabular	Benchmarked against grid and randomized (LHS) sampling methods on the S-slot design problem (auxetic metamaterial structure).	ML models (surrogate, generative, Bayesian optimization)	Sensitive: Properties (underrepresented regions); Target: Designs (shapes, parameters)	Generative models (oversampling, synthetic generation)	Mechanical Design, Metamaterials, Structural Design	Introduces a custom data coverage score as a fairness metric based on Voronoi diagram coverage, representing representational fairness in the design's property space	Web Of Science	2023	Proposes FairGen , a fairness-aware and uncertainty-guided data generation framework for augmenting mechanical design datasets. Demonstrated on auxetic materials, it improves data fairness and model accuracy with fewer samples.
35	2023	FairPRS: Adjusting for Admixed Populations in Polygenic Risk Scores using Invariant Risk Minimization	Pacific Symposium on Biocomputing (PSB) 2023	UK Biobank (ePRS-UKB)	Tabular	Benchmarked using UK Biobank (ePRS-UKB) and diverse simulated datasets. Compared against standard PRS methods (PRSci2) and TL-PRS (transfer learning).	IRM, Autoencoder, MLP	Sensitive: Ancestry (European, African, Asian); Target: Phenotype (height, BMI, HbA1c, HDL, LDL)	Synthetic (Balding-Nichols, PSD, TGP)	Genomics, Precision Medicine., Healthcare	evaluates with metrics such as Net Reclassification Index (NRI), adjusted R ² , Kolmogorov-Smirnov (KS) tests, and ancestry-stratified performance comparisons.	Web Of Science	2023	Proposes FairPRS , a fairness-driven data generation framework using IRM and an autoencoder to produce ancestry-agnostic PRS estimates . Evaluated on UK Biobank and synthetic data, FairPRS improves phenotype prediction while mitigating racial bias.
36	2023	High-resolution synthesis of high-density breast mammograms: Application to improved fairness in deep learning based mass detection	Medical Image Analysis (Elsevier), 2024	OPTIMAM	Image	Benchmarked against real-only and oversampled models using standard datasets like OPTIMAM. Compares to baseline U-Net and U-Net trained with oversampling.	Deep learning (CNN-based mass detection)	Sensitive: Breast density (BI-RADS D); Target: Mass detection	GANs (CycleGAN)	Health care, Vision	Fairness is evaluated using Balanced Accuracy and sensitivity/specificity disaggregated by density class, showing improved detection fairness on BI-RADS D cases.	Web Of Science	2023	Proposes the use of CycleGANs to synthesize high-density FFDM images (BI-RADS D) for underrepresented breast types. The synthetic images are used to improve fairness and generalization in AI mass detection models, especially when real BI-RADS D samples are scarce. Evaluated on multiple datasets (OPTIMAM, CSAW, BCDR, INbreast), with gains in AUC and a radiologist reader study confirming realism.
37	2023	A Deep Generative Recommendation Method for Unbiased Learning from Implicit Feedback	ACM SIGIR ICTIR 2023	MovieLens-1M, Yahoo! R3,	Text	Benchmarked on MovieLens-1M, Yahoo! R3, and synthetic datasets. Compared against MF-Dual, Rel-MF, VAE, and other debiasing baselines.	VAEs (Mult-VAE, VAE-IPS)	Sensitive: Selection bias (popularity, position); Target: Implicit feedback (clicks, items)	VAEs	Recommender Systems	It evaluates fairness via bias-corrected performance metrics (e.g., MAP@5, NDCG@5 under unbiased sampling and IPS corrections), indicating bias mitigation.	Web Of Science	2023	Proposes VAE-IPS , which debiases VAE-based recommendation models using inverse propensity scoring. Combines fairness-aware estimation with generative latent modeling across implicit feedback setups. Outperforms prior models in MAP and NDCG under multiple bias settings.
38	2023	Leveraging Domain Knowledge for Inclusive and Bias-aware Humanitarian Response Entry Classification	International Joint Conference on Artificial Intelligence, 2023	HUMSET	Text	Benchmarked on HUMSET using multiple LLMs (m-BERT, XLM-R, HumBERT), and compared against various bias-mitigation configurations (baseline, CDA, proposed architecture).	LLMs (XLM-RoBERTa, HumBERT)	Sensitive: Societal biases (gender, geography, age); Target: Humanitarian entries (framework categories)	Counterfactual augmentation	NLP	Fairness is measured using Tag-Shift and Overall-Shift metrics based on model sensitivity to protected attribute substitutions. Bias is quantified pre- and post-CDA (Counterfactual Data Augmentation).	Web Of Science	2023	Proposes HumBERT, a domain-specific multilingual LLM for humanitarian response, and introduces HUMSETBIAS, a dataset focused on gender and country bias. Uses counterfactual data augmentation to reduce model bias, while improving performance in both zero-shot and supervised setups. Evaluation shows significant fairness improvements without accuracy loss.
39	2023	Mitigating Voter Attribute Bias for	1st Workshop on Fairness and Bias in AI, ECAI 2023	Crowd Judgment and Jigsaw Toxicity datasets	Tabular	Benchmarked on Crowd Judgment and Jigsaw Toxicity datasets. Compared with	Majority voting, Dawid-Skene (D&S, Soft D&S)	Sensitive: Voter attributes (gender, race); Target: Aggregated	Synthetic data generation	Decision-making, Crowdsourcing, Fairness	Uses metrics like Utility Loss, Representation Disparity, and Fair Aggregation Score to evaluate trade-offs between fairness and accuracy.	Web Of Science	2023	Proposes Soft D&S, a fairness-aware label aggregation model, to address voter attribute bias in crowdsourced labeling. Combines data splitting, sample weighting, and GroupAnno for fairer opinion aggregation. Fairness is evaluated using synthetic

SL	Year	Title	Publication Venue	Dataset	Dataset Type	Descriptions	Classifier Used	Sensitive Attribute	Taxonomy	Domain	Evaluation Metrics	Library	Year	Summary
		Fair Opinion Aggregation				Naïve Aggregation, FairExpert, GroupBalancer, and other bias-correcting techniques.		labels/opinions						and semi-synthetic datasets, including Moral Machine, showing Soft D&S with data splitting and weighted majority voting outperform baselines.
40	2023	On Improving Fairness of AI Models with Synthetic Minority Oversampling Techniques	SIAM International Conference on Data Mining (SDM '23), April 27–29, 2023, Minneapolis, Minnesota, USA	UCI Adult Compas German Credit Medical Expanse Bank data	Tabular		Logistic Regression, SVM, Decision Trees	Sensitive: Sensitive features (gender, race); Target: Labels (hiring, credit approval)	Oversampling (SMOTE, fair-SMOTE)	AI ML Fair (Employment, Healthcare)		Web Of Science	2023	The paper provides theoretical and empirical support for using SMOTE-based oversampling to generate synthetic data that mitigates bias in AI models. It addresses both label and selection bias, offering fairness improvements across multiple datasets and fairness metrics. The method is compared with pre-, in-, and post-processing debiasing techniques and consistently shows strong fairness improvement with minimal accuracy loss.
41	2023	Providing Previously Unseen Users Fair Recommendations Using Variational Autoencoders	17th ACM Conference on Recommender Systems (RecSys 2023)	MovieLens 1M and LastFM 2b	Tabular	Benchmarks on MovieLens 1M and LastFM 2b, comparing against SLIM and base VAE.	VAEs	Sensitive: Demographics (gender, age); Target: Recommendations (items)	VAEs	Recommender Systems	Evaluation Uses AUC to assess sensitive attribute leakage, χ^2 -statistics and Kendall-Tau for group-wise recommendation distribution similarity, and NDCG for utility-fairness trade-off.	Web Of Science	2023	This paper proposes fairness-aware recommender systems using Variational Autoencoders (VAEs) that generate synthetic latent representations for unseen users. It introduces adversarial and split latent setups to reduce demographic bias in recommendations. Fairness is measured using AUC on sensitive attributes, χ^2 -statistics, and Kendall-Tau distance. The system provides competitive fairness and recommendation utility, with an added sampling feature for dynamic fairness-performance tradeoff control.
42	2023	Rectifying Unfairness in Recommendation Feedback Loop	WSDM 2024 (ACM International Conference on Web Search and Data Mining)	MovieLens-1M and Yelp datasets.	Tabular	Benchmarked on MovieLens-1M and Yelp datasets. Compared with baselines including IRM, DebiasRec, InF-VAE, and others.	VAE-based (B-FAIR)	Sensitive: Sensitive context (bias attributes); Target: Relevance scores/recommendations	VAEs	Recommender Systems	Evaluates with fairness metrics such as Exposure Parity (EP), Opportunity Bias (OB), and Catalog Coverage, along with utility metrics like NDCG.	Web Of Science	2023	This paper proposes B-FAIR, a two-stage debiasing framework that rectifies feedback loops in recommendation systems using a new fairness objective called Balanced Fairness Objective (BFO). The model uses synthetic exposure and user feedback to simulate fair training scenarios and employs adversarial learning to enforce fairness in representations. Extensive experiments on synthetic, MovieLens, and Insurance datasets show that B-FAIR significantly improves fairness performance across multiple sensitive groups while maintaining predictive accuracy.
43	2023	SAGAN: Maximizing Fairness using Semantic Attention Based GAN	ICIBA 2023	UCI Adult Dataset	Tabular	Dataset and compares performance with prior adversarial debiasing methods like those in [15], [16].	GAN-based (SAGAN)	Sensitive: race, sex; Target: Predictions (income >50K)	GANs	Algorithmic Decision-making, Finance	Uses Disparate Impact (DI) and Demographic Parity to evaluate fairness. Achieves DI ≈ 0.99 for race and 0.92 for sex.	IEEE	2023	Introduces SAGAN, an adversarial fairness-aware model that uses a Semantic Attention (SA) module to select proxy features. It trains a predictor to maximize accuracy while an adversary is trained to suppress sensitive feature leakage. Evaluated on the UCI Adult Dataset using Disparate Impact (DI) and Demographic Parity, achieving high fairness with minimal accuracy tradeoff.
44	2023	Counterfactual Fairness on Graphs: Augmentations, Hidden Confounders, and Identifiability	Transactions on Machine Learning Research (TMLR)	German Credit, Credit Defaulter, and Synthetic Graphs	Tabular	Experiments on German Credit, Credit Defaulter, and Synthetic Graphs. Compared against GCN, FairGNN, CF-GNNExplainer, and ALFR.	GCN (Graph Convolutional Networks)	Gender (German), Age (Credit), Race (Bail)	VAEs	Finance, Criminal Justice	Evaluates with Counterfactual Fairness metrics (consistency of predictions under counterfactual graph inputs).	Scopus	2023	Proposes a graph-based counterfactual generation framework using latent confounders and counterfactual augmentations. The goal is to ensure counterfactual fairness in graph neural networks by learning representations that remove sensitive attribute influence. Evaluated on credit and synthetic graph datasets using counterfactual fairness measures, outperforming GCN, FairGNN, and ALFR.
45	2023	Data Augmentation via Subgroup Mixup for Improving Fairness	IEEE International Conference on Acoustics, Speech, and Signal	Vanilla Mixup – Adversarial Debiasing – Fair Mixup Penalty (FMP) – Standard Fair Data Augmentation (bootstrapping)	Tabular		Random Forest, MLP	Race (white/nonwhite)	Mixup (interpolation-based)	Education (Law School), Finance, Criminal Justice, Healthcare	Uses Demographic Parity Gap (ΔDP) as the primary fairness metric.	Scopus	2023	Proposes Fair SubGroup Mixup (FSGM), which interpolates between samples across class and group subpopulations to create synthetic data that improves Demographic Parity. Evaluated on synthetic data and the Law School dataset, compared to baselines like adversarial debiasing, vanilla mixup, and fair mixup penalties. Demonstrates consistent improvement in fairness with minor or no loss in accuracy.

SL	Year	Title	Publication Venue	Dataset	Dataset Type	Descriptions	Classifier Used	Sensitive Attribute	Taxonomy	Domain	Evaluation Metrics		Library	Year	Summary
		Processing (ICASSP) 2024	g with group label flips)												
46	2023	Learning Fair Graph Representations via Automated Data Augmentations	International Conference on Learning Representations (ICLR) 2023	NBA, Pokec-z, and Pokec-n	Graph		GNN (Graph Convolutional Networks)	Gender, Race, Region	Automated Augmentations (edge perturbation, feature masking)	Social Networks (Pokec), Sports (NBA), Finance, Criminal Justice, Healthcare	Demographic Parity Gap (ΔDP) and Equal Opportunity Gap (ΔEO).		Scopus	2023	Proposes Graphair, an automated graph augmentation framework that learns to perturb features and edges to minimize demographic and opportunity gaps in downstream predictions. Using adversarial and contrastive losses, it achieves fairness without manual rules. Evaluated on NBA and Pokec datasets, it shows superior fairness-accuracy tradeoff over existing graph fairness methods.
47	2023	Dealing with Data Bias in Classification: Can Generated Data Ensure Representation and Fairness?	International Conference on Artificial Intelligence and Soft Computing (ICAISC) 2023		Tabular	Adult, Bank, COMPAS, German Credit, and compared with Reweighting, DIR, and LFR pre-processing methods.	KNN, LR, DT	Sex, Age, Race, Foreign Worker	Gaussian Copula (statistical generative)	Finance, Criminal Justice	Statistical Parity Difference (SDP) – Average Odds Difference – Normalized Mutual Information (NMI) – Pearson's Correlation (ρ) – Custom black-box fairness metrics		Scopus	2023	Introduces a generative pre-processing method that uses statistical models (e.g., Gaussian copula) to generate synthetic samples that reduce data discrimination. The method evaluates candidate samples using black-box fairness metrics (e.g., SDP, AOD) and iteratively adds those that minimize discrimination. Benchmarked on four fairness-critical datasets, it outperforms traditional pre-processing baselines in fairness with minimal compromise in accuracy. Supports flexible fairness definitions and demonstrates robustness across classifiers and datasets.
48	2023	Distance Correlation GAN: Fair Tabular Data Generation with Generative Adversarial Networks	Springer Nature Switzerland AG 2023	UCI Adult, Bank Marketing, COMPAS, and Law School datasets	Tabular	Benchmarks against TabFairGAN and CRDI on UCI Adult, Bank Marketing, COMPAS, and Law School datasets. Reports fairness and utility trade-offs.	DTC, LR, MLP	Sex, Age, Ethnicity/Race	GAN	Finance, Education, Criminal Justice, Advertising	Uses Demographic Parity to measure fairness in both the synthetic data and the classifier trained on it.		Web of Sceince	2023	The paper proposes a GAN-based model that enforces fairness by minimizing the distance correlation between protected and target attributes. It evaluates fairness using demographic parity metrics on UCI Adult, COMPAS, Law School, and Bank Marketing datasets. The model outperforms baselines like TabFairGAN and CRDI in reducing bias while preserving utility.
49	2023	PreFair: Privately Generating Justifiably Fair Synthetic Data	AAAI 2023	Adult, Compas, Census KDD	Tabular	Benchmarked on Adult, German Credit, and synthetic causal graphs. Compared against Fair-SMOTE, DIR, Causal Path Removal (CPR), and non-private models.	MLP, Linear Regression, Random Forest	Sex, Race, Native Country	Probabilistic Graphical Models (Bayes nets)	Finance, Criminal Justice	Demographic Parity, True Positive Rate Balance (TPR Balance), True Negative Rate Balance (TNR Balance), Conditional Measures		Others	2023	The paper proposes PreFair, a causal modeling framework that synthesizes fair and private data using differentially private training and counterfactual reasoning. It ensures justifiable fairness by restricting influence to admissible causal paths. The method outperforms Fair-SMOTE and CPR on Adult and German datasets in terms of both fairness and utility, while maintaining formal privacy guarantees.
50	2023	FairGen: Fair Synthetic Data Generation	ICML2025	Adult Income and German Credit datasets	Tabular	Benchmarks conducted on Adult Income and German Credit datasets. Compared with raw data and a standard data augmentation method across CTGAN, CopulaGAN, and Gaussian Copula GAN.	Not explicitly named (standard classifiers assumed for evaluation)	Gender, Race	GAN	Finance	BCA, Demographic Parity Ratio (DPR) and Equalized Odds Ratio (EOddR)		Others	2023	The paper introduces a model-agnostic preprocessing pipeline that removes bias-inducing samples or augments training data prior to GAN-based generation. The approach improves fairness metrics (DPR, EOddR) across gender, race, and intersectional subgroups without modifying GAN architectures. Benchmarked on Adult and German Credit datasets, FairGen demonstrates fairer outcomes and improved utility in some settings.
51	2023	Information-Minimizing Generative	Neural Processing Letters, Springer	UCI Adult Income	Tabular	Benchmarks conducted on UCI Adult Income and ProPublica	LR, CNN	Gender, Race, Age	GAN	Finance	ϵ -Fairness (via BER) for generated data, and Demographic Parity and		Others	2023	Proposes FAGAN, which combines a generator, classifier, and adversary in joint adversarial training to minimize information leakage from sensitive to non-sensitive attributes. A latent factor is constructed

SL	Year	Title	Publication Venue	Dataset	Dataset Type	Descriptions	Classifier Used	Sensitive Attribute	Taxonomy	Domain	Evaluation Metrics	Library	Year	Summary	
		Adversarial Network for Fair Generation and Classification				COMPAS datasets. Compared with ACGAN, FairGAN, and FairGAN+.					Equalized Odds for classification fairness.				using ANOVA to preserve utility while ensuring fairness. Evaluated on Adult and COMPAS datasets, FACGAN outperforms FairGAN and FairGAN+ in both fairness (ϵ -Fairness, Demographic Parity, Equalized Odds) and utility.
52	2024	Benchmarking the Fairness of Image Upsampling Methods	ACM FAccT '24	CelebA and FairFace	Image	CelebA is a large-scale face attribute dataset with over 200,000 celebrity images labeled by attributes like gender and age. FairFace is a balanced face dataset of 100,000 images labeled by race, age, and gender to reduce bias in facial recognition tasks.	PULSE, pSp, DDRM	Race	GANs, Diffusion Models	Vision	Assesses fairness using subgroup representation imbalance and feature space distances (e.g., cosine similarity to face embeddings across demographic subgroups).	ACM	2024		The paper proposes a fairness benchmarking framework for generative image upsampling models. It introduces new fairness metrics (RDP, PR, UCPR) and evaluates five upsampling methods on racially biased datasets. The study shows that no current method achieves statistical fairness, despite using fairness-aware training.
53	2024	Can Synthetic Data be Fair and Private? A Comparative Study of Synthetic Data Generation and Fairness Algorithms	LAK' 25	Adult, COMPAS, and German Credit	Tabular	Benchmarked on multiple datasets including Adult, COMPAS, and German Credit, with comparisons to real-data fairness-aware training.	RF, LR, GNB, XGB	Sex, Disability, Race	GANs, LLM	Finance, Criminal justice	Statistical Parity Difference (SPD) – Disparate Impact (DI) – True Positive Rate Gap (TPG) – Demographic Parity Gap (DPG)	ACM	2024		This paper evaluates five synthetic data generators and four pre-processing fairness algorithms across fairness, privacy, and utility. It finds that fairness-aware pre-processing significantly improves synthetic data fairness and highlights DECAF as the best method when privacy and fairness must be balanced.
54	2024	FaceSaliencyAug: Mitigating Geographical, Gender and Stereotypical Biases via Saliency-Based Data Augmentation	Signal, Image and Video Processing, 2024 Springer	four occupation datasets (CEO, Nurse, etc.)	Image	Evaluated on FFHQ, WIKI, IMDB, LFW, UTK, Diverse, and four occupation datasets (CEO, Nurse, etc.), compared with RSMDA and other baselines.	CNNs, ViTs	Gender, Race, Geography	Data Augmentation (Saliency-based)	Vision,	Uses Image-Image Association Score (IIAS) for gender bias and Image Similarity Score (ISS-intra and ISS-cross) for data diversity.	Web Of Science	2024		Proposes FaceSaliencyAug, a saliency-driven augmentation strategy that generates fairer facial image data through selective region masking. Evaluated on CNNs and ViTs, it improves data diversity and reduces gender and geographic biases as measured by ISS and IIAS across real-world facial datasets.
55	2024	Synthetic Dataset Generation for Fairer Unfairness Research	LAK '24 – Learning Analytics and Knowledge Conference	UCI Adult	Tabular	Benchmarked using real-world datasets (e.g., UCI Adult) and evaluated with standard mitigation methods like Reweighting, DIR, and Equalized Odds.	Various (e.g., classifiers for AUC evaluation)	Sex, Race, Socioeconomic Status, Age	Genetic Algorithms (dataset modification)	Finance	Uses Statistical Parity, Calibration Equality, and Overall Accuracy Equality to quantify unfairness in generated datasets.	Web Of Science	2024		Proposes a genetic algorithm-based method to generate synthetic datasets that embed specific types of unfairness. Useful for benchmarking fairness mitigation methods, especially in educational data. Validated across multiple metrics and datasets.
56	2024	Data-Centric Explainable Debiasing for Improving Fairness in Pre-trained Language Models	• Findings of ACL 2024	• Stanford Sentiment Treebank (SST2) • ToxiGen • Bias in Bios	Text	SST2 is a binary sentiment dataset of 4,133 movie reviews with gender labels. ToxiGen contains 38,000 auto-generated toxic text samples labeled for implicit bias. Bias in Bios includes 250,000 biographies	BERT, DistilBERT, RoBERTa	Gender, Race	Counterfactual Data Augmentation	NLP	evaluates fairness explicitly using proposed metrics: FalseRatio/TrueRatio (FR/TR) and TruePositiveRatio (TPR), along with standard fairness indicators (Precision, Recall, F1) for evaluating biases.	Web Of Science	2024		Proposes Data-Debias, an explainability-driven framework that identifies explicit and implicit bias-inducing words and generates synthetic augmented data through word-swapping and attribution-based editing. Trains debiased models with reweighted loss. Evaluated on SST-2, ToxiGen, and Bias in Bios with superior fairness and task performance.

SL	Year	Title	Publication Venue	Dataset	Dataset Type	Descriptions	Classifier Used	Sensitive Attribute	Taxonomy	Domain	Evaluation Metrics	Library	Year	Summary
						across 28 occupations for occupation prediction without explicit occupation mentions.								
57	2024	FairFlow: An Automated Approach to Model-based Counterfactual Data Augmentation for NLP	ECML PKDD 2024, Springer	Bias-in-Bios ECHR Jigsaw	Text	Benchmarked on standard NLP datasets including Bias-in-bios, ECHR, and Jigsaw, comparing against state-of-the-art dictionary-based CDA methods (Hall-M) and generative baselines (Hall-M + BART)..	BERT, BART, ELECTRA	Gender	Counterfactual Data Augmentation (Invertible Flow-based, Transformer-based)	NLP	Uses fairness metrics such as True Positive Rate Difference (TPRD) and False Positive Rate Difference (FPRD) across demographic groups to assess fairness improvement explicitly.	Web Of Science	2024	Introduces FairFlow , a framework that uses flow-based models and BART to generate parallel counterfactual text without manual dictionaries. Demonstrates bias mitigation across three datasets with improved fairness and fluency.
58	2024	From Fake to Real: Pretraining on Balanced Synthetic Images to Prevent Spurious Correlations in Image Recognition	Computer Vision - ECCV 2024, Springer	CelebA-HQ, UTK-Face, and SpuCo Animals	Image	Benchmarked on CelebA-HQ, UTK-Face, and SpuCo Animals, comparing against USB, ASB, and non-synthetic baselines like ERM, GroupDRO, and DFR.	ResNet, ViT	Gender, Environment (Indoors/Outdoors)	Diffusion Models (Stable Diffusion)	Vision	Evaluates fairness via Worst Accuracy (WA) and Balanced Accuracy (BA) across subgroups (intersection of target class and bias group), measuring fair performance across sensitive combinations.	Web Of Science	2024	Proposes FFR , a two-stage training pipeline that first pretrains on balanced synthetic images and then fine-tunes on real data to prevent spurious correlations. Outperforms USB and ASB across three datasets and five bias severities.
59	2024	Imposing Fairness Constraints in Synthetic Data Generation	Journal of Big Data, 2022	Adult, German Credit, and Bank Marketing datasets	Tabular	Benchmarked on Adult, German Credit, and Bank Marketing datasets. Compared to standard CTGAN and a post-processing baseline.	LR, MLP, Linear Regression	Gender, Race	GAN	Socio-economic (Adult), Education (Law School), Finance, Healthcare, Advertising	Evaluates fairness using Disparate Impact (DI), Statistical Parity Difference (SPD), and Accuracy across subgroups.	Web Of Science	2024	Proposes a GAN-based framework that imposes fairness constraints (e.g., counterfactual fairness, information filtering) during synthetic data generation. Evaluates fairness-utility trade-offs on real datasets.
60	2024	Long-Term Fair Decision Making through Deep Generative Models	AAAI 2024	SimLoan	Tabular	Benchmarked on SimLoan (synthetic) and Taiwan credit dataset (semi-synthetic). Compared with MLP, MLP-DP, MLP-EO, and LRLF baselines.	Neural Networks	Race, Gender	Deep Generative Models	Finance (Bank Loans), Education (College Admissions), Employment (Job Placements), Criminal Justice (Recidivism Risks)	Uses 1-Wasserstein distance to measure long-term fairness and direct discrimination for local fairness. Demonstrates fairness improvements over traditional baselines.	Web Of Science	2024	Proposes a three-phase framework combining deep generative modeling and fairness-aware decision learning. Trains an RCGAN to generate synthetic and interventional time series, then uses these to train a classifier balancing long-term and local fairness. Evaluated using a novel 1-Wasserstein fairness metric on synthetic and semi-synthetic datasets, outperforming fairness baselines.
61	2024	Mitigation of Gender Bias in Automatic Facial Non-verbal Behaviors Generation	ACM ICMI 2024 (International Conference on Multimodal Interaction)	(Facial non-verbal time series data)	Image	Benchmarks with its own baseline model (FaceGen), and uses Trueness corpus (real-world discriminatory scenes). Evaluation includes statistical and human studies.	Neural Networks (Gender Discriminator)	Gender	GAN	HCI (Socially Interactive Agents), Healthcare, Employment (Training Simulations)	Gender classification accuracy (as a proxy for gender distinguishability) – DTW (Dynamic Time Warping) distances between male and female distributions – Subjective ratings for believability and coordination	Web Of Science	2024	Proposes FairGenderGen, a speech-driven generative model that produces facial behaviors while reducing gender bias. Uses a gradient reversal layer to make latent features gender-invariant. Evaluation shows reduced gender identifiability (drop from 80.7% to 48.6% accuracy), while preserving perceived believability and coordination in most cases.
62	2024	Realistic Morphology-Preserving Generative Modelling	Nature Machine Intelligence	UK Biobank and ADNI datasets	Image	Benchmarked against HA-GAN , CCE-GAN , and LS-GAN on UK Biobank and ADNI datasets, using image	Segmentation (SynthSeg, FastSurfer)	Age, Pathology, Ethnicity	VAEs, GANs, Diffusion Models	Healthcare (Brain MRI), Medical Imaging	n fidelity, morphological accuracy	Web Of Science	2024	This paper presents a scalable, high-fidelity generative model of 3D brain MRI that preserves both biological and disease-specific morphology. The model generates synthetic samples conditioned on demographic and pathological variables to balance underrepresented groups. Extensive evaluations confirm the morphological accuracy and clinical

SL	Year	Title	Publication Venue	Dataset	Dataset Type	Descriptions	Classifier Used	Sensitive Attribute	Taxonomy	Domain	Evaluation Metrics	Library	Year	Summary	
		of the Brain				fidelity metrics like FID, MMD, MS-SSIM, and morphological preservation.									viability of the synthetic data, including voxel-based morphometry, subcortical volume comparisons, cortical thickness assessments, and successful cross-domain learning from synthetic to real datasets.
63	2024	Toward Unified Data and Algorithm Fairness via Adversarial Data Augmentation and Adaptive Model Fine-tuning	ICDM 2022	CIFAR-10S, CelebA, Adult, COMPAS, and Bank datasets	Tabular	Experiments conducted on CIFAR-10S, CelebA, Adult, COMPAS, and Bank datasets with comparisons to state-of-the-art fairness methods.	DNNs, GNNs, LLMs (LLaMA2, LLaMA3)	Gender, Race, Zip Code	Adversarial Data Augmentation	Socio-economic (Adult), Computer Vision (CIFAR, ImageNet), Healthcare, Legal	Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), Average Odds Difference (AOD).	Web Of Science	2024		Proposes a unified intra-processing fairness framework that addresses both data-level and algorithmic bias. Introduces a novel fairness-aware adversarial data augmentation method and a fine-tuning strategy using weight reactivation. Evaluates SPD, EOD, and AOD, achieving strong fairness-accuracy tradeoffs.
64	2024	Enhancing Tabular GAN Fairness: The Impact of Intersectional Feature Selection	ICMLA 2024	UCI Adult and Diabetes Health Indicator datasets.	Tabular	Experiments conducted on UCI Adult and Diabetes Health Indicator datasets. Models compared include TabFairGAN and CTGAN (standard GAN baselines).	LR, RF	Sex, Race	GAN	Socio-economic (Adult), Healthcare (Diabetes)	Adds intersectional demographic parity constraints into the generator's loss function to generate data that mitigates bias across multiple sensitive attributes (e.g., Gender-Age, Gender-Income).	IEEE	2024		Introduces an intersectionality-based fairness constraint into GANs for generating fair tabular data. Applies modified loss functions in TabFairGAN and CTGAN to enforce demographic parity across intersecting attributes like gender-age and gender-income. Evaluates fairness using Demographic Parity, Equal Opportunity, and Equalized Odds on UCI Adult and Diabetes datasets. The models generate more fair and representative data across subgroups.
65	2024	A Novel Assurance Procedure for Fair Data Augmentation in ML	ECAI 2024	CV employee dataset	Tabular	Uses a public tabular dataset (CV employee dataset). Benchmarks with Random Forest and XGBoost under multiple data representations: original, SMOTE-balanced, protected-feature balanced, and augmented.	RF, SVM, DT	Gender, Race	Similarity Networks (Label Propagation)	Socio-economic, Criminal Justice	Uses Equal Opportunity, Equal Mis-Opportunity, True Positive Rate, and False Positive Rate. Reports improved fairness (~50%) in augmented datasets.	Scopus	2024		Proposes a fairness-aware data augmentation approach using similarity networks and Agent-Based Vector Label Propagation (AVPRA). Synthetic nodes are added near underrepresented subgroups and labeled through propagation. Evaluated on a public CV dataset using fairness metrics (Equal Opportunity, Equal Mis-Opportunity), achieving ~50% fairness improvement. Demonstrates accuracy and SHAP transparency alongside fairness enhancement.
66	2024	Addressing Both Statistical and Causal Gender Fairness in NLP Models	Findings of the Association for Computational Linguistics: North American Chapter (NAACL) 2024	Bias in Bios and WinoGender	Text	Experiments conducted on Bias in Bios and WinoGender, with comparison to models trained with standard and prior augmentation strategies.	BERT, RoBERTa	Gender, Race	Counterfactual Data Augmentation	NLP	Evaluates Statistical Parity (SP) and Natural Direct Effect (NDE) (a causal fairness measure)	Scopus	2024		Proposes a counterfactual data augmentation method that generates gender-swapped text to improve both statistical and causal fairness. Evaluated on Bias in Bios and WinoGender datasets using Statistical Parity and Natural Direct Effect (NDE), the method reduces both bias measures without sacrificing task accuracy.
67	2024	Can Generative AI-based Data Balancing Mitigate Unfairness Issues in ML?	European Workshop on Algorithmic Fairness (EWAF) 2024	German Credit – Berkeley Admission	Tabular		RF, LR	Gender	GANs, VAEs, Diffusion Models, LLMs	Finance, Education	Demographic Parity (DP) – Equalized Odds (EO) – Conditional Use Accuracy Equality (CUAE) Also reports accuracy (ACC) and AUC.	Scopus	2024		Investigates the use of GenAI models (CTGAN, GReaT) to generate synthetic tabular data for balancing protected attributes (like gender). Evaluates fairness improvement in classifiers trained on augmented datasets using metrics like DP, EO, and CUAE. Compared with SMOTE and real-only baselines on German Credit and Berkeley Admission datasets. Results show fairness benefits with GenAI methods.
68	2024	LLM-Guided Counterfactual Data Generation for Fairer AI	ACM Web Conference 2024	CelebA and UTKFace	Image	Human face images (CelebA and UTKFace) with demographic/protected attributes like gender, race, and age	DNNs	Gender, Race, Age	LLMs	Vision	Disparate Impact (DI) – Statistical Parity Difference (SPD) – Equal Opportunity Difference (EOD) – Average Odds Difference (AOD) – Balanced Classification Accuracy (BCA) – Classification Accuracy (CA)	Scopus	2024		Introduces a fairness-driven pipeline that generates counterfactual images guided by LLMs and text-to-image diffusion models. The prompts are derived from model explanation scores (LIME) and fairness metrics (AIF-360). Counterfactuals are used to fine-tune classifiers, achieving measurable bias reduction across CelebA and UTKFace datasets. LLM-generated counterfactuals outperform manual ones in bias mitigation, demonstrating the framework's effectiveness in generating fairer synthetic data.

SL	Year	Title	Publication Venue	Dataset	Dataset Type	Descriptions	Classifier Used	Sensitive Attribute	Taxonomy	Domain	Evaluation Metrics	Library	Year	Summary
69	2024	Charting a Fair Path: FaGGM – Fairness-Aware Generative Graphical Models	International Conference on Artificial Intelligence (AI 2024).	Adult UCI – Compas – Dutch Census – Law School Admission – Credit Card Clients	Tabular		XGBoost	Gender, Race	Graphical Models	Finance, Criminal Justice, Education	Mean Difference (MD) (data fairness) – Statistical Parity Difference (SPD) – Disparate Impact (DI) – Equal Opportunity (EO)	Scopus	2024	Proposes FaGGM, a fairness-aware graph structure learning algorithm that integrates fairness into score-based DAG generation. The method learns fair causal structures and generates synthetic data that reduces group-level discrimination (e.g., SPD, DI, EO). Benchmarked across five tabular datasets and four standard bias mitigation baselines, FaGGM achieves superior fairness improvements (80–97%) with minimal predictive performance loss.
70	2024	Bt-GAN: Generating Fair Synthetic Healthdata via Bias-transforming GANs	Journal of Artificial Intelligence Research (JAIR)	UCI adult, ... Compas, MIMIC-III	Tabular	Benchmarks on MIMIC-III, UCI Adult, COMPAS datasets. Compared with FairGAN, NaiveGAN, Reweighting, and pre-processing baselines.	LR, DNN	Gender, Ethnicity	GAN	Health	evaluates fairness using Demographic Parity, Equal Opportunity, Equalized Odds, and Wasserstein distance on both synthetic and downstream models.	Scopus	2024	The paper introduces Bt-GAN , a GAN-based model for generating fair synthetic health data by transforming biased input samples. It uses a novel bias transformer module and adversarial training to mitigate unfair representations in sensitive attributes. Benchmarked on MIMIC-III, Adult, and COMPAS datasets, it outperforms FairGAN and reweighting baselines on fairness metrics like Demographic Parity, Equal Opportunity, and Equalized Odds .
71	2024	Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making	Future Generation Computer Systems, Elsevier, 2024	Adult and German Credit datasets	Tabular	Benchmarked using Adult and German Credit datasets. Compared against Fair-SMOTE, DIR, and Causal Path Removal (CPR) methods.	LR, DT, RF	Gender, Race, Age	Causal Models (Bayesian Networks)	Finance, Criminal Justice	Demographic Parity, Equal Opportunity, and Counterfactual Fairness Rate (CFR)	Science Direct	2024	The paper presents a causal-based data generation framework that removes discriminatory paths in structural causal models to create counterfactually fair datasets. It demonstrates improved fairness (Demographic Parity, Equal Opportunity, CFR) on Adult and German Credit datasets compared to other preprocessing methods like Fair-SMOTE and DIR.
72	2025	A Novel Metric-Based Counterfactual Data Augmentation with SIL	IJACSA 2025	CrowS-Pairs	Text	CrowS-Pairs: sentence-level bias test dataset with demographic-sensitive sentence pairs (e.g., gender, race).	NLP Models	Gender, Race, Demographic	Counterfactual Data Augmentation (RL-based SIL)	NLP	Statistical Parity Difference (SPD) – Equal Opportunity Difference (EOD) – Conditional Demographic Disparity (CDD) – WEAT, SMART Testing, and the proposed CFRE metric	Web Of Science	2025	This paper introduces the CFRE metric for jointly evaluating fairness and robustness in NLP models. It integrates counterfactual data augmentation with Self-Imitation Learning (SIL) to guide model training using fairness-aware trajectories. Experiments on CrowS-Pairs dataset show improved fairness and robustness over existing methods like WEAT and SMART.
73	2025	Disentangled Contrastive Learning for Fair Graph Representations	Neural Networks, 2025	German, Credit, and Recidivism	Tabular	uses standard datasets like German, Credit, and Recidivism, benchmarking against FairGNN, NIFTY, EDITS, FVGNN, and FairMILE.	MLP	Age, Gender, Race	VAEs	Finance, Criminal Justice	Evaluates fairness explicitly using Statistical Parity (ΔSP) and Equalized Odds (ΔEO) metrics.	Web Of Science	2025	The paper proposes FDGNN, a novel method that achieves fairness in GNNs by disentangling sensitive and non-sensitive attributes using contrastive learning. It utilizes data augmentation and counterfactual generation for fair training, and is validated across three real-world datasets (Credit, German, Recidivism) with superior fairness and efficiency results.
74	2025	Fair ultrasound diagnosis via adversarial protected attribute aware perturbations on latent embeddings	npj Digital Medicine, 2025	USC (public) and QDUS (private) ultrasound datasets.	Image	Benchmarked on the TUSC (public) and QDUS (private) ultrasound datasets. Compared against U-Net, TransUnet, SAM, MedSAM, and fairness baselines like FEBS and InD.	MLP (Discriminator in GAN)	Sex, Age	GAN	Healthcare		Web Of Science	2025	Introduces APPLE , a fairness-aware method that generates GAN-based perturbations in latent space of ultrasound segmentation models to suppress sensitive attribute information, improving equity across sex and age without retraining base models. Validated on real datasets and foundation models like SAM.
75	2025	Balanced Mixed-Type Tabular Data	Transactions on Machine Learning Research (TMLR)	ACS income dataset... acs employee dataset	Tabular		LR, DT, RF, AB, MLP	Sex, Race, Age	Diffusion Models	Finance, Criminal Justice	Evaluates on Demographic Parity Ratio and Equalized	Scopus	2025	Proposes a diffusion-based model that generates fair synthetic

SL	Year	Title	Publication Venue	Dataset	Dataset Type	Descriptions	Classifier Used	Sensitive Attribute	Taxonomy	Domain	Evaluation Matrics	Library	Year	Summary
		<i>Synthesis with Diffusion Models</i>									Odds Ratio, showing fairness improvements over 10%.			tabular data by conditioning on sensitive attributes to ensure balance. Benchmarked on ACS datasets, it achieves up to 10.5% improvement in fairness (Demographic Parity and Equalized Odds) over state-of-the-art baselines like TableDiffusion and TVAE. Focuses on preserving both data utility and fairness across diverse demographic groups.
76	2025	<i>Constructing Fair Latent Space for Intersection of Fairness and Explainability</i>	AAAI 2025	CelebA, CelebA-HQ, and UTKFace with sensitive attributes like gender and age	Image		MLP	Age, Gender	Normalizing Flows, Diffusion Models	Vision	Uses Equalized Odds (EO), Demographic Parity (DP), and Worst Group Accuracy (WGA) across multiple datasets and sensitive attributes.	Scopus	2025	This paper proposes a fairness module that constructs a fair latent space from pretrained generative models using invertible neural networks (INN). The approach disentangles sensitive attributes (e.g., gender, age) from label-relevant features to allow counterfactual explanations and fair generation. Evaluated on CelebA, CelebAHQ, and UTKFace using EO, DP, and WGA. Shows significant fairness improvement over baselines while maintaining performance.