big data and cognitive computing

MDPI

Systematic Review

A Systematic Review of Cross-Population Shifts in Medical Imaging Analysis with Deep Learning

Aminu Musa [1,2,*], Rajesh Prasad [1,3], Peter Onwualu [1] and Monica Hernandez [4,5]

[1] Department of Computer Science, African University of Science and Technology, Abuja 900107, Nigeria; rprasad@aust.edu.ng (R.P.)
[2] Department of Computer Science, Federal University Dutse, Dutse 720101, Nigeria
[3] Computer Science and Engineering Department, Ajay Kumar Garg Engineering College (AKGEC), Ghaziabad 201015, India
[4] Department of Computer Science, University of Zaragoza, 50009 Zaragoza, Spain; mhg@unizar.es
[5] Aragon Institute on Engineering Research (I3A), University of Zaragoza, 50018 Zaragoza, Spain
* Correspondence: musa.aminu@fud.edu.com

Abstract

Deep learning has achieved expert-level performance in medical imaging analysis. However, models often fail to generalize across patient populations due to cross-population domain shifts, distributional differences arising from demographic variability, variations in imaging protocols, scanner hardware, and differences in disease prevalence. This challenge limits the real-world deployment and can increase health inequities. This review systematically examines the nature, causes, and impact of cross-population domain shift in deep learning-based medical imaging analysis. We analyzed 50 peer-reviewed studies from 2020 to 2025, evaluating the proposed methodologies for handling population shifts, the datasets employed, and the metrics used to assess performance. Our findings demonstrate that performance degradation ranged from 10–25% when models were tested on unseen populations, emphasizing the substantial impact of domain shifts on model generalizability. The literature reveals that mitigation strategies broadly fall into two categories: data-centric approaches, such as augmentation and harmonization, and model-centric approaches, including domain adaptation, transfer learning, adversarial learning, multi-task learning, and continual learning. While domain adaptation and transfer learning are the most widely used, their performance gains across populations remain modest, ranging from 5–15%, and are not supported by external validation. Our synthesis reveals a significant reliance on large, publicly available datasets from limited regions, with an underrepresentation of data from low- and middle-income countries. Evaluation practices are inconsistent, with few studies employing standardized external test sets. This review provides a structured taxonomy of mitigation techniques, a refined analysis of domain shift characteristics, and an in-depth critique of methodological challenges. We highlight the urgent need for more geographically and demographically inclusive datasets, adaptable modeling techniques, and standardized evaluation protocols to enable accurate and equitable AI-driven diagnostics across diverse populations. Finally, we outline future research directions to guide the development of robust, generalizable, and fair models for medical imaging analysis.

Keywords: algorithmic fairness; chest X-ray; cross-population domain shift; domain adaptation; deep learning; health equity; medical imaging

Academic Editor: Moulay A. Akhloufi

Received: 20 January 2026
Revised: 19 February 2026
Accepted: 28 February 2026
Published: 4 March 2026

1. Introduction

Medical imaging has transformed clinical diagnostics, providing non-invasive means to detect, monitor, and manage a wide range of diseases [1,2]. Among imaging modalities including X-ray, ultrasound, computed tomography (CT), and magnetic resonance imaging (MRI), X-ray remains the most widely used and accessible modality in low- and middle-income countries (LMICs), due to its cost-effectiveness and diagnostic versatility [3,4].

The integration of artificial intelligence (AI), specifically deep learning (DL), has revolutionized the analysis and interpretation of medical imaging. Convolutional neural networks (CNNs) and related architectures now achieve performance comparable to, and sometimes exceeding, that of expert radiologists [5–7]. These advances have automated the detection and classification of diseases in radiology, oncology, dermatology, and neurology, improving diagnostic accuracy while mitigating radiologist workload [8–12]. The success of such systems depends on the availability of large, high-quality, expert-annotated datasets, which have fueled the rapid evolution of AI-driven healthcare applications [13].

However, these datasets are predominantly collected in high-income Western institutions with advanced digital infrastructure, resulting in the underrepresentation of populations from LMICs. This imbalance introduces biases that compromise model fairness and generalizability across diverse patient populations [14]. Consequently, models trained on single population/race datasets may perform poorly when deployed in settings with different patient demographics, disease prevalence, or imaging practices.

This problem is broadly captured by the concept of domain shift, which occurs when the statistical properties of data used for model training differ from those encountered during real-world deployment [15]. In medical imaging, domain shifts may arise from variations in scanners, acquisition protocols, disease characteristics, annotation standards, and population demographics, often leading to substantial drops in performance and reduced clinical reliability [16]. A particularly challenging and understudied type of domain shift is the cross-population domain shift (CPDS), which occurs when AI models trained on data from one population fail to generalize effectively to another [17]. CPDS reflects global disparities in healthcare resources, imaging technology, and population characteristics [18]. For instance, patients in resource-limited settings often present advanced disease stages or distinct radiological manifestations due to delayed diagnoses, while those in high-income regions benefit from early screening. Such differences alter data distributions and hinder cross-population model transfer [19].

Population diversity in medical datasets poses a significant challenge to developing AI systems that are robust, equitable, and clinically trustworthy [20–22]. Models performing well in specific geographic or demographic contexts frequently yield biased or inaccurate predictions elsewhere, risking exacerbation of health disparities [23]. Addressing CPDS is therefore not only a technical necessity but an ethical imperative for achieving fair and inclusive AI in healthcare [24].

Existing approaches to domain shift mitigation, such as domain adaptation [25], transfer learning [26], continual learning [27], and data augmentation [28], typically assume access to data from both the source and target domains. In reality, such access is rare, particularly across regions separated by socioeconomic or demographic divides. Moreover, most available datasets lack global population representation, further increasing disparities [29].

Although several reviews have examined general domain shift in medical imaging [15,30–35], few have specifically addressed cross-population domain shift. Existing surveys often treat population differences as a secondary issue rather than a central concern. This review fills that gap by offering a comprehensive and systematic synthesis of the literature focused on CPDS in deep learning–based medical imaging analysis.

https://doi.org/10.3390/bdcc10030076

Figure 1 provides the review structure and thematic organization. The review began with an overview of deep learning applications in medical imaging and then progressed to domain shift challenges, a focused discussion on cross-population domain shift (CPDS). It outlines dataset modalities, the nature and evaluation of CPDS, and key mitigation techniques, concluding with a discussion and future research directions. Specifically, this review makes the following contributions:

- It provides the first structured synthesis dedicated to cross-population domain shift (CPDS) in medical imaging.
- It introduces a unified methodological taxonomy of CPDS mitigation strategies, encompassing data-centric and model-centric approaches.
- It critically evaluates the relative effectiveness, assumptions, and limitations of existing techniques in addressing population-level variability.
- It identifies open challenges and outlines future research directions toward building robust, generalizable, and equitable AI systems for healthcare.

The rest of the paper is organized as follows: Section 2 reviews deep learning applications in medical imaging; Section 3 details the review methodology; Section 4 synthesizes findings according to current approaches for managing cross-population shift in medical imaging analysis; and Section 5 concludes with key recommendations and future research directions.

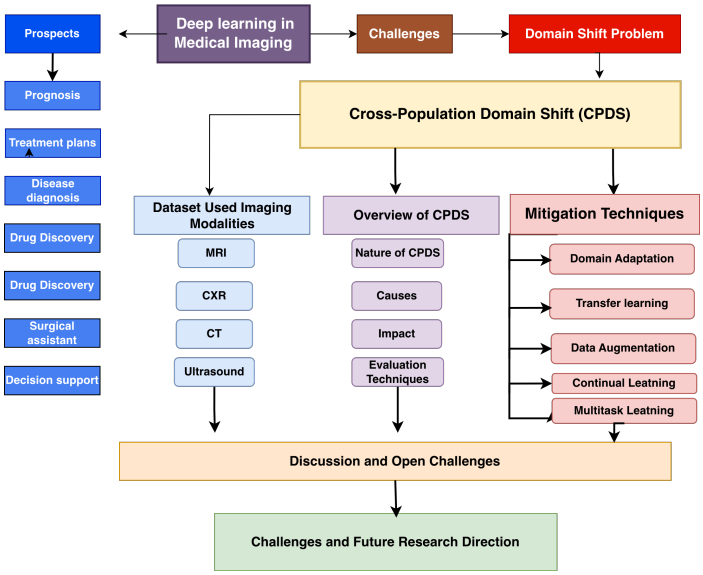


Figure 1. General organization of the review.

2. Overview of Medical Imaging Analysis with Deep Learning

Deep learning (DL) has become a cornerstone of modern data science, driving automation, prediction, and decision support across domains such as healthcare, finance, transportation, and robotics [36]. Its success in medical imaging analysis can be attributed to rapid advances in neural network architectures, increased computational resources, improved optimization algorithms, and the availability of large, annotated datasets [37].

In medical imaging, deep learning models have achieved expert-level accuracy in numerous diagnostic and prognostic tasks. Convolutional neural networks (CNNs) [38] and, more recently, transformer-based architectures [39] can automatically detect, localize, and classify complex imaging patterns from modalities such as X-ray, CT, MRI, and ultrasound [40]. By learning hierarchical representation directly from raw pixels, they outperform classical machine-learning techniques that rely on manually engineered features. Clinically, DL has enabled automated disease detection [41,42], lesion segmentation [43,44],

and outcome prediction [45], supporting radiologists in improving diagnostic accuracy and workflow efficiency.

Despite these achievements, DL systems frequently underperform when applied to data from different institutions or populations, highlighting the persistent challenge of domain shift, a key barrier to reliable clinical translation.

2.1. Clinical Scope of Deep Learning in Imaging

The rapid progress of DL has fueled diverse clinical applications spanning radiology, pathology, ophthalmology, dermatology, cardiology, and neurology. Table 1 summarizes representative studies grouped by specialty, illustrating both the breadth of DL adoption and the methodological diversity of current research.

Table 1. Selected Literature on Deep Learning Applications in Medical Imaging, Grouped by Clinical Application.

Citations	Task	Method	Dataset Used
Radiology			
[5–12,46–56]	Pneumonia, Tuberculosis, COVID-19, Pulmonary and Thoracic disease detection from chest X-ray	CNNs	ChestX-ray14, COVID-19 CXR, PadChest, CheXpert, MIMIC-CXR
Pathology			
[57–60]	Breast cancer classification from histology	CNN	BreakHi/ICD/USCD
[61–65]	Lung cancer subtype classification from WSIs	Inception-v3	TCGA-LUAD/LUSC/NLST
[66,67]	Lymph node metastasis detection	Deep CNN	CAMELYON16
[68,69]	Prediction of breast tumor proliferation	Detectron2	TUPAC16/JPATHOL
Neurology			
[43,70,71]	Brain tumor segmentation from MRI	nnU-Net	BraTS
[44,72–76]	Glioma segmentation from MRI	Deep 3D CNN	BraTS/BRISC
[45,77–79]	Alzheimer's disease prediction	3D CNN, U-Net	ADNI/OASIS
Dermatology			
[80,81]	Skin lesion classification	CNN, Inception-v3	ISIC Archive
[82–85]	Melanoma and lesion detection	Ensembled CNN	HAM10000
Ophthalmology			
[86]	Diabetic retinopathy detection	Inception-v4 CNN	EyePACS
[87]	Diabetic retinopathy detection	CNN	Retinal images
[88]	Glaucoma screening from fundus images	CNN	Fundus images
[77,89]	Detection of retinal diseases	CNN (multi-task)	Singapore National Eye Centre

Table 1 indicates that DL has become integral to imaging analysis across multiple organs and modalities. Within radiology, CNN-based models dominate chest X-ray and CT analysis for pneumonia, tuberculosis, and COVID-19. At the same time, U-Net-derived architectures excel in pathology and neuroimaging for segmentation and subtype classification. Transformer networks are emerging for multimodal feature fusion and report generation. Collectively, these studies demonstrate DL's diagnostic potential but also ex-

pose inconsistencies in dataset scale, labeling quality, and evaluation protocols, factors that complicate model transferability between populations.

2.2. Challenges in Model Generalization

Three obstacles constrain the generalization of DL models in medical imaging. First, high-resolution images from gigapixel histopathology slides to 3-D CT volumes demand substantial computational and storage resources. Second, the scarcity of high-quality annotated data, particularly in LMICs, limits model robustness and reproducibility. Third, unresolved issues around interpretability, bias, and ethical deployment impede clinical trust and regulatory approval. These challenges converge in the phenomenon of domain shift, which occurs when training and deployment data differ in statistical distribution [15].

2.3. Domain Shift in Deep Learning-Based Medical Imaging Analysis

The major challenges in deep learning-based medical imaging analysis are the scarcity and heterogeneity of data. Medical images, especially those in histopathology and advanced radiology, exhibit significant variations in terms of chemical staining, imaging modalities, and digitization protocols [90]. These variations can affect the consistency and quality of the data, making it increasingly difficult for a single deep learning model to generalize across diverse datasets, resulting in a problem termed as domain shift [91]. Domain shift (DS) refers to the deterioration in model performance that occurs when the data distribution in the Source domain (training data) diverges from that of the target domain (testing domain) [14]. DS may occur when imaging data exhibit distinct characteristics resulting from equipment settings, image quality, patient anatomy, disease prevalence, or clinical practices.

Let $\mathcal{X}$ denote the input space (e.g., chest X-ray images) and $\mathcal{Y}$ denote the output space (e.g., disease labels). Domain is defined as a tuple $\mathcal{D} = (\mathcal{X}, P(X))$, where $P(X)$ is the marginal distribution over the input space. A task is defined as $\mathcal{T} = (\mathcal{Y}, f(\cdot))$, where $f : \mathcal{X} \to \mathcal{Y}$ is the predictive function.

Given:
source domain

$$\mathcal{D}_S = (\mathcal{X}_S, P_S(X)) \tag{1}$$

target domain

$$\mathcal{D}_T = (\mathcal{X}_T, P_T(X)) \tag{2}$$

Domain shift occurs when:

$$P_S(X) \neq P_T(X) \quad \text{or} \quad P_S(Y|X) \neq P_T(Y|X) \tag{3}$$

This discrepancy results in degraded performance when a model trained on $\mathcal{D}_S$ is evaluated on $\mathcal{D}_T$.

2.4. Cross-Population Domain Shift (CPDS)

Cross-population domain shift refers to the differences in data distributions between the training $\mathcal{P}_S$ and testing $\mathcal{P}_T$ populations. In medical imaging, CPDS is a critical subset of domain shift that reflects discrepancies between training and testing populations $\mathcal{P}_S$ and $\mathcal{P}_T$. These differences can be in both the input data characteristics and the underlying label distributions, arising from demographic, epidemiological, or socioeconomic factors [89,90]. Such differences commonly include variations in age distribution, sex, race, disease prevalence, and geographic context, all of which limit model generalization across

populations. Unlike general domain shift scenarios, CPDS poses unique ethical and clinical risks, potentially reinforcing health disparities.

As illustrated in Figure 2, CPDS arises at the intersection of covariate shift and concept shift, embedded within the broader domain shift landscape, reflecting simultaneous changes in data distributions and label semantics across populations [18]. While population differences often induce covariate shift through anatomical, physiological, or imaging-distribution changes, cross-population analysis frequently also involves concept shift, where the input–label relationship itself varies across settings. Such a concept shift may stem from heterogeneous diagnostic criteria, differing disease definitions, variable clinical thresholds, or inconsistent annotation practices across hospitals, countries, or healthcare systems [92]. These factors introduce label noise and annotation ambiguity that cannot be addressed through feature alignment or distributional harmonization alone and are particularly pronounced in global health and low-resource contexts.

Figure 2. Relationship between domain shift and CPDS.

Mathematically, CPDS arises specifically from demographic or population-based differences. Let $\mathcal{P}$ represent population-specific attributes such as age, gender, ethnicity, or geography. The domain can be extended to $\mathcal{D}^{\mathcal{P}} = (\mathcal{X}^{\mathcal{P}}, P^{\mathcal{P}}(X))$, where the distributions depend on $\mathcal{P}$.

CPDS occurs when:

$$P^{\mathcal{P}_S}(X) \neq P^{\mathcal{P}_T}(X) \quad \text{or} \quad P^{\mathcal{P}_S}(Y|X) \neq P^{\mathcal{P}_T}(Y|X) \tag{4}$$

where $\mathcal{P}_S$ and $\mathcal{P}_T$ denote different populations (e.g., source: Western, target: African). This suggests that disease manifestations and imaging appearances may differ across populations, thereby affecting model generalization.

Empirical evidence confirms that demographic and infrastructural factors, such as imaging utilization rates, scanner type, and access to care, directly influence AI model predictions [92,93]. Addressing CPDS is therefore essential to ensure fairness, safety, and clinical efficacy across global populations.

2.5. Limitation of Existing Reviews

To conduct a new systematic literature review (SLR), it is essential to establish a clear need for the review [94]. A targeted review of prior surveys reveals that most existing literature examines domain shift broadly rather than focusing on cross-population effects. Matta et al. [31] reviewed domain generalization in medical image classification and proposed a taxonomy based on 77 papers. In contrast, Guan et al. [30] and Kumari et al. [34] focused on domain-adaptation techniques, supervised, semi-supervised, and unsupervised, without addressing population heterogeneity. Hong et al. [15] discussed out-of-distribution (OOD) detection, identifying demographic variation as one OOD source but not its clinical consequences. Overall, no systematic review to date has centered specifically on CPDS.

This gap motivates the present work, which systematically analyzes how population diversity influences the generalizability of DL models in medical imaging and synthesizes methodological advances aimed at mitigating cross-population domain shift.

3. Systematic Review Method

This section outlines the methodology adopted for the systematic literature review (SLR), following the procedures of Kitchenham [94] and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines [95]. The checklist can be found in the Supplementary Materials. The review process, summarized in Figure 3, shows the PRISMA flowchart used for literature search and extraction. The SLR structure comprises three key phases: planning, conducting, and reporting, each designed to ensure transparency, rigor, and reproducibility.

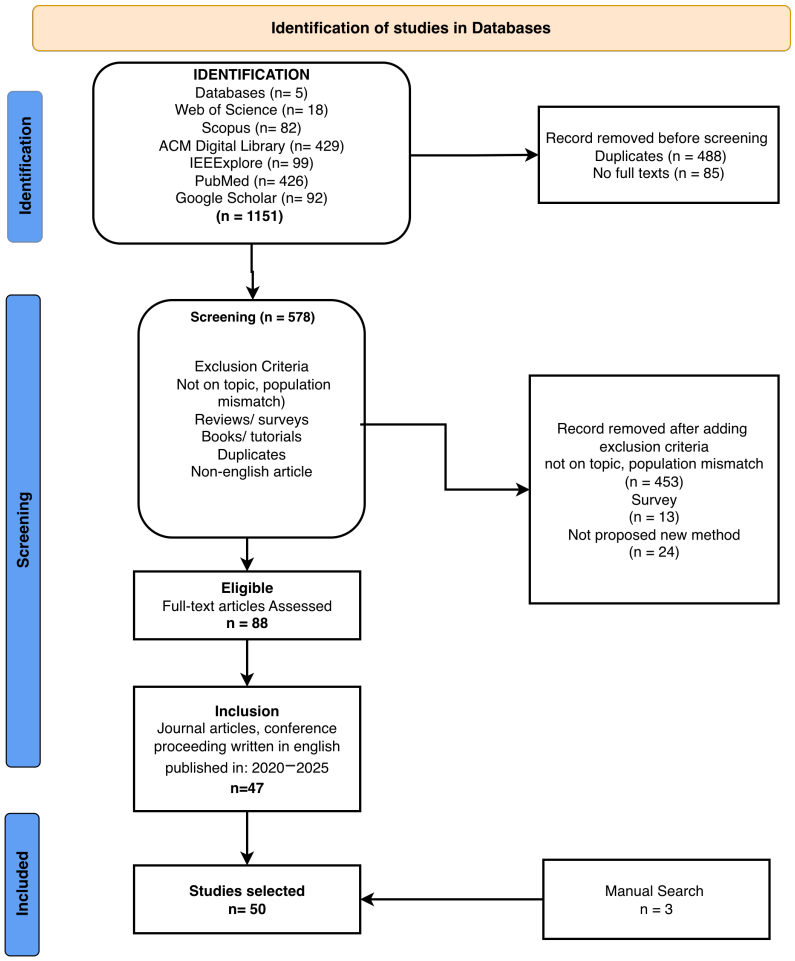


Figure 3. PRISMA Flow Diagram.

3.1. Planning Review

The planning phase established the foundation for a systematic and replicable synthesis. A review protocol was defined to clarify the objectives, search strategy, inclusion/exclusion criteria, and quality-assessment procedures.

3.1.1. Research Questions (RQ)

Four research questions (RQs) guided the review:

- RQ 1: How is cross-population domain shift defined and characterized in deep learning-based medical image analysis literature?
- RQ 2: What deep learning techniques have been proposed to mitigate CPDS?
- RQ 3: What datasets and evaluation metrics are commonly used to assess the effectiveness of these techniques?
- RQ 4: What are the limitations of current approaches, and what future research directions are identified in the literature?

3.1.2. Search Strategy

A systematic search was conducted across five major electronic databases, Scopus, PubMed, IEEE Xplore, Web of Science, and the ACM Digital Library, covering publications from 2020 to 2025. The search string combined keywords as follows: ("medical imaging" OR "medical image analysis" OR radiology OR "diagnostic imaging") AND ("domain shift" OR "dataset shift" OR "distribution shift" OR "domain generalization" OR "domain adaptation" OR "out-of-distribution" OR "external validation" OR "multi-center" OR "multicenter" OR "cross-site" OR "cross-population" OR "cross-institution" OR generalizability OR robustness) AND (population OR demographic* OR race OR ethnicity OR age OR sex OR gender OR geographic* OR regional OR global OR "low-resource" OR LMIC) AND PUBYEAR > 2019.

Duplicates were removed, and additional studies identified through citation tracking and co-author recommendations were manually included. After screening, 50 peer-reviewed studies were selected for detailed synthesis (see Figure 3).

3.1.3. Selection Criteria

Based on the research questions and the objectives of the SLR, the following inclusion and exclusion criteria were outlined and applied to the retrieved literature to screen the selected papers.

Inclusion Criteria:

- Studies employing deep learning methods for medical image analysis across any imaging modality (e.g., X-ray, CT, MRI, ultrasound, histopathology) and explicitly addressing cross-population domain shift.
- Studies explicitly addressing cross-population domain shift.
- Articles published in English.
- Papers published between January 2020–December 2025.
- Studies with clearly defined methodologies and evaluation metrics.

Exclusion Criteria:

- Studies not related to deep learning or medical image analysis.
- Books, Posters, and non-English articles.
- Review and Survey papers.
- Non-peer-reviewed publications.
- Studies with unsound scientific content or vague description of methodology.

Restricting to English-language papers ensured consistent interpretation and accurate synthesis of methodological details, though it may have excluded valuable regional studies from LMICs. This limitation is acknowledged as a potential source of geographic bias. Despite this limitation, the language criterion was necessary to maintain analytical consistency and methodological rigor within the scope of this review.

3.1.4. Quality Assesment

Quality assurance (QA) was performed to evaluate methodological rigor, transparency, and reproducibility, following PRISMA principles adapted for methodological reviews. Given the heterogeneity of DL research designs, existing clinical tools (e.g., AMSTAR-2, ROBIS) were not applicable. Instead, a ten-item checklist (Table 2) was used to assess clarity of objectives, experimental transparency, dataset documentation, and model validation.

Table 2. Quality Assessment (QA) Checklist for Included Studies on Cross-Population Domain Shift.

S/N	Quality Assessment Checklist	Rating (Y/N/P)
1	re the research aims clearly defined?	Y /N/ P
2	Is the study design appropriate to the stated objectives?	Y /N/ P
3	Does the study propose a new methodology?	Y /N/ P
4	Is the dataset adequately described?	Y /N/ P
5	Are the data analysis procedures clearly explained?	Y /N/ P
6	Are all research questions addressed?	Y /N/ P
7	Are results and conclusions logically connected?	Y /N/ P
8	Is the model architecture or experimental pipeline described?	Y /N/ P
9	Was the model validated appropriately (e.g., external or cross-domain testing)?	Y /N/ P
10	Are performance metrics reported transparently?	Y /N/ P

Each criterion was rated as Yes = 1, Partial = 0.5, or No = 0. Studies scoring above 80% were retained for synthesis. Out of 88 eligible papers, 50 met the threshold. Table 3 summarizes the QA outcomes.

Table 3. Quality Assessment (QA) Summary of Included Studies on Cross-Population Domain Shift (n = 50).

S/N	Quality Assessment Question	Y (Yes)	P (Partial)	N (No)
1	Clear research objectives	45	4	1
2	Appropriate study design	41	6	3
3	Clearly defined variables or features	33	10	7
4	Dataset /data collection process transparency	32	11	7
5	Purpose of analysis articulated	37	9	4
6	Research questions adequately addressed	31	10	9
7	Coherence between data and conclusions	29	14	7
8	External or cross-domain validation performed	19	4	29
9	Results and metrics clearly reported	42	5	3
10	Were the results and performance metrics reported clearly and transparently?	38	8	4

Overall, the QA indicated strong methodological clarity (90% defined aims; 84% described architectures) but limited transparency in dataset reporting (64%) and validation (42%). These findings reflect persistent inconsistencies in evaluating generalizability across populations, a central issue in CPDS research. The QA process thus served as an essential quality control step, allowing the review to select studies that demonstrated both methodological rigor and sufficient reporting quality for meaningful synthesis.

3.2. Conducting Review

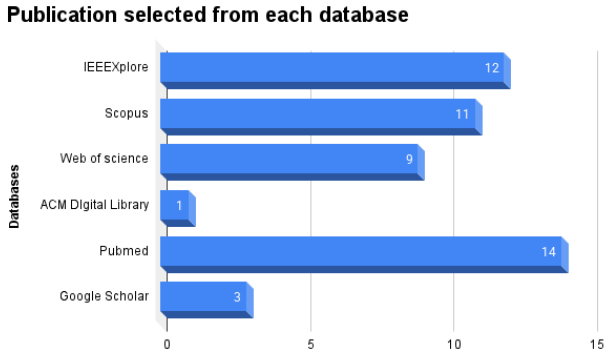
Full texts of all eligible studies were analyzed using a standardized data-extraction form capturing bibliographic details, imaging modality, task, methodology, dataset, type of domain shift, and code availability. Two independent reviewers performed the extraction

to ensure reliability. A data extraction form with specifically 10 items about the research topics was created. Among its contents were the following: (1) article title, (2) year of publication, (3) modality, (4) methodology, (5) task, (6) dataset, (7) domain, (8) shift type, (9) code availability, and (10) dataset availability.

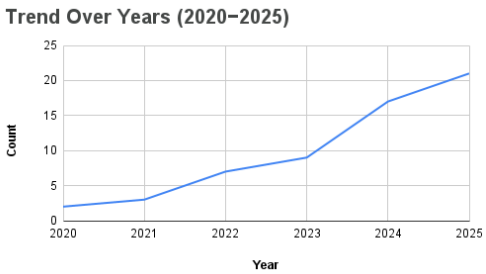
The database search initially retrieved 1146 records. After duplicate removal and preliminary screening, 573 non-relevant papers were excluded. A further 23 studies were removed following full-text review due to lack of relevance or methodological soundness, leaving 50 papers for final synthesis (see Figure 3).

Exploratory Data Analysis (EDA)

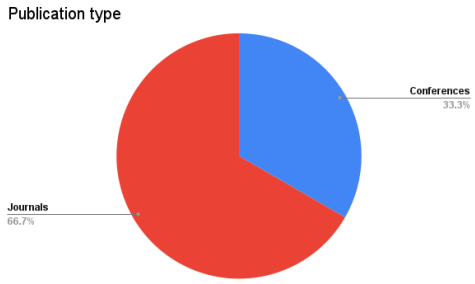
An exploratory data analysis (EDA) of the included studies was conducted to visualize publication trends and thematic distributions (Figure 4b). The analysis examined publication year, source database, and outlet type (journal vs. conference).



(**a**) Distribution of reviewed studies by academic database. Most were retrieved from PubMed and IEEE Xplore, with fewer from Scopus, Web of Science, and ACM Digital Library.



(**b**) Publication trend (2020–2025).



(**c**) Distribution of publication types.

Figure 4. Exploratory overview of the reviewed studies, showing (**a**) database distribution, (**b**) publication trends, and (**c**) publication types.

The EDA revealed that *PubMed* contributed the largest share of studies (14 papers) (See Figure 4a), followed by *IEEE Xplore* (12) and *Scopus* (11). Publication activity increased steadily from 2020 to 2025, peaking in 2024 (see Figure 4b), indicating growing research interest in domain shift and model generalization. Approximately 66% of the papers were published in journals, underscoring a preference for comprehensive peer-reviewed studies over conference proceedings (see Figure 4c).

4. Findings and Discussion

This section presents the key findings of the systematic review, organized according to the research questions (RQ1–RQ4). It summarizes how cross-population domain shift (CPDS) has been defined, categorized, and addressed in deep learning (DL)-based medical image analysis, as well as the datasets and metrics employed in the reviewed studies.

4.1. Characterization of Domain Shift in Medical Image Analysis (RQ1)

RQ1: How is cross-population domain shift defined and characterized in deep learning-based medical image analysis in the literature?

Following Quinonero-Candela et al. [16], domain shift in deep learning models can be broadly categorized into two principal types: *covariate shift*, where the input data distribution changes across domains, and *concept shift*, where the relationship between inputs and outputs differs.

4.1.1. Covariate Shift

Covariate shift occurs when $P_S(X) \neq P_T(X)$ while $P_S(Y|X) = P_T(Y|X)$, meaning the input distributions differ but the label relationship remains stable [96]. In medical imaging, this is common due to variations in scanner types, acquisition parameters, or patient demographics. For example, CXR images from high-resolution hospital scanners differ from those obtained with portable units used in low-resource settings. CPDS, arising from demographic factors such as age, race, sex, or geography, is considered a subset of covariate shift.

4.1.2. Concept Shift

Concept shift occurs when $P_S(Y|X) \neq P_T(Y|X)$, i.e., when the input–label relationship itself changes, often due to inconsistent diagnostic criteria or annotation practices [97]. For instance, two hospitals may apply different thresholds for diagnosing pneumonia, leading to label inconsistencies despite similar imaging appearances.

Both covariate and concept shifts can significantly undermine the generalization of deep learning models, as shown in Figure 5. Models trained in one domain may not adapt to different data distributions or labeling standards in other domains [14,33,98]. Beyond mere failure on new data, models can learn spurious correlations, ike associating chest drains with pneumothorax or relying on hospital-specific markers that hold in one setting but break down in others. This highlights another cause of CPDS, explaining it as more than just a distribution shift [99].

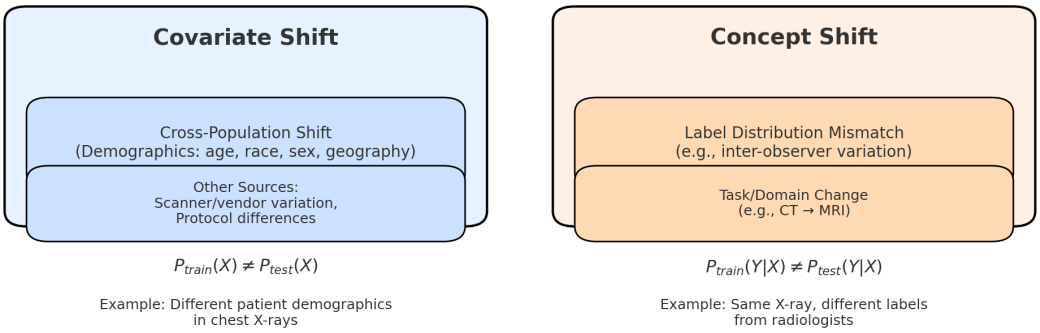


Figure 5. Difference between Covariate and Concept Shift.

4.1.3. Cross-Population Domain Shift

Cross-population domain shift refers to performance degradation that occurs when models trained on data from one population are applied to data from another population with different demographic, biological, or clinical characteristics [100]. Importantly, CPDS

cannot be attributed exclusively to a single category of domain shift. Instead, it often emerges at the intersection of covariate and concept shift. Formally, CPDS may arise when:

$$P^{\mathcal{P}_S}(X) \neq P^{\mathcal{P}_T}(X) \quad \text{or} \quad P^{\mathcal{P}_S}(Y|X) \neq P^{\mathcal{P}_T}(Y|X), \tag{5}$$

where $\mathcal{P}_S$ and $\mathcal{P}_T$ denote source and target populations, respectively.

While CPDS is frequently dominated by covariate shift due to demographic-driven differences in imaging characteristics, concept shift can also play a critical role. Population-specific disease manifestations, clinical workflows, and labeling practices can alter the semantic interpretation of visual patterns, thereby changing the conditional label distribution [101]. Therefore, CPDS should be understood as a compound form of domain shift that encompasses both statistical distributional changes and semantic inconsistencies.

By explicitly situating CPDS at the overlap between covariate and concept shift, this framework clarifies why mitigation strategies focused solely on statistical alignment are often insufficient. It highlights the need for population-aware modeling, robust evaluation protocols, and bias-aware learning strategies that account for both biological variability and population-dependent label semantics [35].

These categories of domain shifts, therefore, provide a structured framework for understanding how domain shift arises and how different shifts fall under the main broad categories. Table 4 presents the categorization of studies on domain shift.

Table 4. Characterization of Domain Shift in Medical Imaging Literature.

Domain Shift Type	Specific Shift Subtype	References	Modality
Covariate Shift	**Population Shift** (age, race, sex, geography)	[14,18,27,33,89,90,92,96,99,102–112]	X-ray, WSI, MRI, CT, Ultrasound
	Equipment variation (scanner, vendor, resolution)	[41,113–122]	Chest X-ray, MRI, CT, Fundus, Dermoscopy
	Acquisition protocol/Lighting/ Image quality	[120]	Chest X-ray
Concept Shift	Label distribution mismatch/ inter-observer variation	[22,29,123–132]	X-ray, CT
	Task/domain change (e.g., CT → MRI)	[98,133–137]	Retinal, skin lesions, CT, MRI, Chest X-ray

Our analysis identified approximately 20 studies explicitly addressing CPDS, primarily using X-ray, MRI, and CT modalities (Table 4). CPDS is most frequently framed as population shift within covariate shift, reflecting demographic variability and institutional heterogeneity [18,90]. Studies typically assess performance degradation across multi-institutional or cross-regional datasets (e.g., U.S., European, and Asian cohorts), revealing the compounding effects of demographic and equipment differences. The literature consistently links CPDS to model underperformance on underrepresented populations, underscoring its importance for equitable AI in healthcare.

In addition to population-level factors, the literature also highlights related technical sources of covariate shift, such as equipment variation, acquisition protocols, and image quality, that indirectly influence cross-population generalization [116,118,120]. This diversity of imaging modality used in studies of CPDS reflects both the ubiquity of these modalities in clinical workflows and their susceptibility to demographic and institutional variability.

4.2. Techniques for Mitigating CPDS (RQ2)

RQ2: What deep learning techniques have been proposed to mitigate CPDS?

From the reviewed studies, a taxonomy of mitigation techniques was proposed based on the methodology used and the underlying strategy adopted to address CPDS as pre-

sented in Figure 6. The mitigation strategies were clustered into three broad causes of CPDS: Prevalence/label shift, Protocol variations, and Anatomical variations.

Figure 6 presents a taxonomy of Cross-Population Domain Shift (CPDS) mitigation techniques organized according to the underlying source of population-induced shift, reflecting the clinical and demographic realities encountered when deploying medical imaging models across diverse populations. We distinguish three primary categories: prevalence and label distribution shift, protocol variation, and anatomical and biological variability.

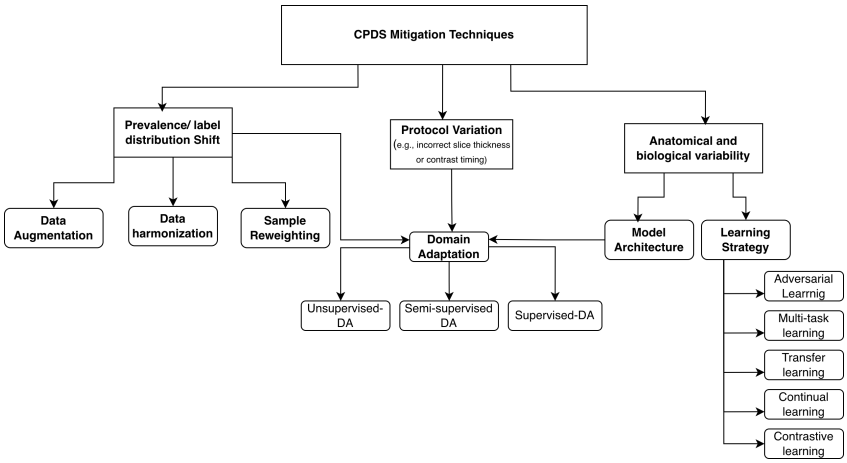


Figure 6. Taxonomy of CPDS mitigation techniques from reviewed literature: Techniques are grouped by the primary source of population-induced shift they target: (i) Prevalence/label shift methods correct class imbalance or diagnostic threshold differences; (ii) Protocol variation methods address scanner or acquisition heterogeneity via representation alignment; (iii) Anatomical/biological variability methods learn representations robust to population-specific physiology or disease presentation.

The first category, prevalence and label distribution shift, captures scenarios where disease frequencies, class imbalance, or diagnostic thresholds differ across populations. Mitigation strategies in this group, such as data augmentation, data harmonization, and sample reweighting, aim to correct distributional imbalances rather than alter model structure. The second category, protocol variation, addresses differences arising from imaging acquisition practices, such as slice thickness or contrast timing, which frequently co-vary with geographic or institutional context. Here, domain adaptation approaches spanning unsupervised, semi-supervised, and supervised settings are commonly employed to align representations across acquisition environments.

The third category, anatomical and biological variability, reflects population-specific differences in physiology, morphology, and disease manifestation that cannot be adequately addressed through statistical alignment alone. Methods in this group emphasize model architecture and learning strategies, including adversarial learning, multi-task learning, transfer learning, continual learning, and contrastive learning, to learn representations that are robust to biologically meaningful variation rather than suppressing it.

Adversarial and representation-learning methods are categorized under anatomical/biological variability because their objective is not merely to align marginal feature distributions, but to learn features that remain predictive despite population-specific anatomical or pathological differences, a requirement distinct from harmonizing scanner-induced artifacts.

By structuring CPDS mitigation strategies around the nature of population shift rather than algorithmic convenience, this taxonomy highlights why methods effective for scanner or protocol harmonization may fail under true demographic shift. It provides a clinically grounded framework for selecting mitigation strategies that align with the specific challenges posed by cross-population generalization in medical imaging.

4.2.1. Data-Centric Approaches

Data-centric approaches aim to reduce domain divergence by enriching or normalizing the data distribution by modifying or augmenting the available data to better reflect the variability of the target domain. They do not change the model architecture but attempt to reduce the distribution gap through data manipulation.

Data Augmentation:

Is the process of creating a synthetic version of the data through randomized transformations (flips, rotations, noise) used to increase dataset diversity and promote feature invariance [138]. GAN-based augmentation and Fourier-domain perturbations further mimic inter-site variability, improving generalization across hospitals and scanners. This can be achieved by defining a standard augmentation T that modifies an input x but preserves its original label y, as shown in Equation (6)

$$(x', y') = (T(x), y) \tag{6}$$

Equation (7) demonstrates how to incorporate data augmentation into the training objective by expressing the loss as an expectation over both data and augmentation distributions.

$$\mathcal{L}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{E}_{T\sim\mathcal{T}}[\ell(f_\theta(T(x)), y)]] \tag{7}$$

Stacke [33] addressed domain shift in histopathology images by applying a color and intensity data augmentation pipeline to the Camelyon-17 dataset, aiming to reduce color variability and enhance model generalization across slides from different medical centers. Similarly, Garrucho [118] conducted a multi-center study on deep learning models for mammography, introducing a single-source training approach where models are trained on data from one domain and tested on previously unseen centers. Their results revealed that all detectors experienced performance declines when applied to new domains. However, the model with data augmentation consistently outperformed the cross-domain approaches, demonstrating improved generalization without relying on target-domain images.

While augmentation can improve robustness, it can also introduce or amplify biases if not applied carefully. If the original dataset lacks fundamental diversity (e.g., is missing entire demographic groups), augmentation cannot create meaningful new information and may simply reinforce existing imbalances [139].

Data Harmonization:

Data homogenization and intensity normalization align image characteristics across populations or equipment types [42,140]. These approaches enhance consistency in appearance and reduce scanner- or region-induced bias, though they require access to multi-population data.

Overall, studies report moderate but consistent gains in AUC and Dice scores when employing augmentation or harmonization. For example, BigAug [102] and ASPECTS [116] improved cross-domain segmentation accuracy by 3–8%, while local fine-tuning on regional data (e.g., Thai CXRs) increased AUROC to 0.91 [107].

Yin [42] addressed domain shift in pulmonary nodule detection by aligning CT scan features from different scanners. They proposed adversarial frequency alignment (AFA), which uses frequency attention maps and adversarial training to create domain-invariant representations while preserving image fidelity. Their approach, tested on a reorganized LUNA16 dataset (LUNA-DG), outperformed other domain generalization methods and offers a promising strategy to improve model generalization across diverse imaging settings.

Data harmonization techniques are not widely used because they require the availability of data from both populations (source and target domains). As such, the methods might not be suitable for situations where target domain data is not available due to privacy or underrepresentation reasons. Table 5 presents a summary of data-centric techniques for mitigating CPDS.

Table 5. Summary of data-centric approaches for CPDS mitigation in medical image analysis.

Article	Category	Methodology	Findings/Performance
Zhang [102]	Data-centric	Deep stacked transformation (BigAug) using long augmentation pipeline mimicking domain shifts; encourages feature invariance.	Improves generalization to unseen modalities and institutions.
Chen [115]	Data-centric	GAN-based augmentation of MR images for realistic variability.	Improves Dice score and generalization to out-of-domain sequences.
Zhu [98]	Data-centric	Shared latent Gaussian mixture to generate synthetic bridge images.	Smoother modality adaptation and better segmentation.
Garrucho [118]	Data-centric	Single-centre training evaluated on unseen centres.	Outperforms cross-domain pooling in 4/5 centres.
Li [116]	Data-centric	Fourier-based augmentation via amplitude spectrum diversification.	Boosts unseen-modality performance with low computational cost.
Chamveha [107]	Data-centric	Local retraining of CNN on target population.	AUROC = 0.91; reduces workload by 5.6%.
Xue [22]	Data-centric	Cross-centre lung-region detection evaluation.	Improves mAP from 0.5 to 0.95.

Table 5 shows that data-centric strategies like augmentation, data harmonization, and cross-domain fine-tuning effectively reduce covariate shift (CPDS) in medical imaging. Methods such as deep CNNs, GAN-based augmentation, and BigAug improve prediction accuracy, Dice scores, and model robustness on new domains. Additionally, combining datasets from different populations and local retraining further boosts generalizability. Overall, data diversity and augmentation are key to overcoming CPDS challenges.

4.2.2. Model-Centric Approaches

Model-centric approaches modify model architectures or learning strategies to better handle the variability of the target domain. They do not operate on the data. They operate at the training stage to enhance domain robustness. Five categories were identified.

Domain Adaptation:

Domain adaptation (DA) is a technique aimed at reducing distributional discrepancy between source and target domains. This can be implemented through multiple strategies, including discrepancy minimization (e.g., MMD, CORAL), feature normalization, or adversarial training. Supervised, semi-supervised, and unsupervised DA have all been explored [33]. Among these, Adversarial DA, where a discriminator enforces domain confusion, is the most effective and widely used DA technique [18,103]. Musa et al. achieved a 28% accuracy gain on Nigerian CXRs using supervised ADA, while He et al. [114] improved average AUC by 7–12% using a Wasserstein-discrepancy DA model.

The main goal is to learn a feature representation f that predicts well and is indistinguishable between domains, using an objective function with two components.

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{task}}(f(x_S), y_S) + \lambda \cdot \mathcal{L}_{\text{domain}}(f(x_S), f(x_T)) \quad (8)$$

where:

$\mathcal{L}_{\text{task}}$ is the task loss (e.g., cross-entropy) on labeled source data (x_S, y_S),
$\mathcal{L}_{\text{domain}}$ measures the distance between source $f(x_S)$ and target $f(x_T)$,
λ is a hyperparameter that balances the two objectives.

The domain discrepancy $\mathcal{L}_{\text{domain}}$ is measured using methods like Maximum Mean Discrepancy (MMD), adversarial losses (DANN), or statistical moment matching (CORAL) [141,142].

Domain adaptation is particularly relevant in CPDS imaging, where variability arises from demographics and access to target data is somewhat limited or unavailable [114]. The majority of studies in CPDS utilize one form of DA to minimize the domain shift between source and target data [18,21,103,123,143].

Domain Generalization (DG) resembles Domain Adaptation (DA) but aims to learn domain-invariant features without access to target domain data. This approach addresses the common clinical challenge where data from new hospitals is unavailable during training, making DG both more realistic and demanding [35].

Transfer Learning:

Transfer learning aims to reuse a model developed for one task (the source task) as the starting point for a model on a second, related task (the target task). Fine-tuning pre-trained models (e.g., Resnet) on small local datasets substantially improves cross-population performance [107,144]. Transfer learning is widely adopted due to its simplicity and efficiency, with accuracy improvements ranging from 5–10% in low-data settings.

Ghafoorian [145] showed that re-training a CNN model for brain lesion detection using transfer learning with target population data greatly improved accuracy. Hansun [24] used transfer learning, model ensembling, and a rejection mechanism for TB detection, boosting sensitivity and reducing false positives. These studies demonstrate that transfer learning effectively improves cross-population model performance, with limited target data.

Adversarial Learning:

Adversarial learning (AL) is a training paradigm based on a minimax optimization scheme in which competing networks (e.g., feature extractor and discriminator) are trained to enforce invariance or robustness [146]. Adversarial domain adaptation (e.g., DANN-style methods) lies at the intersection of DA and AL; it uses adversarial learning as a mechanism to achieve the objective of domain adaptation [104]. In such cases, adversarial training is not a separate mitigation category but rather an implementation strategy for distribution alignment. However, adversarial learning can also be employed beyond domain adaptation. For example, to suppress sensitive demographic attributes, enforce fairness constraints, or improve robustness to anatomical variability, without explicitly performing source–target alignment [115]. AL with Generative Adversarial Networks (GANs) was also employed together with domain adaptation or data augmentation to train a model that learns domain-invariant feature representations. However, adversarial DA is the most widely used AL approach, in which a domain discriminator is trained to distinguish source vs. target features. In contrast, the feature extractor is trained to confuse the discriminator, thereby extracting features common to both domains [14]. The objective function for adversarial training is described in (9)

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (9)$$

where the discriminator (D) tries to maximize this function (make it good at spotting fakes). The generator (G) tries to minimize this function (make D fail).

GAN-based frameworks, such as CycleGANs and TUNA-Net [143], generate synthetic images in the style of the target population to bridge visual gaps. These methods achieved near-supervised AUCs (e.g., 96.3% vs. 98.1%) on pediatric pneumonia detection, demonstrating strong robustness to demographic differences.

This concept of synthetic data generation has also been applied in other contexts. For example, Mahmoud [147] utilized a GAN with reverse flow to generate synthetic endoscopy images for underrepresented groups, thereby improving accuracy and robustness. Such augmented diversity can mitigate domain shift by exposing the model to a wider variety of appearances.

In the unsupervised setting (no target labels), this generative image-level adaptation approach yielded the best performance among several DA methods compared to other methods [106]. Valliani [148] uses Adversarial domain adaptation to improve model performance on radiographic data derived from multiple out-of-sample healthcare populations. Their adapted models outperformed non-adapted models. Beyond adversarial training, some works use discrepancy-based DA.

Adversarial augmentation techniques introduce domain variability that helps models learn invariant features and improve robustness to domain shifts, though they can be computationally demanding and require careful validation of synthetic images. Adversarial domain adaptation remains the most widely used approach for mitigating cross-population domain shift (CPDS).

Multi-Task Learning (MTL):

MTL jointly trains auxiliary tasks (e.g., segmentation + classification) to learn shared representations and improve generalization [108]. While some studies report modest generalization improvements, results vary depending on task alignment and data diversity [149]. For instance, one approach is to add a domain classification head to the network (predicting the source of the data) in parallel with disease prediction, allowing the model to be explicitly aware of domain differences [150].

However, some studies suggest that the effectiveness of MTL for CPDS depends on the type of data and nature of the task [151]. Additionally, Jie et al. [152] found that adding a domain-classification task without adversarial inversion yielded only minor gains. The shared representations in their multi-task model did not significantly enhance cross-population generalization.

Continual Learning (CL):

Continual learning updates models sequentially with new population data while retaining prior knowledge. Techniques such as Elastic Weight Consolidation (EWC) and Learning without Forgetting (LwF) mitigate catastrophic forgetting. improved target-domain AUC by over 10% without full retraining, suggesting continual adaptation as a practical strategy for evolving clinical data [27].

By adjusting to changes in distribution over time and making sure that models remain up to date with a variety of populations, continuous learning can help address CPDS. This is particularly important in practice because models deployed in hospitals might require retraining regularly when they are exposed to new patient demographics. However, continual learning techniques require access to the target domain data, which limits their potential in certain places where new population data is not available.

Collectively, model-centric approaches, especially adversarial DA and transfer learning, are the most widely used and empirically successful methods for CPDS mitigation. They directly target population variability and consistently outperform non-adaptive baselines. The summary of model-centric approaches is presented in Table 6.

Table 6. Summary of model-centric approaches for CPDS mitigation and reported outcomes.

Study	Technique (s)	Modality/Domain Context	Reported Outcomes
[18]	Supervised adversarial domain adaptation (domain-adversarial training with a discriminator) to learn domain-invariant features.	Chest X-ray data from three populations.	Improved target performance with Accuracy ≈ 90%, AUC ≈ 0.96 on Nigerian test set, outperforming baseline models.
[74]	Multitask learning for glioma segmentation and IDH genotyping to mitigate data heterogeneity	Brain MRI	The proposed multi-task network, MTTU-Net, improves the Dice score by 1.23% for glioma segmentation, and improves the AUC and accuracy by 2.13% and 4.2%, respectively.
[150]	Federated learning for domain generalization (collaborative training across institutions without data pooling).	Multiple public CXR datasets.	The model outperformed single-dataset training in cross-institution tests, yielding higher AUROC on unseen hospital data.
[138]	Neural-style transfer to overcome the generalization limitations of deep learning segmentation models	3D cardiovascular MRI	Improved dice score by 29.9% compared to the baseline models.
[147]	Unsupervised reverse domain adaptation for synthetic medical images via adversarial training	Endoscopy	improved AUC by 78.7% by using reverse domain adaptation.
[148]	Ameliorating the effect of dataset shift using generative adversarial networks (GANs)	chest radiography	The model achieved an average internal test AUC of 78.07% and an average external test AUC of 71.43%.
[146]	unsupervised domain adaptation to transfer the discriminative knowledge obtained from the source to the target domain without labels	prostate cancer histopathology WSI	The network achieved 76.9% accuracy, and the TCGA network achieved 83.0% accuracy, both outperforming the 73.5% baseline.
[24]	Ensemble transfer learning + rejection mechanism (EfficientNet ensemble with post-hoc uncertainty rejection).	TB detection from CXR; Montgomery (USA) vs. Shenzhen (China) datasets.	Achieved 94.9% (Montgomery) and 92.8% (Shenzhen) accuracy, TB detection improved from 84–88%.
[144]	Baseline cross-population evaluation (no adaptation; trained/tested across populations).	CXR abnormality screening across national datasets.	Observed major performance drop across populations, confirming CPDS impact and the need for adaptation.
[22]	Multi-source training and feature visualization for domain generalization.	Lung region detection, Public CXRs sources.	Training on merged data improved performance (mAP 0.954 → 0.978 on JSRT).
[105]	Attention-guided partial domain adaptation (aligning shared classes via attention).	Pneumonia CXR datasets.	Improved target-domain accuracy/AUC vs. baseline unsupervised DA.
[103]	Unsupervised domain adaptation (UDA) via feature alignment (Wasserstein distance minimization).	Cardiomegaly detection in CXR.	Adapted models outperformed source-only models on AUC, confirming the effectiveness of DA.
[137]	Dynamic extension networks with class-boundary alignment.	CXR classification (public vs. local hospital datasets).	Reduced domain gap, improved target-domain accuracy while preserving source-domain performance.
[27]	Continual learning (Joint Training, EWC, LwF) for sequential adaptation.	ChestX-ray14 → MIMIC-CXR transfer.	Improved target performance and reduced forgetting. JT AUC = 92.66%, LwF = 82.37%.
[153]	Multi-task Supervised Contrastive Learning (MTSCL) framework.	CXR datasets for lung segmentation, COVID-19, and abnormality detection.	AUC improvement of 1.03–5.33%, average accuracy 75.19%.
[142]	End-to-end CNN-GNN for multi-label thoracic disease classification.	NIH ChestX-ray14 dataset.	Achieved mean AUC = 82.66%, outperforming standard CNN baselines.
[148]	GAN-based domain adaptation.	CXR data from four medical centers.	Target AUC improved from 71.4% to 73.8%, showing benefit of adversarial adaptation.
[111]	Selective optimization of underserved groups to reduce bias.	MIMIC-CXR (target) and local datasets.	Improved fairness across gender, age, and ethnicity without loss of overall performance.

4.3. Common Datasets and Evaluation Metrics in CPDS Studies (RQ3)

RQ3: What datasets and evaluation metrics are commonly used to assess the effectiveness of CPDS mitigation techniques?

In this subsection, we discussed common datasets and evaluation metrics used to assess the impact of CPDS.

4.3.1. Datasets

The studies on cross-population domain shift draw upon a variety of medical imaging datasets, often pairing a large source dataset with a different-population target dataset from diverse hospitals (primarily North American), which serve as rich source domains. The most frequently used source datasets include NIH ChestX-ray14 [154], CheXpert [46], and MIMIC-CXR [155], representing large, Western hospital cohorts. Target datasets often come from different geographic or demographic regions, such as PadChest (Spain) [156], VinDr-CXR (Vietnam) [157], Shenzhen (China), or locally collected datasets from LMICs (e.g., Nigeria, Thailand). This pattern highlights the dominance of Northern data sources and the relative scarcity of African and South Asian datasets.

Cross-population studies typically involve at least two datasets: one from a well-established international database (e.g., NIH, MIMIC, CheXpert) and another from a different region or patient group (e.g., country-specific collections, such as those from India, China, Nigeria, or pediatric cohorts). Table 7 lists the public datasets used in each study from our selection.

Table 7. A Sample of Open-Source Medical Imaging Datasets and Repositories.

Dataset Name	Primary Authors/Institution	Modality	Approx. Size	URL
MIMIC-CXR	Johnson et al. (MIT, BIDMC), USA	Chest X-ray (CXR)	377,110 images	https://physionet.org/content/mimic-cxr/2.0.0/ (accessed on 10 December 2025).
CheXpert	Irvin, et al. (Stanford AIMI), USA	Chest X-ray (CXR)	224,316 images	https://stanfordmlgroup.github.io/competitions/chexpert/ (accessed on 21 November 2025).
LIDC-IDRI	Armato et al. (TCIA Consortium), USA	Lung CT	1018 cases	https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI (accessed on 10 November 2025)
BraTS Challenge	Menze et al. (BraTS Consortium), Intl.	Multi-modal MRI (T1, T1c, T2, FLAIR)	Varies by year	https://www.med.upenn.edu/cbica/brats/ (accessed on 10 November 2025)
OASIS	Marcus et al. (Washington Univ.), USA	Brain MRI	3 datasets	https://www.oasis-brains.org/ (accessed on 08 December 2025)
fastMRI	Zbontar et al. (NYU Langone Health & FAIR), USA	Knee & Brain MRI	8.4 TB total	https://fastmri.med.nyu.edu/ (accessed on 20 November 2025)
Camelyon16/17	Ehteshami Bejnordi et al. (Radboud UMC), Netherlands	Whole-Slide Images (Histopathology)	399 slides (Cam16)	https://camelyon16.grand-challenge.org/ (accessed on 10 October 2025)
PadChest	Bustos et al. (Univ. of Valencia), Spain	Chest X-ray (CXR)	160,868 images	https://bimcv.cipf.es/bimcv-projects/padchest/ (accessed on 21 October 2025)
ChestXray-14 (NIH)	Wang et al. (NIH Clinical Center), USA	Chest X-ray (CXR)	112,120 images	https://nihcc.app.box.com/v/ChestXray-NIHCC (accessed on 21 October 2025)
TCIA (Repository)	NCI/TCIA Consortium, USA	Various (CT, MRI, PET, WSI, etc.)	100s of collections (hosts intl.)	https://www.cancerimagingarchive.net/ (accessed on 10 October 2025)
ADNI	ADNI Consortium (PI: Mueller, S.G., et al.), USA/Canada	MRI, PET	Varies by phase	https://adni.loni.usc.edu/ (accessed on 16 December 2025)
COVID-19 Image Data Collection, Intl.	Cohen, J. P., Morrison, P., & Dao, L.	CXR, CT	1100+ images	https://github.com/ieee8023/covid-chestxray-dataset (accessed on 16 December 2025)
TCGA (Repository)	NCI/NHGRI (TCGA Research Network), USA	Histopathology, MRI, CT	>20,000 cases	https://wiki.cancerimagingarchive.net/display/Public/TCGA+Collections (accessed on 25 December 2025)
COVID19-CT-Dataset	Zhao, J., Zhang, Y., et al. China	CT	349 CT scans	https://github.com/UCSD-AI4H/COVID-CT (accessed on 25 December 2025)

Table 7. *Cont.*

Dataset Name	Primary Authors/Institution	Modality	Approx. Size	URL
Guangzhou Pediatric CXR (Pneumonia)	Kermany, D. S., et al. (Guangzhou Women & Children's Med. Ctr.), China	Chest X-ray (CXR)	5863 images	https://data.mendeley.com/datasets/rscbjbr9sj/2 (accessed on 10 October 2025)
Nigeria CXR Dataset	Musa et al. (Kaggle), Nigeria	Chest X-ray (CXR)	2000 images	https://www.kaggle.com/datasets/aminumusa/nigeria-chest-x-ray-dataset (accessed on 21 November 2025)
JSRT Dataset	Horry, et al. (JSRT), Japan	Chest X-ray (CXR) lung nodules	154 images	https://wiki.cancerimagingarchive.net/display/Public/JSRT (accessed on 10 December 2025)

In Table 7, we summarize the selection of prominent open-source medical imaging datasets and repositories that were used for benchmarking in CPDS studies. The table categorized each dataset by name, primary authors or institution, and the specific imaging modality, which includes diverse types like Chest X-ray (CXR), Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and whole-slide histopathology. Furthermore, the table quantifies the approximate size of each dataset, identifies its country of origin, and provides a direct URL for access, offering a valuable resource for researchers seeking data for conditions ranging from lung diseases and brain tumors to Alzheimer's and COVID-19.

4.3.2. Evaluation Metrics

Despite the increasing application of deep learning in medical imaging analysis, the assessment of model performance across heterogeneous populations continues to lack standardization. While there is no single metric to measure and quantify domain shift across all scenarios, researchers have relied on statistical distribution distance to measure domain shift (e.g., maximum mean discrepancy (MMD), which often correlates with model performance degradation.

The majority of reviewed studies (85%) predominantly utilize conventional evaluation metrics such as accuracy, AUC, and F1-score [14,27,99,107,118,142,151]. However, these metrics often fail to comprehensively measure performance degradation arising from cross-population domain shifts. For example, AUC is often the most reported performance metric due to its robustness to class imbalance. Accuracy, sensitivity, specificity, F1-score, and G-Mean complement AUC in evaluating model reliability across domains [103]. A strong CPDS mitigation method minimizes AUC or accuracy drop from source to target domains, often a 5–20% gap without adaptation, reduced to <10% after applying domain-adaptive methods [18,114]. Statistical significance tests are occasionally mentioned to confirm that improvements resulting from adaptation are significant (e.g., $p < 0.05$). By and large, though, the focus is on practical metrics that quantify generalization [111]. Therefore, to move toward equity, performance must be measured with fairness-aware metrics that go beyond overall accuracy. While standard metrics are essential, they can mask significant performance disparities across subgroups.

Table 8 outlines the primary evaluation metrics used in each study. As seen, AUC and accuracy are nearly universal, often showing how the existing work falls short in quantifying bias and equity. State-of-the-art fairness metrics, such as Positive-Sum Fairness, Group Benefit Equality, and intersectional fairness, such as FAIR-MED, should be employed [158].

Additionally, while AUC is less sensitive to class imbalance, it frequently does not capture disparities in predictive performance among demographic subgroups, thereby overlooking issues of fairness. Sensitivity and specificity, though clinically significant, may also fluctuate substantially across populations due to intrinsic differences in underlying data distributions. Table 8 summarizes the most commonly employed evaluation

metrics, detailing both their primary functions and notable shortcomings when applied in the context of domain shifts. These constraints underscore the critical need to select evaluation measures that reflect not only predictive accuracy but also fairness, calibration, and generalizability across diverse demographic cohorts. In future research, the adoption of robust, context-sensitive metrics, such as subgroup-disaggregated reporting and calibration-aware scores, will be essential to promote equitable and reliable deployment of AI in medical imaging.

Table 8. Comparison of Evaluation Metrics used in CPDS studies.

Metric	Purpose	Limitations Under Domain Shift
AUC (Area Under the Curve)	Measures the ability of the model to distinguish between classes. Useful under class imbalance.	May not reflect fairness across subgroups if the threshold is fixed.
Accuracy	Overall proportion of correct predictions can be misleading under imbalance.	Inflated if one class dominates; not sensitive to imbalance.
F1 Score	Harmonic mean of precision and recall. Balances false positives and negatives.	Sensitive to class distribution; unstable when one class is rare.
Sensitivity (Recall)	True positive rate. High sensitivity means fewer false negatives.	Performance may vary greatly across subgroups; not population-invariant.
Specificity	True negative rate. High specificity reduces false positives.	Less meaningful if the negative class is rare or the context is skewed.
Precision (PPV)	Proportion of correct identifications. Affected by prevalence.	Can be unstable across domains with different prevalence rates.
Balanced Accuracy	Average of sensitivity and specificity. Handles imbalance better than plain accuracy.	More robust, but still doesn't account for subgroup fairness.
Youden's Index	Summarizes the performance of sensitivity + Specificity-1.	Rarely used, doesn't capture the calibration.
Matthews Correlation Coefficient (MCC)	takes all four confusion matrix categories into account. More informative under an imbalance.	Not widely adopted in medical DL; complex to interpret.
Brier Score	Measures the accuracy of probabilistic predictions. Lower is better.	Assumes probabilistic output; hard to calibrate across population.

4.4. Limitations and Research Gaps

RQ4: What limitations and research gaps remain in current CPDS literature?
Despite notable progress, the review identifies several persistent gaps:

- **Limited external validation:** Most studies evaluate models on a single target dataset, lacking rigorous cross-population testing.
- **Data imbalance and underrepresentation:** African, South American, and low-resource populations remain largely absent from public medical datasets.
- **Inconsistent evaluation protocols:** Performance metrics and data splits vary widely, hindering reproducibility.
- **Overreliance on Western datasets:** Nearly 80% of reviewed studies used NIH, CheXpert, or MIMIC-CXR as their primary source.
- **Ethical and fairness considerations:** Few studies quantify or mitigate demographic bias explicitly.

Future research should prioritize developing inclusive global datasets, standardized cross-population benchmarks, and fairness-aware domain adaptation techniques. Continuous learning and federated paradigms offer promising directions for scalable, privacy-preserving generalization across diverse clinical environments.

4.5. Discussion

This review provides an exhaustive synthesis of how cross-population domain shift (CPDS) primarily manifests and is mitigated in deep learning–based medical image analysis. CPDS fa as a covariate shift driven by demographic differences, such as age, sex,

and ethnicity, which alter image distributions and degrade model performance across cohorts. Variations in equipment, acquisition protocols, and disease prevalence compound this effect, leading to notable accuracy drops when algorithms are applied to data from new hospitals or regions. Thus, improving model generalization requires addressing both population heterogeneity and imaging variability. While numerous techniques have been proposed, no single strategy completely resolves CPDS. Even state-of-the-art models remain biased toward specific demographic groups, such as paediatric or minority populations, confirming a persistent research gap [22]. To ensure fair and robust AI, both model development and validation must explicitly consider demographic diversity and dataset representativeness.

4.5.1. Methods for Addressing Cross-Population Domain Shift

Mitigation strategies can be broadly classified into *data-centric* and *model-centric* approaches. Data-centric methods enhance or harmonize datasets to reflect target populations more accurately. Advanced augmentation, synthetic data generation (e.g., GAN-based style transfer), and inter-site normalization reduce variability in imaging appearance. For instance, Choi et al. [28] demonstrated that GAN-generated image variants improve robustness to unseen domains.

Model-centric approaches, by contrast, adjust architectures or the learning process to mitigate CPDS. Domain adaptation, especially adversarial adaptation using discriminators, remains the most popular strategy, as shown by Lafarge et al. [104] for histopathology and by subsequent works in CXR analysis using models like TUNA-Net. Multi-task learning, where networks jointly learn disease classification and auxiliary tasks (e.g., demographic or segmentation), has also shown promise in encouraging generalizable representations [159]. Transfer learning continues to be effective when limited target-domain data are available, while continual learning enables progressive adaptation as new populations emerge [27].

Federated learning further extends this principle by enabling distributed training across institutions without sharing sensitive data, enabling collaborative model training without centralized data pooling. These studies [150,160,161] have demonstrated improved generalization using federated learning, where models were evaluated across multiple hospitals. Collectively, the studies indicate that combining multiple strategies, diverse data, adaptive models, and privacy-preserving collaboration yields the best progress toward cross-population robustness.

Consequently, beyond algorithmic factors, CPDS is strongly shaped by privacy regulations and data governance frameworks that restrict access to demographically diverse medical data [162]. Regulations such as the General Data Protection Regulation (GDPR) in Europe and the Health Insurance Portability and Accountability Act (HIPAA) in the United States limit cross-institutional and cross-border data sharing, fragmenting datasets along geographic and institutional lines [163,164]. While essential for patient protection, this fragmentation reduces exposure to underrepresented populations during model development, thereby exacerbating CPDS [97,160]. Privacy-preserving approaches such as federated learning partially address data-sharing constraints but introduce additional challenges, including non-identically distributed client populations, heterogeneous labeling practices, and limited access to demographic attributes for fairness evaluation, which may amplify bias if not carefully managed [158,165]. Collectively, these factors emphasize that CPDS is not only a technical challenge but also a structural and regulatory one, necessitating solutions that jointly consider algorithmic robustness and policy-aware data governance.

Although many mitigation techniques have demonstrated success under general domain shift, particularly scanner, vendor, or protocol-induced variability [35,41,114,116,118], their effectiveness does not reliably translate to cross-population domain shift (CPDS).

Mitigation methods that perform well under general domain shift often fail under cross-population domain shift (CPDS). CPDS reflects not only distributional differences but also biologically and clinically meaningful variation across populations that alters the relationship between image features and diagnostic labels [151,166]. Statistical alignment techniques, such as adversarial adaptation and image harmonization, are effective when domain differences primarily affect appearance (e.g., scanner or protocol shifts) but rely on the assumption of invariant label semantics, which frequently breaks under CPDS [99,106,130,136,146].

While feature-alignment methods such as adversarial domain adaptation effectively minimize distributional divergence in input features (i.e., addressing covariate shift), they operate under the assumption that the conditional label distribution $P(Y \mid X)$ remains invariant across domains. In CPDS scenarios, however, concept shift frequently arises due to population-specific disease manifestations, heterogeneous diagnostic criteria, or varying clinical thresholds [92,99]. When the semantic relationship between imaging features and diagnostic labels differs across populations, aligning feature distributions without accounting for label-shift can inadvertently suppress clinically relevant, population-specific biomarkers or amplify annotation biases [14,90]. Consequently, mitigation strategies for CPDS must incorporate mechanisms that explicitly model or adapt to changes in $P(Y \mid X)$, such as label-aware adaptation, uncertainty-calibrated prediction, or subgroup-specific threshold optimization [100,137].

In contrast, the results from the reviewed literature indicate that representation-centric approaches, including self-supervised learning, multi-task learning, and continual adaptation, better handle CPDS by promoting clinically meaningful invariance while allowing controlled population-specific learning [22,27,107,138,145,153,159,165]. These observations highlighted the need for population-aware method selection and disaggregated evaluation, as aggregate performance metrics often mask failures in underrepresented groups.

4.5.2. Datasets and Evaluation Metrics Benchmarks

Datasets

Beyond simple underrepresentation, important geographic differences in disease epidemiology and imaging practice contribute directly to CPDS [83,101]. Large datasets such as ChestX-ray14 [154], CheXpert [46], and MIMIC-CXR [155] primarily reflect adult populations with higher prevalence of chronic cardiopulmonary conditions and standardized digital acquisition protocols. In contrast, datasets from LMIC contexts, including VinDr-CXR [157] and locally collected African or Asian cohorts, often contain higher proportions of infectious diseases such as tuberculosis and advanced pneumonia, frequently captured at later clinical stages [18]. Differences in scanner type, portability, exposure control, and image digitization further alter contrast, noise characteristics, and anatomical visibility, inducing covariate shift [21]. Additionally, variations in diagnostic thresholds and annotation practices across regions introduce potential concept shift [98]. These combined epidemiological and technical disparities illustrate that CPDS reflects structural healthcare and infrastructure differences rather than a purely statistical distribution gap.

Evaluation Metrics

In the reviewed literature, the evaluation metrics predominantly utilized are aggregate performance metrics, including Area Under the Curve (AUC), accuracy, F1-score, sensitivity, specificity, and the Dice coefficient. These aggregate metrics continue to serve as the primary evaluation criteria within CPDS research [35,151,159,165]. This review reveals that fairness-aware and subgroup-level evaluation is rarely operationalized in practice [164]. Among the reviewed studies included in this review, the majority (85%) reported only aggregate

metrics, without stratifying performance by demographic or population subgroup. A smaller subset of studies (15%) reported subgroup-disaggregated performance, typically stratified by age, sex, or acquisition site, most commonly in the form of subgroup-specific AUC or sensitivity.

Fairness Evaluation

However, aggregate metrics often fail to capture performance disparities across demographic subgroups. Accuracy can be unreliable in imbalanced datasets, and AUC, while robust to imbalance, may obscure fairness issues. For instance, a model reported over 75% accuracy simply by predicting most of the cases as negative, given the low prevalence of positive cases in cancer classification with pathology and genomic features [131]. Additionally, Chamveha et al. [107] reported clear AUC degradation when testing CXR models across datasets, illustrating how traditional metrics underestimate domain shift.

To systematically evaluate algorithmic fairness in CPDS, researchers should move beyond aggregate scores to assess group fairness metrics, including Demographic Parity (equal prediction rates across groups), Equalized Odds (equal true positive and false positive rates), and Calibration (consistent risk interpretation across groups) [167,168]. In clinical contexts, Equalized Odds and Calibration are often more ethically appropriate than Demographic Parity, as the latter may ignore legitimate differences in disease prevalence across populations [169]. However, applying these metrics in CPDS scenarios presents challenges, particularly regarding the availability of robust protected attribute data (e.g., race, socioeconomic status) and the risk of perpetuating historical biases present in electronic health records [170]. To actively promote algorithmic fairness through evaluation, we recommend a three-tiered approach: (1) Stratified Reporting, where performance metrics are mandatorily disaggregated by key demographic variables during validation; (2) Threshold Optimization, where decision thresholds are tuned per subgroup to equalize error rates rather than maximizing global accuracy; and (3) Uncertainty Quantification, ensuring that model confidence is calibrated across groups to prevent over-reliance on predictions for underrepresented populations [171]. By embedding these fairness-aware metrics into the standard evaluation pipeline, researchers can transition from passive bias detection to active bias mitigation, ensuring CPDS tools enhance rather than exacerbate health inequities.

Consequently, explicit use of formal fairness metrics was rare. Only 6 studies (12%) employed metrics such as equalized odds, demographic parity, or related fairness criteria, despite frequent conceptual references to fairness and bias mitigation [165,166,172,173]. Even among studies reporting subgroup performance, few assessed disparities systematically or evaluated worst-group performance, and none adopted fairness metrics as a primary evaluation endpoint. These findings indicate a clear gap between the growing theoretical emphasis on fairness in CPDS and its practical implementation in empirical studies [101].

Table 6 presents evaluation metrics across the reviewed literature, distinguishing between aggregate-only evaluation, subgroup-stratified reporting, and formal fairness metric usage. The results indicate that most CPDS studies remain evaluated under assumptions of population homogeneity, potentially masking clinically meaningful performance disparities across demographic groups.

Benchmark Datasets

Despite widespread adoption of benchmarking datasets, inconsistencies in these benchmarks persist. Large datasets such as NIH ChestX-ray14 and MIMIC-CXR are commonly used as sources, but few standardized cross-population benchmarks exist. Studies typically

assemble their own multi-institutional combinations, hindering comparability [118]. Establishing shared benchmarks (e.g., training on Dataset A, testing on Dataset B) would enable fairer comparisons and accelerate reproducibility.

Figure 7 demonstrates that the majority of publicly available medical imaging datasets are sourced primarily from the United States and China, resulting in the marked underrepresentation of regions such as Africa and Latin America. This geographic concentration introduces significant bias to existing benchmarks, thereby limiting the generalizability and fairness of evaluations in medical imaging research. The analysis also identifies a subset of datasets classified as 'Intl', which were compiled through collaborations across multiple countries or institutions. For example, Wang et al. [174] developed an 'Intl' dataset that aggregates imaging data from diverse international sources. Such datasets are especially significant because they offer a more representative sample of global populations and can partially address the lack of regional diversity. Nevertheless, the overall availability of public medical imaging datasets remains insufficient on a global scale. Consequently, the majority of studies reported in the literature rely on private datasets, largely due to concerns regarding patient privacy and data security [68,75,139].

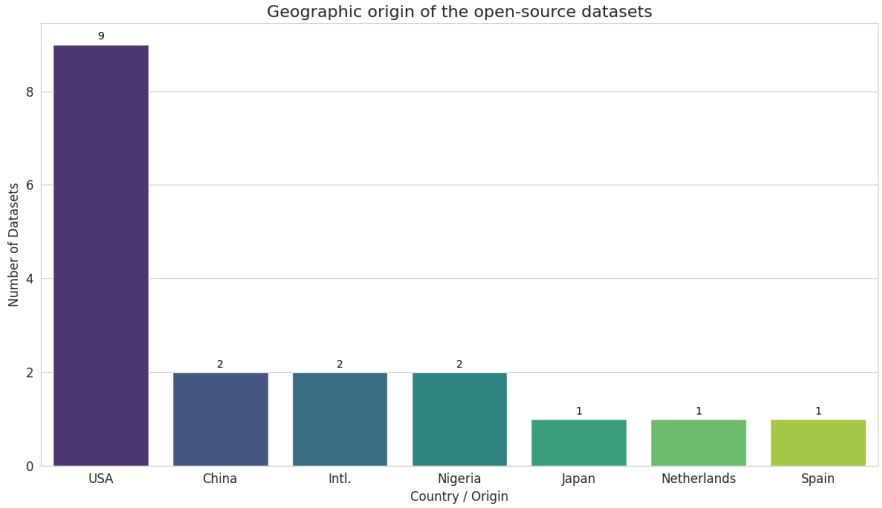


Figure 7. Dataset Representation and Geographic Bias.

4.5.3. Applications and Real-World Implementations

Real-world deployment remains the ultimate validation of CPDS mitigation. Several studies have conducted external evaluations, where models trained on one population are tested on independent sites or regions. Results consistently show performance degradation without adaptation, as seen with CheXNet [9], which achieved radiologist-level accuracy on NIH data but underperformed on external datasets. Conversely, federated and multi-center collaborations have shown that training with diverse data improves generalization [118]. Commercial CAD systems for TB and COVID-19 now increasingly emphasize global validation, signaling a move toward fairer AI tools. Nonetheless, most CXR models remain research prototypes pending comprehensive cross-site trials [175].

4.5.4. Challenges and Future Directions

Persistent challenges include the scarcity of representative datasets, the risk of algorithmic bias, and limited interpretability. Most available datasets overrepresent Western or East Asian populations, resulting in models that perform poorly in low- and middle-income settings. Addressing this requires inclusive data curation and the careful validation of synthetic augmentation. Fairness-aware training, as noted by Lee et al. [92], must become a routine part of model development to mitigate race- or gender-related disparities.

Additionally, model evaluation must look beyond aggregate performance; future CPDS studies should adopt fairness-aware evaluation protocols tailored to cross-population deployment. At a minimum, performance should be reported using subgroup-disaggregated sensitivity, specificity, and AUC across relevant demographics (e.g., age, sex, ethnicity, and geography), alongside disparity measures such as worst-group performance and inter-group performance gaps. Where annotations permit, formal fairness criteria, including equalized odds and equal opportunity, are particularly suitable for clinical tasks, as they assess parity in error rates and true positive rates across populations. Subgroup-stratified calibration error should also be reported to ensure probabilistic predictions remain clinically reliable across populations. Collectively, these metrics offer a practical and clinically grounded baseline for evaluating fairness and robustness under CPDS.

Explainability and transparency are equally vital: clinicians must understand model reasoning to trust AI outputs. Future systems should integrate interpretability modules, visual heatmaps, and textual explanations. Emerging frameworks such as federated and continual learning will likely dominate as they enable models to evolve with new data while preserving privacy. Finally, standardized benchmarks and fairness-aware evaluation protocols with fairness metrics, such as demographic parity and equalized odds, were not seen in the literature review; this needs to be addressed in the future to measure progress consistently.

Future work should prioritize: (i) test-time adaptation frameworks that enable on-the-fly model calibration using limited target-domain data without full retraining [130]; (ii) parameter-efficient fine-tuning of foundation models (e.g., MedSAM, BiomedCLIP) for population-specific adaptation while preserving general knowledge; (iii) causal representation learning to disentangle population-invariant disease features from confounding demographic factors; and (iv) federated continual learning protocols that support privacy-preserving, sequential adaptation across heterogeneous clinical sites [162].

5. Conclusions

This systematic review has examined the growing body of literature addressing cross-population domain shift in deep learning–based medical imaging analysis. Across modalities, from radiography and CT to MRI, ultrasound, and histopathology, our findings confirm that models trained on one population often fail to generalize to others, primarily due to demographic, geographic, and technical variability. Differences in patient age, sex, ethnicity, scanner type, and acquisition protocol introduce systematic distributional shifts that reduce performance when models are applied to new cohorts.

A wide range of mitigation strategies has been explored, including data augmentation, transfer learning, domain adaptation, adversarial training, and harmonization. Our taxonomy shows that while individual methods can yield measurable improvements, no single strategy alone can fully address CPDS. Instead, the most successful solutions integrate diverse data sources, adaptive learning algorithms, and rigorous external evaluation.

By synthesizing 50 peer-reviewed studies, this review highlights key methodological trends, commonly used datasets, and evaluation metrics such as AUC, accuracy, and F1-score. We identify persistent challenges, including data imbalance, limited representation from low- and middle-income countries, and inconsistent validation practices that hinder progress toward equitable AI in healthcare. To achieve fairness and reliability, the research community must prioritize inclusive data curation, transparent model evaluation, and standardized benchmarks that explicitly test cross-population generalization.

Ultimately, overcoming cross-population domain shift is not only a technical burden but also a moral responsibility. Ensuring that AI-based diagnostic systems perform equitably across all populations is crucial to realizing the full potential of medical imag-

ing analysis with AI, thereby delivering accurate, trustworthy, and universally beneficial healthcare innovations.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/bdcc10030076/s1, File S1: PRISMA checklist.

Author Contributions: A.M.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Visualization, Writing original draft, Writing review, editing, and Validation. R.P.: Conceptualization, Methodology, Project administration, Supervision, Data curation, and Validation. P.O.: Conceptualization, Methodology, Data curation, and Supervision. M.H.: Conceptualization, Supervision, Resource, Methodology, Writing review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Spanish Ministry of Science, Innovation, and Universities project PID2022-138703OB-I00 (Trust-BEyE), Government of Aragon Order ECU/1871/2023 PROY-B50-24, RICORS network of inflammatory diseases from Carlos III Health Institute Network RD24/0007/0022, and Government of Aragon COS2MOS research group T64 23R. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability Statement: No new data were created or analyzed in this study.

Acknowledgments: During the preparation of this manuscript, the authors used language-editing tools, including Grammarly and ChatGPT 5.0, to assist with grammar and clarity. All content generated or revised with these tools was thoroughly reviewed, verified, and edited by the authors, who take full responsibility for the final version of the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Suetens, P. *Fundamentals of Medical Imaging*; Cambridge University Press: Cambridge, UK , 2017.
2. Hussain, S.; Mubeen, I.; Ullah, N.; Shah, S.S.U.D.; Khan, B.A.; Zahoor, M.; Ullah, R.; Khan, F.A.; Sultan, M.A. Modern diagnostic imaging technique applications and risk factors in the medical field: A review. *BioMed Res. Int.* **2022**, *2022*, 5164970. [CrossRef]
3. E Kalkman, G.B. Melanoma. *Clin. Radiol.* **2004**, *59*, 313–326. [CrossRef]
4. Lukande, R.; Kyokunda, L.T.; Bedada, A.G.; Milner, D. Pathology for Thoracic Conditions in Low- and Middle-Income Countries. *Thorac. Surg. Clin.* **2022**, *32*, 299–306. [CrossRef]
5. Ibrahim, A.U.; Ozsoz, M.; Serte, S.; Al-Turjman, F.; Yakoi, P.S. Pneumonia classification using deep learning from chest X-ray images during COVID-19. *Cogn. Comput.* **2024**, *16*, 1589–1601. [CrossRef]
6. Chen, K.C.; Yu, H.R.; Chen, W.S.; Lin, W.C.; Lee, Y.C.; Chen, H.H.; Jiang, J.H.; Su, T.Y.; Tsai, C.K.; Tsai, T.A.; et al. Diagnosis of common pulmonary diseases in children by X-ray images and deep learning. *Sci. Rep.* **2020**, *10*, 17374. [CrossRef]
7. Sajed, S.; Sanati, A.; Garcia, J.E.; Rostami, H.; Keshavarz, A.; Teixeira, A. The effectiveness of deep learning vs. traditional methods for lung disease diagnosis using chest X-ray images: A systematic review. *Appl. Soft Comput.* **2023**, *147*, 110817. [CrossRef]
8. Jaiswal, A.K.; Tiwari, P.; Kumar, S.; Gupta, D.; Khanna, A.; Rodrigues, J.J. Identifying pneumonia in chest X-rays: A deep learning approach. *Measurement* **2019**, *145*, 511–518. [CrossRef]
9. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv* **2017**, arXiv:1711.05225.
10. Khan, E.; Rehman, M.Z.U.; Ahmed, F.; Alfouzan, F.A.; Alzahrani, N.M.; Ahmad, J. Chest X-ray Classification for the Detection of COVID-19 Using Deep Learning Techniques. *Sensors* **2022**, *22*, 1211. [CrossRef] [PubMed]
11. Rahman, T.; Khandakar, A.; Kadir, M.A.; Islam, K.R.; Islam, K.F.; Mazhar, R.; Hamid, T.; Islam, M.T.; Kashem, S.; Mahbub, Z.B.; et al. Reliable Tuberculosis Detection Using Chest X-Ray With Deep Learning, Segmentation and Visualization. *IEEE Access* **2020**, *8*, 191586–191601. [CrossRef]
12. Hussain, E.; Hasan, M.; Rahman, M.A.; Lee, I.; Tamanna, T.; Parvez, M.Z. CoroDet: A deep learning based classification for COVID-19 detection using chest X-ray images. *Chaos Solitons Fractals* **2021**, *142*, 110495. [CrossRef]
13. Yu, K.H.; Beam, A.L.; Kohane, I.S. Artificial intelligence in healthcare. *Nat. Biomed. Eng.* **2018**, *2*, 719–731. [CrossRef]
14. Aminu, M.; Mariya, I.A.; Adamu, K.H.; Monica, H.; Yusuf, L. Analyzing Cross-Population Domain Shift in Chest X-Ray Image Classification and Mitigating the Gap with Deep Supervised Domain Adaptation. In *Medical Image Computing and Computer Assisted Intervention MICCAI 2024*; Springer Nature Switzerland: Cham, Switzerland, 2024; pp. 585–595.

15. Hong, Z.; Yue, Y.; Chen, Y.; Cong, L.; Lin, H.; Luo, Y.; Wang, M.H.; Wang, W.; Xu, J.; Yang, X.; et al. Out-of-distribution Detection in Medical Image Analysis: A survey. *arXiv* **2024**, arXiv:2404.18279.
16. Quinonero-Candela, J.; Sugiyama, M.; Schwaighofer, A.; Lawrence, N.D. *Dataset Shift in Machine Learning*; MIT Press: Cambridge, MA, USA, 2022.
17. Yang, Y.; Zhang, H.; Katabi, D.; Ghassemi, M. Change is hard: A closer look at subpopulation shift. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 39584–39622.
18. Musa, A.; Prasad, R.; Hernandez, M. Addressing cross-population domain shift in chest X-ray classification through supervised adversarial domain adaptation. *Sci. Rep.* **2025**, *15*, 11383. [CrossRef] [PubMed]
19. Waite, S.; Scott, J.; Colombo, D. Narrowing the gap: Imaging disparities in radiology. *Radiology* **2021**, *299*, 27–35. [CrossRef]
20. Moreno-Torres, J.G.; Raeder, T.; Alaiz-Rodríguez, R.; Chawla, N.V.; Herrera, F. A unifying view on dataset shift in classification. *Pattern Recognit.* **2012**, *45*, 521–530. [CrossRef]
21. Stacke, K.; Eilertsen, G.; Unger, J.; Lundström, C. A closer look at domain shift for deep learning in histopathology. *arXiv* **2019**, arXiv:1909.11575. [CrossRef]
22. Xue, Z.; Yang, F.; Rajaraman, S.; Zamzmi, G.; Antani, S. Cross Dataset Analysis of Domain Shift in CXR Lung Region Detection. *Diagnostics* **2023**, *13*, 1068. [CrossRef] [PubMed]
23. Dhar, T.; Dey, N.; Borra, S.; Sherratt, R.S. Challenges of deep learning in medical image analysis—improving explainability and trust. *IEEE Trans. Technol. Soc.* **2023**, *4*, 68–75. [CrossRef]
24. Hansun, S.; Argha, A.; Alinejad-Rokny, H.; Alizadehsani, R.; Gorriz, J.M.; Liaw, S.T.; Celler, B.G.; Marks, G.B. A New Ensemble Transfer Learning Approach with Rejection Mechanism for Tuberculosis Disease Detection. *IEEE Trans. Radiat. Plasma Med. Sci.* **2024**, *9*, 433–446. [CrossRef]
25. Farahani, A.; Voghoei, S.; Rasheed, K.; Arabnia, H.R. A brief review of domain adaptation. In *Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 877–894. [CrossRef]
26. Patricia, N.; Caputo, B. Learning to learn, from transfer learning to domain adaptation: A unifying perspective. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1442–1449.
27. Lenga, M.; Schulz, H.; Saalbach, A. Continual learning for domain adaptation in chest x-ray classification. In *Medical Imaging with Deep Learning*; PMLR: London, UK, 2020; pp. 413–423.
28. Choi, J.; Kim, T.; Kim, C. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 6830–6840.
29. Schäfer, R.; Nicke, T.; Höfener, H.; Lange, A.; Merhof, D.; Feuerhake, F.; Schulz, V.; Lotz, J.; Kiessling, F. Overcoming data scarcity in biomedical imaging with a foundational multi-task model. *Nat. Comput. Sci.* **2024**, *4*, 495–509. [CrossRef] [PubMed]
30. Guan, H.; Liu, M. Domain Adaptation for Medical Image Analysis: A Survey. *IEEE Trans. Biomed. Eng.* **2022**, *69*, 1173–1185. [CrossRef]
31. Matta, S.; Lamard, M.; Zhang, P.; Le Guilcher, A.; Borderie, L.; Cochener, B.; Quellec, G. A systematic review of generalization research in medical image classification. *Comput. Biol. Med.* **2024**, *183*, 109256. [CrossRef]
32. Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; Loy, C.C. Domain generalization: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 4396–4415. [CrossRef]
33. Stacke, K.; Eilertsen, G.; Unger, J.; Lundstrom, C. Measuring Domain Shift for Deep Learning in Histopathology. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 325–336. [CrossRef]
34. Suruchi Kumari, P.S. Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives. *Comput. Biol. Med.* **2024**, *170*, 107912. [CrossRef] [PubMed]
35. Yoon, J.S.; Oh, K.; Shin, Y.; Mazurowski, M.A.; Suk, H.I. Domain generalization for medical image analysis: A review. *Proc. IEEE* **2024**, *112*, 1583–1609. [CrossRef]
36. Dong, S.; Wang, P.; Abbas, K. A survey on deep learning and its applications. *Comput. Sci. Rev.* **2021**, *40*, 100379. [CrossRef]
37. Anaya-Isaza, A.; Mera-Jiménez, L.; Zequera-Diaz, M. An overview of deep learning in medical imaging. *Inform. Med. Unlocked* **2021**, *26*, 100723. [CrossRef]
38. Ayeni, J. Convolutional neural network (CNN): The architecture and applications. *Appl. J. Phys. Sci* **2022**, *4*, 42–50. [CrossRef]
39. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, 6000–6010. . . [CrossRef]
40. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 53. [CrossRef] [PubMed]

41. Gunasinghe, H.; McKelvie, J.; Koay, A.; Mayo, M. Domain Generalisation for Glaucoma Detection in Retinal Images from Unseen Fundus Cameras. In *Intelligent Information and Database Systems*; Springer Nature Switzerland: Cham, Switzerland, 2022; pp. 421–433. [CrossRef]
42. Yin, B.; Sun, M.; Zhang, J.; Liu, W.; Liu, C.; Wang, Z. AFA: Adversarial frequency alignment for domain generalized lung nodule detection. *Neural Comput. Appl.* **2022**, *34*, 8039–8050. [CrossRef]
43. Walsh, J.; Othmani, A.; Jain, M.; Dev, S. Using U-Net network for efficient brain tumor segmentation in MRI images. *Healthc. Anal.* **2022**, *2*, 100098. [CrossRef]
44. Li, Q.; Gao, Z.; Wang, Q.; Xia, J.; Zhang, H.; Zhang, H.; Liu, H.; Li, S. Glioma Segmentation Using a Novel Unified Algorithm in Multimodal MRI Images. *IEEE Access* **2018**, *6*, 9543–9553. [CrossRef]
45. Saratxaga, C.L.; Moya, I.; Picón, A.; Acosta, M.; Moreno-Fernandez-de Leceta, A.; Garrote, E.; Bereciartua-Perez, A. MRI Deep Learning-Based Solution for Alzheimer's Disease Prediction. *J. Pers. Med.* **2021**, *11*, 902. [CrossRef]
46. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 590–597. [CrossRef]
47. Alshmrani, G.M.M.; Ni, Q.; Jiang, R.; Pervaiz, H.; Elshennawy, N.M. A deep learning architecture for multi-class lung diseases classification using chest X-ray (CXR) images. *Alex. Eng. J.* **2023**, *64*, 923–935. [CrossRef]
48. Bharati, S.; Podder, P.; Mondal, M.R.H. Hybrid deep learning for detecting lung diseases from X-ray images. *Inform. Med. Unlocked* **2020**, *20*, 100391. [CrossRef]
49. Al-Sheikh, M.H.; Al Dandan, O.; Al-Shamayleh, A.S.; Jalab, H.A.; Ibrahim, R.W. Multi-class deep learning architecture for classifying lung diseases from chest X-Ray and CT images. *Sci. Rep.* **2023**, *13*, 19373. [CrossRef]
50. Kim, S.; Rim, B.; Choi, S.; Lee, A.; Min, S.; Hong, M. Deep learning in multi-class lung diseases' classification on chest X-ray images. *Diagnostics* **2022**, *12*, 915. [CrossRef]
51. Ibrahim, D.M.; Elshennawy, N.M.; Sarhan, A.M. Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases. *Comput. Biol. Med.* **2021**, *132*, 104348. [CrossRef]
52. Bhosale, Y.H.; Patnaik, K.S. PulDi-COVID: Chronic obstructive pulmonary (lung) diseases with COVID-19 classification using ensemble deep convolutional neural network from chest X-ray images to minimize severity and mortality rates. *Biomed. Signal Process. Control* **2023**, *81*, 104445. [CrossRef]
53. Gabruseva, T.; Poplavskiy, D.; Kalinin, A. Deep Learning for Automatic Pneumonia Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 350–351.
54. Tilve, A.; Nayak, S.; Vernekar, S.; Turi, D.; Shetgaonkar, P.R.; Aswale, S. Pneumonia detection using deep learning approaches. In Proceedings of the 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 24–25 February 2020; pp. 1–8.
55. Szepesi, P.; Szilágyi, L. Detection of pneumonia using convolutional neural networks and deep learning. *Biocybern. Biomed. Eng.* **2022**, *42*, 1012–1022. [CrossRef]
56. Goyal, S.; Singh, R. Detection and classification of lung diseases for pneumonia and Covid-19 using machine and deep learning techniques. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *14*, 3239–3259. [CrossRef]
57. Spanhol, F.A.; Oliveira, L.S.; Petitjean, C.; Heutte, L. A Dataset for Breast Cancer Histopathological Image Classification. *IEEE Trans. Biomed. Eng.* **2016**, *63*, 1455–1462. [CrossRef]
58. Liu, M.; Hu, L.; Tang, Y.; Wang, C.; He, Y.; Zeng, C.; Lin, K.; He, Z.; Huo, W. A deep learning method for breast cancer classification in the pathology images. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 5025–5032. [CrossRef] [PubMed]
59. Golatkar, A.; Anand, D.; Sethi, A. Classification of Breast Cancer Histology Using Deep Learning. In *Image Analysis and Recognition*; Springer International Publishing: Cham, Switzerland, 2018; pp. 837–844.
60. Shahidi, F.; Daud, S.M.; Abas, H.; Noor Azurati Ahmad, N.M. Breast Cancer Classification Using Deep Learning Approaches and Histopathology Image: A Comparison Study. *IEEE Access* **2020**, *8*, 187531–187552. [CrossRef]
61. Coudray, N.; Ocampo, P.S.; Sakellaropoulos, T.; Narula, N.; Snuderl, M.; Fenyö, D.; Moreira, A.L.; Razavian, N.; Tsirigos, A. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* **2018**, *24*, 1559–1567. [CrossRef] [PubMed]
62. Wang, L. Deep Learning Techniques to Diagnose Lung Cancer. *Cancers* **2022**, *14*, 5569. [CrossRef] [PubMed]
63. Kriegsmann, M.; Haag, C.; Weis, C.A.; Steinbuss, G.; Warth, A.; Zgorzelski, C.; Muley, T.; Winter, H.; Eichhorn, M.; Eichhorn, F.; et al. Deep Learning for the Classification of Small-Cell and Non-Small-Cell Lung Cancer. *Cancers* **2020**, *12*, 1604. [CrossRef]
64. Asuntha, A.; Srinivasan, A. Deep learning for lung Cancer detection and classification. *Multimed. Tools Appl.* **2020**, *79*, 7731–7762. [CrossRef]
65. Kanavati, F.; Toyokawa, G.; Momosaki, S.; Rambeau, M.; Kozuma, Y.; Shoji, F.; Yamazaki, K.; Takeo, S.; Iizuka, O.; Tsuneki, M. Weakly-supervised learning for lung carcinoma classification using deep learning. *Sci. Rep.* **2020**, *10*, 9297. [CrossRef]

66. Lee, J.H.; Ha, E.J.; Kim, J.H. Application of deep learning to the diagnosis of cervical lymph node metastasis from thyroid cancer with CT. *Eur. Radiol.* **2019**, *29*, 5452–5457. [CrossRef]
67. Wang, D.; Khosla, A.; Gargeya, R.; Irshad, H.; Beck, A.H. Deep Learning for Identifying Metastatic Breast Cancer. *arXiv* **2016**, arXiv:1606.05718. [CrossRef]
68. Amerikanos, P.; Maglogiannis, I. Image Analysis in Digital Pathology Utilizing Machine Learning and Deep Neural Networks. *J. Pers. Med.* **2022**, *12*, 1444. [CrossRef]
69. Hussain, S.I.; Toscano, E. Optimized deep learning for mammography: Augmentation and tailored architectures. *Information* **2025**, *16*, 359. [CrossRef]
70. Pereira, S.; Pinto, A.; Alves, V.; Silva, C.A. Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. *IEEE Trans. Med. Imaging* **2016**, *35*, 1240–1251. [CrossRef] [PubMed]
71. Isensee, F.; Petersen, J.; Klein, A.; Zimmerer, D.; Jaeger, P.F.; Kohl, S.; Wasserthal, J.; Koehler, G.; Norajitra, T.; Wirkert, S.; et al. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. *arXiv* **2019**, arXiv:1809.10486.
72. Vafaeikia, P.; Wagner, M.W.; Hawkins, C.; Tabori, U.; Ertl-Wagner, B.B.; Khalvati, F. MRI-Based End-To-End Pediatric Low-Grade Glioma Segmentation and Classification. *Can. Assoc. Radiol. J.* **2023**, *75*, 153–160. [CrossRef] [PubMed]
73. Naser, M.A.; Deen, M.J. Brain tumor segmentation and grading of lower-grade glioma using deep learning in MRI images. *Comput. Biol. Med.* **2020**, *121*, 103758. [CrossRef]
74. Cheng, J.; Liu, J.; Kuang, H.; Wang, J. A Fully Automated Multimodal MRI-Based Multi-Task Learning for Glioma Segmentation and IDH Genotyping. *IEEE Trans. Med. Imaging* **2022**, *41*, 1520–1532. [CrossRef]
75. Pemberton, H.G.; Wu, J.; Kommers, I.; Müller, D.M.J.; Hu, Y.; Goodkin, O.; Vos, S.B.; Bisdas, S.; Robe, P.A.; Ardon, H.; et al. Multi-class glioma segmentation on real-world data with missing MRI sequences: Comparison of three deep learning algorithms. *Sci. Rep.* **2023**, *13*, 18911. [CrossRef]
76. Fateh, A.; Rezvani, Y.; Moayedi, S.; Rezvani, S.; Fateh, F.; Fateh, M.; Abolghasemi, V. BRISC: Annotated Dataset for Brain Tumor Segmentation and Classification. *Sci. Data* **2026**. [CrossRef] [PubMed]
77. Li, F.; Chen, H.; Liu, Z.; Zhang, X.D.; Jiang, M.S.; Wu, Z.Z.; Zhou, K.Q. Deep learning-based automated detection of retinal diseases using optical coherence tomography images. *Biomed. Opt. Express* **2019**, *10*, 6204. [CrossRef] [PubMed]
78. Helaly, H.A.; Badawy, M.; Haikal, A.Y. Deep Learning Approach for Early Detection of Alzheimer's Disease. *Cogn. Comput.* **2021**, *14*, 1711–1727. [CrossRef] [PubMed]
79. Thangavel, P.; Natarajan, Y.; Sri Preethaa, K. EAD-DNN: Early Alzheimer's disease prediction using deep neural networks. *Biomed. Signal Process. Control* **2023**, *86*, 105215. [CrossRef]
80. Zhang, J.; Xie, Y.; Xia, Y.; Shen, C. Attention Residual Learning for Skin Lesion Classification. *IEEE Trans. Med. Imaging* **2019**, *38*, 2092–2103. [CrossRef]
81. Yap, J.; Yolland, W.; Tschandl, P. Multimodal skin lesion classification using deep learning. *Exp. Dermatol.* **2018**, *27*, 1261–1267. [CrossRef]
82. Nida, N.; Irtaza, A.; Javed, A.; Yousaf, M.H.; Mahmood, M.T. Melanoma lesion detection and segmentation using deep region based convolutional neural network and fuzzy C-means clustering. *Int. J. Med. Inform.* **2019**, *124*, 37–48. [CrossRef] [PubMed]
83. Ahmed, N.; Tan, X.; Ma, L. A new method proposed to Melanoma-skin cancer lesion detection and segmentation based on hybrid convolutional neural network. *Multimed. Tools Appl.* **2022**, *82*, 11873–11896. [CrossRef]
84. Adegun, A.A.; Viriri, S. Deep learning-based system for automatic melanoma detection. *IEEE Access* **2019**, *8*, 7160–7172. [CrossRef]
85. Thepade, S.D.; Shukla, S. Enhancing melanoma skin cancer detection through feature fusion of pre-trained deep convolutional neural network ResNet50 and Thepade sorted block truncation coding. *SN Comput. Sci.* **2024**, *5*, 426. [CrossRef]
86. Safi, H.; Safi, S.; Hafezi-Moghadam, A.; Ahmadieh, H. Early detection of diabetic retinopathy. *Surv. Ophthalmol.* **2018**, *63*, 601–608. [CrossRef] [PubMed]
87. Pires, R.; Avila, S.; Wainer, J.; Valle, E.; Abramoff, M.D.; Rocha, A. A data-driven approach to referable diabetic retinopathy detection. *Artif. Intell. Med.* **2019**, *96*, 93–106. [CrossRef] [PubMed]
88. Fu, H.; Cheng, J.; Xu, Y.; Zhang, C.; Wong, D.W.K.; Liu, J.; Cao, X. Disc-Aware Ensemble Network for Glaucoma Screening From Fundus Image. *IEEE Trans. Med. Imaging* **2018**, *37*, 2493–2501. [CrossRef]
89. Ting, D.S.W.; Cheung, C.Y.L.; Lim, G.; Tan, G.S.W.; Quang, N.D.; Gan, A.; Hamzah, H.; Garcia-Franco, R.; San Yeo, I.Y.; Lee, S.Y.; et al. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA* **2017**, *318*, 2211. [CrossRef]
90. Bercea, C.I.; Puyol-Antón, E.; Wiestler, B.; Rueckert, D.; Schnabel, J.A.; King, A.P. Bias in Unsupervised Anomaly Detection in Brain MRI. In *Clinical Image Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging*; Springer Nature: Cham, Switzerland, 2023; pp. 122–131. [CrossRef]
91. Chan, H.; K.Samala, R.; Hadjiiski, L.M.; Zhou, C. Deep Learning in Medical Image Analysis. *Adv. Exp. Med. Biol.* **2020**, *1213*, 3–21. [CrossRef]

92. Lee, T.; Puyol-Antón, E.; Ruijsink, B.; Aitcheson, K.; Shi, M.; King, A.P. An Investigation into the Impact of Deep Learning Model Choice on Sex and Race Bias in Cardiac MR Segmentation. In *Clinical Image-Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging*; Springer Nature: Cham, Switzerland, 2023; pp. 215–224. [CrossRef]
93. Smith-Bindman, R.; Kwan, M.L.; Marlow, E.C.; Theis, M.K.; Bolch, W.; Cheng, S.Y.; Bowles, E.J.A.; Duncan, J.R.; Greenlee, R.T.; Kushi, L.H.; et al. Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000–2016. *JAMA* **2019**, *322*, 843. [CrossRef]
94. Barbara, K.; Stuart, C. *Procedures for Undertaking Systematic Reviews*; Technical Report, EBSE Technical Report EBSE-2007-01; Empirical Software Engineering National ICT Australia Ltd.: Eveleigh, Australia, 2004.
95. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
96. Lin, R.; Gholipour, A.; Thiran, J.P.; Karimi, D.; Kebiri, H.; Cuadra, M.B. Cross-age and cross-site domain shift impacts on deep learning-based white matter fiber estimation in newborn and baby brains. In Proceedings of the 2024 IEEE International Symposium on Biomedical Imaging (ISBI), Athens, Greece, 27–30 May 2024; pp. 1–5.
97. Li, J.L.; Hsu, C.F.; Chang, M.C.; Chen, W.C. A Comprehensive Review of Machine Learning Advances on Data Change: A Cross-Field Perspective. *arXiv* **2024**, arXiv:2402.12627. [CrossRef]
98. Zhu, Y.; Tang, Y.; Tang, Y.; Elton, D.C.; Lee, S.; Pickhardt, P.J.; Summers, R.M., Cross-domain Medical Image Translation by Shared Latent Gaussian Mixture Model. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020*; Springer International Publishing: Cham, Switzerland, 2020; pp. 379–389. [CrossRef]
99. Banerjee, I.; Bhattacharjee, K.; Burns, J.L.; Trivedi, H.; Purkayastha, S.; Seyyed-Kalantari, L.; Patel, B.N.; Shiradkar, R.; Gichoya, J. "Shortcuts" Causing Bias in Radiology Artificial Intelligence: Causes, Evaluation, and Mitigation. *J. Am. Coll. Radiol.* **2023**, *20*, 842–851. [CrossRef]
100. Krois, J.; Garcia Cantu, A.; Chaurasia, A.; Patil, R.; Chaudhari, P.K.; Gaudin, R.; Gehrung, S.; Schwendicke, F. Generalizability of deep learning models for dental image analysis. *Sci. Rep.* **2021**, *11*, 6102. [CrossRef]
101. Su, Z.; Guo, J.; Yang, X.; Wang, Q.; Coenen, F.; Huang, K. Navigating distribution shifts in medical image analysis: A survey. *arXiv* **2024**, arXiv:2411.05824.
102. Zhang, L.; Wang, X.; Yang, D.; Sanford, T.; Harmon, S.; Turkbey, B.; Wood, B.J.; Roth, H.; Myronenko, A.; Xu, D.; et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Trans. Med. Imaging* **2020**, *39*, 2531–2540. [CrossRef]
103. Thiam, P.; Lausser, L.; Kloth, C.; Blaich, D.; Liebold, A.; Beer, M.; Kestler, H.A. Unsupervised domain adaptation for the detection of cardiomegaly in cross-domain chest X-ray images. *Front. Artif. Intell.* **2023**, *6*, 1056422. [CrossRef]
104. Lafarge, M.W.; Pluim, J.P.W.; Eppenhof, K.A.J.; Moeskops, P.; Veta, M. Domain-Adversarial Neural Networks to Address the Appearance Variability of Histopathology Images. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer International Publishing: Cham, Switzerland, 2017; pp. 83–91. [CrossRef]
105. Liu, W.; Ni, Z.; Chen, Q.; Ni, L. Attention Guided Partial Domain Adaptation for Automated Pneumonia Diagnosis From Chest X-Ray Images. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 5848–5859. [CrossRef] [PubMed]
106. Ghimire, S.; Kashyap, S.; Wu, J.T.; Karargyris, A.; Moradi, M. Learning Invariant Feature Representation to Improve Generalization Across Chest X-Ray Datasets. In *Machine Learning in Medical Imaging*; Springer International Publishing: Cham, Switzerland, 2020; pp. 644–653. [CrossRef]
107. Chamveha, I.; Tongdee, T.; Saiviroonporn, P.; Chaisangmongkon, W. Local Adaptation Improves Accuracy of Deep Learning Model for Automated X-Ray Thoracic Disease Detection: A Thai Study. *arXiv* **2020**, arXiv: 2004.10975.
108. Kordnoori, S.; Sabeti, M.; Mostafaei, H.; Seyed, S.; Banihashemi, A. Advancing COVID-19 Infection Diagnosis: Integrating Segmentation and Classification via Deep Multi-Task Learning Model for Lung CT Scan Images. In Proceedings of the AI SOFT, Macao, China, 19–25 August 2023.
109. Blaivas, M.; Blaivas, L.N.; Tsung, J.W. Deep Learning Pitfall: Impact of Novel Ultrasound Equipment Introduction on Algorithm Performance and the Realities of Domain Adaptation. *J. Ultrasound Med.* **2021**, *41*, 855–863. [CrossRef] [PubMed]
110. Brown, A.; Tomasev, N.; Freyberg, J.; Liu, Y.; Karthikesalingam, A.; Schrouff, J. Detecting shortcut learning for fair medical AI using shortcut testing. *Nat. Commun.* **2023**, *14*, 4314. [CrossRef]
111. Chen, X.; Wang, T.; Zhou, J.; Song, Z.; Gao, X.; Zhang, X. Evaluating and mitigating bias in AI-based medical text generation. *Nat. Comput. Sci.* **2025**, *5*, 388–396. [CrossRef]
112. Graham, S.; Minhas, F.; Bilal, M.; Ali, M.; Tsang, Y.W.; Eastwood, M.; Wahab, N.; Jahanifar, M.; Hero, E.; Dodd, K.; et al. Screening of normal endoscopic large bowel biopsies with interpretable graph learning: A retrospective study. *Gut* **2023**, *72*, 1709–1721. [CrossRef] [PubMed]
113. Oliveira, H.N.; Ferreira, E.; Santos, J.A.D. Truly Generalizable Radiograph Segmentation With Conditional Domain Adaptation. *IEEE Access* **2020**, *8*, 84037–84062. [CrossRef]

114. He, B.; Chen, Y.; Zhu, D.; Xu, Z. Domain adaptation via Wasserstein distance and discrepancy metric for chest X-ray image classification. *Sci. Rep.* **2024**, *14*, 2690. [CrossRef] [PubMed]
115. Chen, C.; Qin, C.; Qiu, H.; Ouyang, C.; Wang, S.; Chen, L.; Tarroni, G.; Bai, W.; Rueckert, D. Realistic Adversarial Data Augmentation for MR Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020*; Springer International Publishing: Cham, Switzerland, 2020; pp. 667–677. [CrossRef]
116. Li, Y.; He, N.; Huang, Y. Single Domain Generalization via Spontaneous Amplitude Spectrum Diversification. In *Resource-Efficient Medical Image Analysis*; Springer Nature: Cham, Switzerland, 2022; pp. 32–41. [CrossRef]
117. Zhang, R.; Xu, Q.; Huang, C.; Zhang, Y.; Wang, Y. Semi-Supervised Domain Generalization for Medical Image Analysis. In Proceedings of the 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), Kolkata, India, 28–31 March 2022; pp. 1–5. [CrossRef]
118. Garrucho, L.; Kushibar, K.; Jouide, S.; Diaz, O.; Igual, L.; Lekadir, K. Domain generalization in deep learning based mass detection in mammography: A large-scale multi-center study. *Artif. Intell. Med.* **2022**, *132*, 102386. [CrossRef] [PubMed]
119. Xu, Z.; Liu, D.; Yang, J.; Raffel, C.; Niethammer, M. Robust and Generalizable Visual Representation Learning via Random Convolutions. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
120. Wang, H.; Xia, Y. Domain-ensemble learning with cross-domain mixup for thoracic disease classification in unseen domains. *Biomed. Signal Process. Control* **2023**, *81*, 104488. [CrossRef]
121. Zhou, W.; Guan, G.; Yi, Y.; Cui, W.; Chen, Y. Progressive pseudo-labels enhancement for source-free domain adaptation medical image segmentation. *Biomed. Signal Process. Control* **2025**, *109*, 108053. [CrossRef]
122. Ilyas, T.; Ahmad, K.; Arsa, D.M.S.; Jeong, Y.C.; Kim, H. Enhancing medical image analysis with unsupervised domain adaptation approach across microscopes and magnifications. *Comput. Biol. Med.* **2024**, *170*, 108055. [CrossRef]
123. Chen, C.; Dou, Q.; Chen, H.; Heng, P.A. Semantic-Aware Generative Adversarial Nets for Unsupervised Domain Adaptation in Chest X-Ray Segmentation. In *Machine Learning in Medical Imaging*; Springer International Publishing: Cham, Switzerland, 2018; pp. 143–151. [CrossRef]
124. Tang, Y.; Tang, Y.; Sandfort, V.; Xiao, J.; Summers, R.M., TUNA-Net: Task-Oriented UNsupervised Adversarial Network for Disease Recognition in Cross-domain Chest X-rays. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*; Springer International Publishing: Cham, Switzerland, 2019; pp. 431–440. [CrossRef]
125. Li, X.; Desrosiers, C.; Liu, X. Deep Neural Forest for Out-of-Distribution Detection of Skin Lesion Images. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 157–165. [CrossRef]
126. Oliveira, H.; Mota, V.; Machado, A.M.; dos Santos, J.A. From 3D to 2D: Transferring knowledge for rib segmentation in chest X-rays. *Pattern Recognit. Lett.* **2020**, *140*, 10–17. [CrossRef]
127. Utama, N.P.; Muzakki, M.F. Utilizing Generative Adversarial Network for Synthetic Image Generation to Address Imbalance Challenges in Chest X-Ray Image Classification. *J. ICT Res. Appl.* **2023**, *17*, 373. [CrossRef]
128. Nguyen, P.; Rathod, A.; Chapman, D.; Prathapan, S.; Menon, S.; Morris, M.; Yesha, Y. Active Semi-Supervised Learning via Bayesian Experimental Design for Lung Cancer Classification Using Low Dose Computed Tomography Scans. *Appl. Sci.* **2023**, *13*, 3752. [CrossRef]
129. Hermoza, R.; Maicas, G.; Nascimento, J.C.; Carneiro, G. Censor-Aware Semi-supervised Learning for Survival Time Prediction from Medical Images. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022*; Springer Nature: Cham, Switzerland, 2022; pp. 213–222. [CrossRef]
130. Bai, T.; Zhang, Z.; Zhao, C.; Luo, X. A Novel Pseudo-Labeling Approach for Cell Detection Based on Adaptive Threshold. In *Bioinformatics Research and Applications*; Springer International Publishing: Cham, Switzerland, 2021; pp. 254–265. [CrossRef]
131. Qiu, L.; Zhao, L.; Hou, R.; Zhao, W.; Zhang, S.; Lin, Z.; Teng, H.; Zhao, J. Hierarchical multimodal fusion framework based on noisy label learning and attention mechanism for cancer classification with pathology and genomic features. *Comput. Med. Imaging Graph.* **2023**, *104*, 102176. [CrossRef]
132. Zhu, J.; Bolsterlee, B.; Song, Y.; Meijering, E. Improving cross-domain generalizability of medical image segmentation using uncertainty and shape-aware continual test-time domain adaptation. *Med. Image Anal.* **2025**, *101*, 103422. [CrossRef]
133. Nandy, J.; Hs, W.; Le, M.L. Distributional Shifts In Automated Diabetic Retinopathy Screening. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 255–259. [CrossRef]
134. Pacheco, A.G.C.; Sastry, C.S.; Trappenberg, T.; Oore, S.; Krohling, R.A. On Out-of-Distribution Detection Algorithms with Deep Neural Skin Cancer Classifiers. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 3152–3161. [CrossRef]
135. Calli, E.; Ginneken, B.V.; Sogancioglu, E.; Murphy, K. FRODO: An In-Depth Analysis of a System to Reject Outlier Samples From a Trained Neural Network. *IEEE Trans. Med. Imaging* **2023**, *42*, 971–981. [CrossRef] [PubMed]
136. Zuyu, Z.; Yan, L.; Byeong-Seok, S. C2-GAN: Content-consistent generative adversarial networks for unsupervised domain adaptation in medical image segmentation. *Med. Phys.* **2022**, *49*, 6491–6504. [CrossRef]

137. Chen, Y.; Wang, D.; Zhu, D.; Xu, Z.; He, B. Unsupervised domain adaptation of dynamic extension networks based on class decision boundaries. *Multimed. Syst.* **2024**, *30*, 80. [CrossRef]
138. Ma, C.; Ji, Z.; Gao, M. Neural Style Transfer Improves 3D Cardiovascular MR Image Segmentation on Inconsistent Data. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*; Springer International Publishing: Cham, Switzerland, 2019; pp. 128–136. [CrossRef]
139. Angelakis, A.; Rass, A. A data-centric approach to class-specific bias in image data augmentation. *arXiv* **2024**, arXiv:2403.04120.
140. Kim, H.; Lee, S.; Shim, W.J.; Choi, M.S.; Cho, S. Homogenization of multi-institutional chest X-ray images in various data transformation schemes. *J. Med. Imaging* **2023**, *10*, 061103. [CrossRef]
141. Kousiga, T.; Nithya, P. An Improving Lung Disease Detection by Combining Ensemble Deep Learning and Maximum Mean Discrepancy Transfer Learning. *Int. J. Intell. Eng. Syst.* **2024**, *17*, 294–306. [CrossRef]
142. Lee, Y.W.; Huang, S.K.; Chang, R.F. CheXGAT: A disease correlation-aware network for thorax disease diagnosis from chest X-ray images. *Artif. Intell. Med.* **2022**, *132*, 102382. [CrossRef] [PubMed]
143. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
144. Das, D.; Santosh, K.; Pal, U. Cross-population train/test deep learning model: Abnormality screening in chest X-rays. In Proceedings of the 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Rochester, MN, USA, 28–30 July 22020; pp. 514–519.
145. Ghafoorian, M.; Mehrtash, A.; Kapur, T.; Karssemeijer, N.; Marchiori, E.; Pesteie, M.; Guttmann, C.R.; De Leeuw, F.E.; Tempany, C.M.; Van Ginneken, B.; et al. Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2017; pp. 516–524.
146. Ren, J.; Hacihaliloglu, I.; Singer, E.A.; Foran, D.J.; Qi, X. Adversarial Domain Adaptation for Classification of Prostate Histopathology Whole-Slide Images. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*; Springer International Publishing: Cham, Switzerland, 2018; pp. 201–209. [CrossRef]
147. Mahmood, F.; Chen, R.; Durr, N.J. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. *IEEE Trans. Med. Imaging* **2018**, *37*, 2572–2581. [CrossRef]
148. Valliani, A.A.; Gulamali, F.F.; Kwon, Y.J.; Martini, M.L.; Wang, C.; Kondziolka, D.; Chen, V.J.; Wang, W.; Costa, A.B.; Oermann, E.K. Deploying deep learning models on unseen medical imaging using adversarial domain adaptation. *PLoS ONE* **2022**, *17*, e0273262. [CrossRef]
149. Zou, L. Meta-learning basics and background. In *Meta-Learning*; Elsevier: Amsterdam, The Netherlands, 2023; pp. 1–22. [CrossRef]
150. Tayebi Arasteh, S.; Kuhl, C.; Saehn, M.J.; Isfort, P.; Truhn, D.; Nebelung, S. Enhancing domain generalization in the AI-based analysis of chest radiographs with federated learning. *Sci. Rep.* **2023**, *13*, 22576. [CrossRef] [PubMed]
151. Musa, A.; Prasad, R.; Hassan, M.; Hamada, M.; Ilu, S.Y. FairCXRnet: A Multi-Task Learning Model for Domain Adaptation in Chest X-Ray Classification for Low Resource Settings. *Eng. Proc.* **2025**, *107*, 16. [CrossRef]
152. Cai, J.; Li, H.; Tan, M.; He, B.; Li, H. Cross-modal generalizable medical image segmentation with dual-domain deformable transformer and multitask adaptation. *Expert Syst. Appl.* **2025**, *277*, 127249. [CrossRef]
153. Chen, G.Y.; Lin, C.T. Multi-task supervised contrastive learning for chest X-ray diagnosis: A two-stage hierarchical classification framework for COVID-19 diagnosis. *Appl. Soft Comput.* **2024**, *155*, 111478. [CrossRef]
154. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
155. Johnson, A.E.; Pollard, T.J.; Berkowitz, S.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.y.; Mark, R.G.; Horng, S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **2019**, *6*, 317. [CrossRef]
156. Bustos, A.; Pertusa, A.; Salinas, J.M.; De La Iglesia-Vaya, M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Med. Image Anal.* **2020**, *66*, 101797. [CrossRef]
157. Nguyen, H.Q.; Lam, K.; Le, L.T.; Pham, H.H.; Tran, D.Q.; Nguyen, D.B.; Le, D.D.; Pham, C.M.; Tong, H.T.; Dinh, D.H.; et al. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Sci. Data* **2022**, *9*, 429. [CrossRef]
158. Bahamazava, K.; O'Reilly, R. FAIR-MED: Bias Detection and Fairness Evaluation in Healthcare Focused XAI. In *World Conference on Explainable Artificial Intelligence*; Springer: Cham, Switzerland, 2025; pp. 380–401.
159. Mannan, S.; Shekar, B.H.; Taljeh, M.Z. PulmoMTL-Net: A multitask learning framework for enhanced segmentation and classification of pulmonary diseases. *Int. J. Comput. Appl.* **2025**, *47*, 611–630. [CrossRef]
160. Lai, H.; Luo, Y.; Li, B.; Lu, J.; Yuan, J. Bilateral proxy federated domain generalization for privacy-preserving medical image diagnosis. *IEEE J. Biomed. Health Inform.* **2024**, *29*, 2784–2797. [CrossRef]

161. Sun, Y.; Chong, N.; Ochiai, H. Feature distribution matching for federated domain generalization. In *Asian Conference on Machine Learning*; PMLR: London, UK, 2023; pp. 942–957.
162. Li, Y.; Wang, X.; Zeng, R.; Donta, P.K.; Murturi, I.; Huang, M.; Dustdar, S. Federated domain generalization: A survey. *Proc. IEEE* **2025**, *113*, 370–410. [CrossRef]
163. Tran, A.T.; Zeevi, T.; Payabvash, S. Strategies to improve the robustness and generalizability of deep learning segmentation and classification in neuroimaging. *BioMedInformatics* **2025**, *5*, 20. [CrossRef]
164. Mehta, R.; Shui, C.; Arbel, T. Evaluating the fairness of deep learning uncertainty estimates in medical image analysis. In *Proceedings of the Medical Imaging with Deep Learning*; PMLR: London, UK, 2024; pp. 1453–1492.
165. Tian, Y.; Wen, C.; Shi, M.; Afzal, M.M.; Huang, H.; Khan, M.O.; Luo, Y.; Fang, Y.; Wang, M. Fairdomain: Achieving fairness in cross-domain medical image segmentation and classification. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2024; pp. 251–271.
166. Sufian, M.A.; Alsadder, L.; Hamzi, W.; Zaman, S.; Sagar, A.S.; Hamzi, B. Mitigating Algorithmic Bias in AI-Driven Cardiovascular Imaging for Fairer Diagnostics. *Diagnostics* **2024**, *14*, 2675. [CrossRef]
167. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–35. [CrossRef]
168. Rajkomar, A.; Hardt, M.; Howell, M.D.; Corrado, G.; Chin, M.H. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med.* **2018**, *169*, 866–872. [CrossRef] [PubMed]
169. Corbett-Davies, S.; Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv* **2018**, arXiv:1808.00023.
170. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **2019**, *366*, 447–453. [CrossRef]
171. Ahmad, M.A.; Patel, A.; Eckert, C.; Kumar, V.; Teredesai, A. Fairness in machine learning for healthcare. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual, 6–10 July 2020; pp. 3529–3530.
172. Wang, R.; Kuo, P.C.; Chen, L.C.; Seastedt, K.P.; Gichoya, J.W.; Celi, L.A. Drop the shortcuts: Image augmentation improves fairness and decreases AI detection of race and other demographics from medical images. *EBioMedicine* **2024**, *102*, 105047. [CrossRef] [PubMed]
173. Yang, Y.; Zhang, H.; Gichoya, J.W.; Katabi, D.; Ghassemi, M. The limits of fair medical imaging AI in real-world generalization. *Nat. Med.* **2024**, *30*, 2838–2848. [CrossRef] [PubMed]
174. Wang, X.; Zhang, J.; Yang, S.; Xiang, J.; Luo, F.; Wang, M.; Zhang, J.; Yang, W.; Huang, J.; Han, X. A generalizable and robust deep learning algorithm for mitosis detection in multicenter breast histopathological images. *Med. Image Anal.* **2023**, *84*, 102703. [CrossRef] [PubMed]
175. Bashyam, V.M.; Doshi, J.; Erus, G.; Srinivasan, D.; Abdulkadir, A.; Singh, A.; Habes, M.; Fan, Y.; Masters, C.L.; Maruff, P. Deep generative medical image harmonization for improving cross-site generalization in deep learning predictors. *J. Magn. Reson. Imaging* **2022**, *55*, 908–916. [CrossRef]