

Supplementary Material

for the article

*Fairness in Federated Medical Imaging:
A Systematic Review Through the Dual Fairness Lens*

Pengyang Yu^{1,*}, Zhongping Dong¹, Sahraoui Dhelim²,
Chun-Mei Feng^{1,*}, M. Tahar Kechadi¹

¹School of Computer Science, University College Dublin, Dublin, Ireland

²School of Computing, Dublin City University, Dublin, Ireland

*Corresponding authors: pengyang.yu@ucdconnect.ie; chunmei.feng@ucd.ie

This Electronic Supplementary Material accompanies the main manuscript submitted to Artificial Intelligence Review. Sections, tables, and figures in this document carry the “S” prefix (e.g., Section S1, Table S1). Cross-references to the main manuscript use the unprefix numbering (e.g., “Figure 1 of the main manuscript”).

S1 Per-Database Search Documentation

This section reports the exact search strings, scope/filter settings, and hit counts for each electronic database used in the systematic review (cf. main manuscript, Section 2.1, “Search Strategy”). The search strategy combined three concept groups—federated learning, fairness, and medical imaging—using Boolean operators, with database-specific syntax adaptations.

The original screening was performed between September 2025 and December 2025 across six databases (Web of Science, ACM Digital Library, Scopus, IEEE Xplore, PubMed/MEDLINE, and Google Scholar), with a supplementary update in February 2026 to capture publications appearing after the initial screening window. Google Scholar was used as a relevance-ranked supplementary source without structured per-database hit retention, as it does not support bulk export of search results. For reproducibility documentation, we additionally executed the equivalent search in Embase as part of the revision-stage verification; Embase provides comparable academic-database coverage with structured Boolean retrieval and bulk export functionality, and is accessible through the authors’ institutional library (University College Dublin).

S1.1 Search Strings and Hit Counts

Table S1 reports the per-database documentation. Rows 1–5 correspond to the five structured-retrieval databases of the original screening (counts confirmed by a verification re-execution during the revision stage). Row 6 (Google Scholar) records the supplementary relevance-ranked role. Row 7 (Embase) documents the reproducibility-compatible substitute executed during the revision stage.

Table S1: Per-database search strings, scope/filter settings, and hit counts.

Database	Scope / Filter	Search String	Hits
Web of Science	Topic (TS=: title, abstract, keywords)	TS= (("federated learning" OR "federated optimization" OR "distributed learning" OR "collaborative learning") AND ("fairness" OR "bias" OR "equity" OR "demographic parity" OR "equal opportunity" OR "disparate impact") AND ("medical imaging" OR "clinical imaging" OR "radiology" OR "pathology" OR "dermatology" OR "ophthalmology" OR "segmentation" OR "classification" OR "diagnosis"))	304
ACM Digital Library	Abstract / Anywhere; E-Publication 2017–2026	[[Abstract: "federated learning"] OR [Abstract: "federated optimization"] OR [Abstract: "distributed learning"] OR [Abstract: "collaborative learning"]] AND [[All: "fairness"] OR [All: "bias"] OR [All: "equity"] OR [All: "demographic parity"] OR [All: "equal opportunity"] OR [All: "disparate impact"]] AND [[All: "medical imaging"] OR [All: "clinical imaging"] OR [All: "radiology"] OR [All: "pathology"] OR [All: "dermatology"] OR [All: "ophthalmology"] OR [All: "segmentation"] OR [All: "classification"] OR [All: "diagnosis"]] AND [E-Publication Date: (01/01/2017 TO 12/31/2026)]	661
Scopus	TITLE-ABS-KEY; English language; 2017–2026 intended (PUBYEAR bounds effective 2018–2025; see Notes)	TITLE-ABS-KEY((federated learning OR federated optimization OR distributed learning OR collaborative learning)) AND TITLE-ABS-KEY(("fairness" OR "bias" OR "equity" OR demographic parity OR equal opportunity OR disparate impact)) AND TITLE-ABS-KEY((medical imaging OR clinical imaging OR "radiology" OR "pathology" OR "dermatology" OR "ophthalmology" OR "segmentation" OR "classification" OR "diagnosis")) AND (LIMIT-TO(SUBJAREA,"COMP")) AND PUBYEAR > 2017 AND PUBYEAR < 2026	666

Continued on next page

Table S1 (continued)

Database	Scope / Filter	Search String	Hits
IEEE Xplore	Abstract	((("Abstract":"federated learning" OR "Abstract":"federated optimization" OR "Abstract":"distributed learning" OR "Abstract":"collaborative learning") AND ("Abstract":"fairness" OR "Abstract":"bias" OR "Abstract":"equity" OR "Abstract":"demographic parity" OR "Abstract":"equal opportunity" OR "Abstract":"disparate impact") AND ("Abstract":"medical imaging" OR "Abstract":"clinical imaging" OR "Abstract":"radiology" OR "Abstract":"pathology" OR "Abstract":"dermatology" OR "Abstract":"ophthalmology" OR "Abstract":"segmentation" OR "Abstract":"classification" OR "Abstract":"diagnosis")))	114
PubMed / MEDLINE	Title/Abstract ([tiab])	("federated learning"[tiab] OR "federated optimization"[tiab] OR "distributed learning"[tiab] OR "collaborative learning"[tiab]) AND ("fairness"[tiab] OR "bias"[tiab] OR "equity"[tiab] OR "demographic parity"[tiab] OR "equal opportunity"[tiab] OR "disparate impact"[tiab]) AND ("medical imaging"[tiab] OR "clinical imaging"[tiab] OR "radiology"[tiab] OR "pathology"[tiab] OR "dermatology"[tiab] OR "ophthalmology"[tiab] OR "segmentation"[tiab] OR "classification"[tiab] OR "diagnosis"[tiab])	103
Subtotal: structured-retrieval databases (rows 1–5)			1,848
Google Scholar	Relevance- ranked supplementary screening; bulk export not supported	Used as a supplementary relevance-ranked source. Per-database hit count was not retained at the original screening stage due to the absence of bulk export functionality in Google Scholar.	—

Continued on next page

Table S1 (continued)

Database	Scope / Filter	Search String	Hits
Embase [†]	emtree expansion + free-text; broad scope	(‘federated learning’/exp OR ‘federated learning’ OR ‘federated optimization’ OR ‘distributed learning’ OR ‘collaborative learning’/exp OR ‘collaborative learning’) AND (‘fairness’/exp OR ‘fairness’ OR ‘bias’/exp OR ‘bias’ OR ‘equity’/exp OR ‘equity’ OR ‘demographic parity’ OR ‘equal opportunity’ OR ‘disparate impact’) AND (‘medical imaging’/exp OR ‘medical imaging’ OR ‘clinical imaging’ OR ‘radiology’/exp OR ‘radiology’ OR ‘pathology’/exp OR ‘pathology’ OR ‘dermatology’/exp OR ‘dermatology’ OR ‘ophthalmology’/exp OR ‘ophthalmology’ OR ‘segmentation’/exp OR ‘segmentation’ OR ‘classification’/exp OR ‘classification’ OR ‘diagnosis’/exp OR ‘diagnosis’)	215

[†] Embase was executed during the revision stage to provide reproducibility-compatible documentation of the supplementary search role originally served by Google Scholar. Embase is not part of the original screening cohort; the 132 included studies reported in the main manuscript are based on the original six-database screening summarized in Figure 1 of the main manuscript.

S1.2 Notes on Search Documentation

Reconciliation with Figure 1 of the main manuscript. The aggregate of the five structured-retrieval databases (rows 1–5 of Table S1) is 1,848 records at revision-stage re-execution, whereas Figure 1 of the main manuscript reports 1,847 records at the identification stage. The one-record differential could not be attributed to a specific entry at re-execution time and most plausibly reflects minor database re-indexing between the original screening (September–December 2025) and the revision-stage verification (2026); it does not affect any downstream screening count, which proceeds from the archived 1,847-record identification set.

Scopus year-filter boundary. The search was designed to cover publication years 2017–2026 in all databases. In Scopus, the archived query expressed this range as `PUBYEAR > 2017 AND PUBYEAR < 2026`, whose strict-inequality semantics effectively cover publication years 2018–2025 within that database. Records published in 2017 or 2026 remained discoverable through the other four structured-retrieval databases (which applied no such bounds), the supplementary relevance-ranked screening, and citation chaining; the inclusion window stated in Section 2.2 of the main manuscript (January 2016 – February 2026) was enforced at the screening stage rather than by the Scopus filter.

Embase as a documentation substitute, not an additional source of included studies.

The Embase query was executed during the revision stage to provide an audit trail for the supplementary search space originally covered by Google Scholar at the screening stage. The 215 Embase records were not introduced into the main synthesis; the 132 final included studies remain based on the original six-database screening reported in Figure 1.

Reproducibility confirmation. A verification re-execution of the five structured-retrieval database queries during the revision stage returned the hit counts shown in Table S1 (Web of Science: 304; ACM: 661; Scopus: 666; IEEE Xplore: 114; PubMed/MEDLINE: 103), consistent with the original screening archive within standard database-indexing variation.

S2 Complete List of Catalogued Fairness-Aware or Fairness-Adapted FL Methods

Table S2 provides the complete list of the 20 fairness-aware or fairness-adapted federated learning (FL) methods catalogued in this review, organized by intervention location (client-side, server-side, communication-based). For each method we report its primary fairness scope (CF: collaboration fairness; GF: group fairness; DF $\approx$: partial dual-fairness coverage), whether the fairness mechanism is FL-native or a general fairness technique adapted to / evaluated in an FL setting, and a concise assignment rationale. A method was counted only if it explicitly proposes, adapts, evaluates, or operationalizes a fairness mechanism in an FL context; methods that are solely domain-generalization, personalization, communication-efficiency, privacy-efficiency, or foundation-model oriented were excluded unless they carry an explicit fairness-aware FL mechanism. Table 2 of the main manuscript presents a representative subset of these methods; this table is the complete enumeration.

Table S2: Complete list of catalogued fairness-aware or fairness-adapted FL methods.

Category	Method (Ref.)	Scope	FL	Native/ Adapted	Assignment rationale
Client-side	FairMixup [1]	GF		Adapted	Local mixup-based augmentation enforcing a smoothness fairness constraint; evaluated as an FL baseline in FlexFair, not originally FL-native.
	Wu et al. [15]	GF	✓	Native	Sharpness-aware local optimization minimizing demographic performance gaps on medical imaging tasks.
	FairLoRA [7]	GF	✓	Native	Fairness-aware LoRA/adaptor training for demographic equity (FairFedMed).
	DRO [4]	GF		Adapted	Distributionally robust worst-group objective; general technique applied as a local fairness mechanism in FL.
	Zafar et al. [18]	GF		Adapted	Decision-boundary covariance fairness constraint incorporable as a local loss term; non-FL-native.
Server-side	q-FFL [8]	CF	✓	Native	Rewighted global objective (q -parameterized) equalizing client-level performance.
	AFL [9]	CF	✓	Native	Agnostic/minimax FL optimizing worst-case client performance.
	Proportional FL [5]	CF	✓	Native	Game-theoretic proportional fairness across institutions in federated histopathology.

Continued on next page

Table S2 (continued)

Category	Method (Ref.)	Scope	FL	Native/ Adapted	Assignment rationale
	Adaptive reweighting [17]	CF	✓	Native	Adaptive reweighting of institutional contributions for performance fairness in medical FL.
	Cui et al. [2]	CF	✓	Native	Gradient-based penalization of inter-client performance disparity and inconsistency.
	FairFed [3]	GF	✓	Native	Fairness-aware aggregation reweighting clients by local fairness-gap measurements.
	FedMinMax [10]	GF	✓	Native	Server-side minimax optimization over demographic groups.
	FlexFair [16]	DF $\approx$	✓	Native	Variance-regularization unifying EA (CF) with DP/EO (GF); evaluated on medical imaging.
	FAIR-FATE [12]	GF	✓	Native	Fairness-aware aggregation with momentum-based estimator.
	FedFACT [21]	GF	✓	Native	Controllable group-fairness calibration with certified guarantees (one-shot post-processing).
	LoGoFair [20]	DF $\approx$	✓	Native	Post-processing Bayes-optimal calibration of local and global fairness (not training-time joint).
	DFFL [11]	CF	✓	Native	Equality $\times$ collaboration fairness at the client level.
	FedUFO [19]	DF $\approx$	✓	Native	Unified multi-level fairness (client/sample/attribute/class) via federated DRO with convergence guarantees.
Comm.	mFairFL [13]	DF $\approx$	✓	Native	Differential multipliers resolving gradient conflict across rounds for multidimensional fairness.
	FedSAC [14]	CF	✓	Native	Dynamic submodel allocation for bounded collaboration fairness under client heterogeneity.

Scope key: CF = collaboration fairness across clients/institutions; GF = group fairness across demographic groups; DF $\approx$ = partial dual-fairness coverage (no method provides a joint CF+GF theoretical guarantee under non-IID conditions). Of the four DF $\approx$ methods, three (FlexFair, FedUFO, mFairFL) engage both dimensions through training-time joint optimization; LoGoFair does so through post-processing calibration and is therefore distinguished in the main-manuscript discussion. FL column: ✓ denotes an FL-native fairness mechanism; “Adapted” denotes a general fairness technique adapted to or evaluated within an FL setting.

S3 Per-Study Quality Assessment of the 132 Included Studies

As stated in Section 2.2 of the main manuscript, each included study was evaluated against five criteria adapted from Kitchenham et al. [6]: (Q1) clarity of research objectives, (Q2) appropriateness of the experimental design, (Q3) reproducibility of the methodology, (Q4) validity of statistical analysis, and (Q5) relevance to dual fairness in federated medical imaging. Each criterion was scored on a three-point scale (Y = fully met, P = partially met, N = not met, contributing 1.0, 0.5, and 0.0 respectively to the per-study Total). Studies were stratified into three tiers (High: $\text{Total} \geq 4.0$; Medium: $2.5 \leq \text{Total} < 4.0$; Low: $\text{Total} < 2.5$) and used to weight the narrative synthesis. As stated in the main text, no study was excluded on the basis of quality scores.

S3.1 Aggregate Statistics

Table S3 summarizes the per-criterion distribution across the 132 included studies, and Table S4 reports the resulting tier distribution. The mean Total score was 3.00 (median 3.0, range 1.5–5.0), corresponding to the center of the Medium tier.

Table S3: Per-criterion distribution of quality scores across the 132 included studies. Y = fully met; P = partially met; N = not met. Mean is computed as $(\text{count}_Y \cdot 1.0 + \text{count}_P \cdot 0.5)/132$.

Criterion	Y (%)	P (%)	N (%)	Mean
Q1. Clarity of research objectives	132 (100.0)	0 (0.0)	0 (0.0)	1.00
Q2. Appropriateness of experimental design	68 (51.5)	62 (47.0)	2 (1.5)	0.75
Q3. Reproducibility of methodology	35 (26.5)	72 (54.5)	25 (18.9)	0.54
Q4. Validity of statistical analysis	23 (17.4)	78 (59.1)	31 (23.5)	0.47
Q5. Relevance to dual fairness in federated medical imaging	6 (4.5)	51 (38.6)	75 (56.8)	0.24

Table S4: Tier distribution of the 132 included studies. Tier assignment reflects topical relevance to the review’s three-scope focus (federated learning + fairness + medical imaging) integrated with reporting quality, not absolute paper impact. Highly cited foundational works that address only one scope (e.g., FedAvg) appear in the Medium tier when reporting quality on Q1–Q4 is strong but Q5 is N.

Tier	Count	Share	Threshold
High	29	22.0%	$\text{Total} \geq 4.0$
Medium	72	54.5%	$2.5 \leq \text{Total} < 4.0$
Low	31	23.5%	$\text{Total} < 2.5$

The distribution reflects the heterogeneous composition of the included literature. Approximately 22% of studies (High tier) constitute three-scope core methods or multi-site clinical FL trials with explicit fairness analysis; these were given the most detailed treatment in the narrative synthesis. Approximately 55% (Medium tier) span two of the three scopes—most commonly FL fairness methods on non-medical data (e.g., q-FFL, AFL, FairFed) and medical FL methods without an explicit fairness objective (e.g., FedBN, HarmoFL). The remaining 24% (Low tier) comprise FL foundational references, fairness theory, and methodology citations; their inclusion ensures that the review’s conceptual scaffolding is fully documented, while the lower tier signals their indirect bearing on the dual fairness focus.

S3.2 Per-Study Assessment Table

Table S5 reports the per-criterion assessment for each of the 132 included studies, grouped by tier and ordered alphabetically by first author within each tier. The Notes column records the

strongest evidence underlying the assigned scores. Full citation keys are resolved in the main bibliography.

Table S5: Per-criterion quality assessment for the 132 included studies (Section 2.2; criteria adapted from Kitchenham, 2009). Scores: Y = fully met (1.0); P = partially met (0.5); N = not met (0.0). Tier: High (Total ≥ 4.0); Medium ($2.5 \leq \text{Total} < 4.0$); Low (Total < 2.5). Source: Supplementary data file `ESM_2_quality_data.xlsx`.

Study (First Year)	Author,	Q1	Q2	Q3	Q4	Q5	Tot.	Notes (key evidence)
High Tier (Total ≥ 4.0) — 29 studies								
Bai (2024)		Y	Y	Y	Y	N	4.0	FlexLoRA addresses heterogeneous FL LLM...
Cui (2021)		Y	Y	Y	P	P	4.0	FCFL targets fairness/performance in FL
Dayan (2021)		Y	Y	P	Y	Y	4.5	20-site COVID medical FL trial
Duan (2025)		Y	Y	Y	P	P	4.0	Fair FL post-processing over Adult, ACS and...
Gichoya (2022)		Y	Y	P	Y	P	4.0	Race-recognition medical imaging study
Jiang (2022)		Y	Y	Y	P	P	4.0	HarmoFL targets heterogeneous medical-image FL
Jiang (2025)		Y	Y	Y	Y	P	4.5	UltraFedFM is federated ultrasound FM
Kaissis (2021)		Y	Y	Y	Y	P	4.5	PriMIA demonstrates privacy-preserving FL on...
Li (2020)		Y	Y	Y	P	P	4.0	q-FFL targets fair FL resource allocation
Li (2021)		Y	Y	Y	P	P	4.0	Ditto: clear fairness/robustness aims
Li (2021)		Y	Y	Y	P	P	4.0	FedBN medical/general non-IID FL
Li (2026)		Y	Y	Y	P	Y	4.5	FairFedMed/FairLoRA targets group fairness in...
Liu (2021)		Y	Y	Y	P	P	4.0	FedDG uses 4 fundus/6 prostate sites, DG...
Lu (2022)		Y	Y	Y	Y	P	4.5	HistoFL evaluates RCC/BRCA pathology tasks with...
Lu (2022)		Y	Y	P	Y	P	4.0	Seven-institution ROP imaging FL versus...
Ogier (2022)		Y	Y	Y	P	P	4.0	FLamby healthcare FL benchmark
Ogier (2023)		Y	Y	P	Y	P	4.0	Real-world histology FL trial across hospitals...
Pati (2022)		Y	Y	P	Y	P	4.0	71-site glioblastoma FL study
Seyyed-Kalantari (2021)		Y	P	Y	Y	P	4.0	CXR underdiagnosis fairness across three datasets
Seyyed-Kalantari (2021)		Y	Y	Y	Y	P	4.5	CheXclusion audits CXR fairness
Su (2024)		Y	Y	P	Y	P	4.0	mFairFL addresses group and client fairness in FL
Tayebi (2024)		Y	Y	Y	Y	P	4.5	Privacy/fairness medical imaging study
Vaidya (2024)		Y	Y	Y	Y	P	4.5	Computational pathology bias study
Wu (2024)		Y	Y	Y	P	Y	4.5	FedISM defines imaging-quality fairness in...
Xing (2025)		Y	Y	Y	Y	Y	5.0	FlexFair: explicit federated medical fairness
Yan (2024)		Y	Y	P	P	Y	4.0	Fed-LWR explicitly targets performance fairness...
Yan (2026)		Y	Y	Y	P	P	4.0	TM-FAL has two medical datasets, nine baselines...
Yang (2024)		Y	Y	Y	Y	P	4.5	Fair medical imaging generalization across six...
Zhang (2025)		Y	Y	Y	P	P	4.0	FedFACT controls global/local fairness in FL
Medium Tier ($2.5 \leq \text{Total} < 4.0$) — 72 studies								

(continued on next page)

(continued from previous page)

Study (First Year)	Author,	Q1	Q2	Q3	Q4	Q5	Tot.	Notes (key evidence)
Adnan (2022)		Y	Y	P	P	P	3.5	Medical histopathology FL+DP case study with...
Babakniya (2023)		Y	Y	P	P	N	3.0	SLoRA targets FL PEFT for NLP
Bagdasaryan (2019)		Y	Y	P	N	N	2.5	DP-SGD disparate impact across image/text/FL...
Binns (2018)		Y	P	P	P	N	2.5	Conceptual fairness paper linking ML definitions...
Burlina (2021)		Y	P	N	Y	P	3.0	Retinal bias mitigation study using modified...
Castelnovo (2022)		Y	P	P	P	N	2.5	Fairness-metrics clarification paper
Chen (2018)		Y	P	P	Y	N	3.0	Fairness decomposition with Adult, MIMIC...
Chen (2024)		Y	Y	P	N	P	3.0	FEAL uses five multicenter medical image...
Chen (2025)		Y	Y	P	N	N	2.5	FedBiP studies heterogeneous one-shot FL
Chouldechova (2017)		Y	P	P	P	N	2.5	Fairness trade-off analysis using recidivism data
Collins (2021)		Y	Y	P	N	N	2.5	FedRep personalized FL
Corbett-Davies (2023)		Y	P	P	P	N	2.5	Fairness theory and taxonomy with proofs and case...
Cummings (2019)		Y	P	P	P	N	2.5	Theoretical privacy-fairness compatibility paper...
Dai (2024)		Y	Y	Y	P	N	3.5	Co-Boosting one-shot FL on five datasets with...
Daneshjou (2021)		Y	P	N	P	P	2.5	Dermatology AI scoping review assesses dataset...
Daneshjou (2022)		Y	P	P	Y	P	3.5	Dermatology fairness benchmark evaluates three...
Dwork (2012)		Y	P	P	P	N	2.5	Foundational individual-fairness theory
Dwork (2014)		Y	P	P	P	N	2.5	Differential privacy foundations monograph
Esipova (2023)		Y	Y	Y	P	N	3.5	DPSGD fairness mechanism evaluated on MNIST...
Ezzeldin (2023)		Y	Y	P	P	P	3.5	FairFed centralizes group fairness in FL
Fraz (2012)		Y	Y	P	N	N	2.5	Retinal vessel ensemble segmentation
Groh (2021)		Y	P	P	P	P	3.0	Fitzpatrick17k links dermatology images and...
Gu (2025)		Y	Y	P	N	P	3.0	DFFL targets data fairness under non-IID FL
Guan (2024)		Y	P	P	P	P	3.0	Survey plus ADNI experiment
Gulshan (2016)		Y	P	P	Y	N	3.0	Retinal DL validation has clear aims, two...
Hardt (2016)		Y	P	P	P	N	2.5	Foundational fairness criterion with proofs and...
Hashimoto (2018)		Y	P	P	P	N	2.5	DRO fairness method with proofs, simulation and...
Hernandezperez (2024)		Y	P	Y	P	N	3.0	BCN20000 dermoscopy dataset
Hosseini (2023)		Y	Y	P	N	Y	3.5	Prop-FFL explicitly targets fair hospital...
Hu (2022)		Y	Y	Y	P	N	3.5	LoRA objective clear
Huang (2025)		Y	Y	P	P	N	3.0	FOL one-shot personalized FL
Kairouz (2021)		Y	P	P	P	N	2.5	Comprehensive FL survey
Kaissis (2020)		Y	P	N	P	P	2.5	Perspective on secure privacy-preserving...
Kang (2023)		Y	Y	P	N	P	3.0	FedISCA one-shot medical FL
Karimireddy (2020)		Y	Y	P	P	N	3.0	SCAFFOLD clear theory

(continued on next page)

(continued from previous page)

Study (First Year)	Author,	Q1	Q2	Q3	Q4	Q5	Tot.	Notes (key evidence)
Kleinberg (2017)		Y	P	P	P	N	2.5	Theoretical fairness impossibility result with...
Kusner (2017)		Y	P	P	P	N	2.5	Counterfactual fairness theory with proofs and...
Larrazabal (2020)		Y	Y	P	N	P	3.0	Gender-bias medical imaging study
Li (2020)		Y	Y	Y	P	N	3.5	FedProx clear aims
Lin (2020)		Y	Y	P	P	N	3.0	FedDF heterogeneous model fusion FL
Liu (2025)		Y	P	P	P	N	2.5	OFL survey/practical guide with taxonomy and...
Locatello (2019)		Y	Y	P	N	N	2.5	Fairness-disentanglement study with many...
Ma (2025)		Y	Y	P	N	P	3.0	One-shot medical FL classification
McMahan (2017)		Y	Y	Y	P	N	3.5	FedAvg seminal FL method
Mohri (2019)		Y	P	P	P	P	3.0	AFL gives clear minimax FL objective
Nam (2022)		Y	Y	Y	P	N	3.5	FedPara communication-efficient FL
Nguyen (2022)		Y	P	N	P	P	2.5	Ocular imaging FL survey
Obermeyer (2019)		Y	N	P	Y	N	2.5	Racial bias study has clear empirical aim...
Papadaki (2022)		Y	P	Y	P	P	3.5	FedMinMax studies group fairness in FL
Pfohl (2021)		Y	Y	Y	P	N	3.5	Clinical risk fairness benchmark
Qi (2025)		Y	Y	P	N	P	3.0	DFFL targets equality and collaboration fairness...
Qiu (2024)		Y	Y	P	P	N	3.0	FedTPG evaluates federated CLIP prompts on nine...
Rieke (2020)		Y	P	N	P	P	2.5	Perspective on federated learning for digital...
Salazar (2023)		Y	P	Y	P	P	3.5	FAIR-FATE targets fair FL on four tabular...
Salazar (2026)		Y	P	P	P	P	3.0	Survey of group fairness in FL with structured...
Sheller (2020)		Y	Y	P	P	P	3.5	Medical-image FL for BraTS institutions
Sun (2024)		Y	Y	P	P	N	3.0	Privacy-preserving FL LoRA with GLUE/GSM8K...
Tran (2021)		Y	Y	P	Y	N	3.5	DP fair deep learning on Income/Bank/Compas with...
Tschandl (2018)		Y	P	P	P	N	2.5	HAM10000 dermoscopy dataset descriptor
Wang (2020)		Y	Y	Y	P	N	3.5	FedNova addresses objective inconsistency
Wang (2024)		Y	Y	P	P	P	3.5	FedSAC targets collaborative fairness in FL
Xu (2024)		Y	P	P	P	P	3.0	Systematic review of fairness in medical image...
Yan (2023)		Y	Y	P	N	P	3.0	Self-supervised FL for heterogeneous medical...
Zafar (2017)		Y	Y	P	N	N	2.5	Fair classifier constraints on Adult/Bank with...
Zhang (2020)		Y	P	P	Y	N	3.0	Clinical BERT bias study on MIMIC notes
Zhang (2022)		Y	Y	P	N	N	2.5	DENSE data-free one-shot FL
Zhang (2024)		Y	Y	P	N	P	3.0	FedUFO unifies client/attribute fairness in FL...
Zhang (2024)		Y	Y	Y	P	N	3.5	FedSD2C one-shot FL with three image datasets...
Zhang (2025)		Y	Y	N	P	P	3.0	LoGoFair optimizes local/global fair FL on Adult...
Zhang (2026)		Y	Y	P	N	N	2.5	TOFA has clear one-shot FL-VLM aims, multiple...
Zhou (2023)		Y	Y	P	P	P	3.5	FedFA feature augmentation for non-IID FL...

(continued on next page)

(continued from previous page)

Study (First Year)	Author,	Q1	Q2	Q3	Q4	Q5	Tot.	Notes (key evidence)
Zhu (2021)		Y	Y	Y	P	N	3.5	FedGen targets heterogeneous FL
Low Tier (Total < 2.5) — 31 studies								
Abadi (2016)		Y	P	P	N	N	2.0	DP-SGD has clear privacy objective, several...
Adamson (2019)		Y	P	N	P	N	2.0	Short perspective on cancer-diagnosis labels and...
Amato (2026)		Y	P	N	P	N	2.0	One-shot FL survey with taxonomy/challenges and...
Bandi (2019)		Y	P	P	N	N	2.0	CAMELYON17 challenge uses five centers and many...
Barocas (2023)		Y	P	N	P	N	2.0	Book-length fairness ML survey
Chen (2020)		Y	P	N	N	N	1.5	Commentary on AI and health disparities
Chen (2021)		Y	P	N	P	N	2.0	Healthcare ML ethics review with fairness...
Chuang (2021)		Y	P	P	N	N	2.0	Fair mixup targets centralized group fairness on...
Farrand (2020)		Y	P	P	N	N	2.0	Studies DP, imbalance and fairness on CelebA...
Fioretto (2022)		Y	P	N	P	N	2.0	Survey of DP-fairness interaction
Guan (2022)		Y	P	N	P	N	2.0	Medical image domain-adaptation survey with clear...
Guha (2019)		Y	P	P	N	N	2.0	One-shot FL study
Guo (2024)		Y	P	P	N	N	2.0	PromptFL uses CLIP prompt FL on seven image...
Hoover (2000)		Y	P	P	N	N	2.0	Retinal vessel localization with hand-labeled...
Jha (2020)		Y	N	P	N	N	1.5	Open Kvasir-SEG dataset with clear aim and...
Li (2023)		Y	P	N	P	N	2.0	General FL systems survey
Mehrabi (2021)		Y	P	N	P	N	2.0	Fairness survey with taxonomy, not original...
Mitchell (2021)		Y	P	N	P	N	2.0	Algorithmic fairness review
Pessach (2022)		Y	P	N	P	N	2.0	Comprehensive fairness ML survey
Price (2019)		Y	P	N	P	N	2.0	Legal/policy review on medical big-data privacy
Pujol (2020)		Y	P	P	N	N	2.0	Fair decision-making with DP census data and...
Rajkomar (2018)		Y	P	N	P	N	2.0	Conceptual health-equity fairness article with...
Rajpurkar (2022)		Y	P	N	P	N	2.0	Broad AI-in-health review
Staal (2004)		Y	P	P	N	N	2.0	Retinal vessel segmentation method
Topol (2019)		Y	P	N	P	N	2.0	Narrative medical AI review with clear scope
Tran (2021)		Y	P	P	N	N	2.0	Analyzes DP data releases under fairness lens for...
Vázquez (2017)		Y	P	P	N	N	2.0	Colonoscopy benchmark with public data, splits...
Verma (2018)		Y	P	N	P	N	2.0	Fairness-definition survey
Willemink (2020)		Y	P	N	P	N	2.0	Medical imaging data-preparation review with...
Xu (2021)		Y	P	N	P	N	2.0	Survey of FL for healthcare informatics
Zou (2023)		Y	P	N	N	P	2.0	Perspective on race prediction from medical images

S3.3 Reproducibility of the Assessment

The complete scoring rationale—including per-criterion evidence quotations for each of the 132 studies—is provided as raw data in the supplementary file `ESM_2_quality_data.xlsx`. The Excel workbook contains three sheets: (i) the per-study assessment with embedded scoring formulas and data validation, (ii) auto-computed aggregate statistics, and (iii) the scoring guide documenting the decision rules applied for each criterion. The scoring decisions for one-scope foundational references (e.g., FedAvg = N on Q5) follow a strict scope-counting interpretation: a paper receives Y on Q5 only if all three scopes (federated learning, fairness, and medical imaging) are core contributions; two scopes yield P; one or zero scopes yield N. Under this interpretation, the relatively high proportion of Q5 = N (56.8%) is the expected outcome and reflects the review’s inclusive scope: the included literature deliberately spans foundational FL methods, fairness theory, and clinical applications, in addition to the three-scope core methods that constitute the High tier.

References

- [1] Ching-Yao Chuang, Jiahao Liang, Antonio Torralba, and Stefanie Jegelka. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*, 2021.
- [2] Sen Cui, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. Addressing algorithmic disparity and performance inconsistency in federated learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 26091–26102, 2021.
- [3] Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7494–7502, 2023.
- [4] Tatsunori B Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.
- [5] S Maryam Hosseini, Milad Sikaroudi, Morteza Babaie, and HR Tizhoosh. Proportionally fair hospital collaborations in federated learning of histopathology images. *IEEE Transactions on Medical Imaging*, 42(7):1982–1995, 2023. doi: 10.1109/TMI.2023.3234450.
- [6] Barbara Kitchenham, O Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. Systematic literature reviews in software engineering—a systematic literature review. *Information and Software Technology*, 51(1):7–15, 2009. doi: 10.1016/j.infsof.2008.09.009.
- [7] Minghan Li, Congcong Wen, Yu Tian, Min Shi, Yan Luo, Hao Huang, Yi Fang, and Mengyu Wang. FairFedMed: Benchmarking group fairness in federated medical imaging with FairLoRA. *IEEE Transactions on Medical Imaging*, 45(4):1337–1351, 2026. doi: 10.1109/TMI.2025.3622522.
- [8] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020.
- [9] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019.
- [10] Afroditi Papadaki, Natalia Martinez, Martin Bertran, Guillermo Sapiro, and Miguel Rodriguez. Minimax demographic group fairness in federated learning. In *ACM Conference on Fairness, Accountability, and Transparency*, 2022. doi: 10.1145/3531146.3533081.

- [11] Kaiyue Qi, Tongjiang Yan, Pengcheng Ren, Jianye Yang, and Jialin Li. DFFL: A dual fairness framework for federated learning. *Computer Communications*, 235:108104, 2025. doi: 10.1016/j.comcom.2025.108104.
- [12] Teresa Salazar, Miguel Fernandes, Hélder Araújo, and Pedro Henriques Abreu. FAIR-FATE: Fair federated learning with momentum. In *Computational Science—ICCS 2023*, volume 14073 of *Lecture Notes in Computer Science*. Springer, 2023. doi: 10.1007/978-3-031-35995-8_37.
- [13] Cong Su, Guoxian Yu, Jun Wang, Hui Li, Qingzhong Li, and Han Yu. Multi-dimensional fair federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15083–15090, 2024. doi: 10.1609/aaai.v38i13.29430.
- [14] Zihui Wang, Zheng Wang, Lingjuan Lyu, Zhaopeng Peng, Zhicheng Yang, Chenglu Wen, Rongshan Yu, Cheng Wang, and Xiaoliang Fan. FedSAC: Dynamic submodel allocation for collaborative fairness in federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pages 3299–3310, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671748. URL <https://doi.org/10.1145/3637528.3671748>.
- [15] Nannan Wu, Zhuo Kuang, Zengqiang Yan, and Li Yu. From optimization to generalization: Fair federated learning against quality shift via inter-client sharpness matching. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 5199–5207, 2024. doi: 10.24963/ijcai.2024/575.
- [16] Huijun Xing, Rui Sun, Jinke Ren, Jun Wei, Chun-Mei Feng, Xuan Ding, Zilu Guo, Yu Wang, Yudong Hu, Wei Wei, Xiaohua Ban, Chuanlong Xie, Yu Tan, Xian Liu, Shuguang Cui, Xiaohui Duan, and Zhen Li. Achieving flexible fairness metrics in federated medical imaging. *Nature Communications*, 16(1):3342, 2025. doi: 10.1038/s41467-025-58549-0.
- [17] Yunlu Yan, Lei Zhu, Yuexiang Li, Xinxing Xu, Rick Siow Mong Goh, Yong Liu, Salman Khan, and Chun-Mei Feng. A new perspective to boost performance fairness for medical federated learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 13–23. Springer, 2024. doi: 10.1007/978-3-031-72117-5_2.
- [18] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR, 2017.
- [19] Fengda Zhang, Zitao Shuai, Kun Kuang, Fei Wu, Yueting Zhuang, and Jun Xiao. Unified fair federated learning for digital healthcare. *Patterns*, 5(1):100907, 2024. doi: 10.1016/j.patter.2023.100907.
- [20] Li Zhang, Chaochao Chen, Zhongxuan Han, Qiyong Zhong, and Xiaolin Zheng. LoGoFair: Post-processing for local and global fairness in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22470–22478, 2025. doi: 10.1609/aaai.v39i21.34404.
- [21] Li Zhang, Zhongxuan Han, Xiaohua Feng, Anpeng Wu, Kun Kuang, and Qiang Yang. FedFACT: A provable framework for controllable group-fairness calibration in federated learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.