Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 9, No. 1, January 2027, pp: 156-180 e-ISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

A Multi-Source Harmonization Pipeline for Chest X-ray, Computed Tomography, and Brain MRI to Construct a Plausibility-Constrained Training Corpus for Medical Image Synthesis

Said Laharraf, and Khaoula Addakiri

Department of Biomedical Sciences and Engineering, Higher Institute of Health Sciences (ISSS), Hassan I University, Settat, Morocco

Corresponding author: Said Laharraf (laharraf.said4408.doc@uhp.ac.ma), **Author(s) Email:** Khaoula Addakiri (khaoula.addakiri@uhp.ac.ma).

Abstract

Background: Generative adversarial networks for medical image synthesis can produce images that look realistic yet are anatomically implausible. Constraining a generator against this failure requires a reference distribution of tissue composition, which in turn requires a harmonized, leakage-free, multi-modality training corpus that public archives do not provide, because each was released with its own format, intensity convention, labelling taxonomy and partitioning policy.

Aims: The objective of this study is to design, implement and validate the data-engineering stage of the IntrinsicX programme: a reproducible pipeline that converts five public archives (NIH ChestX-ray14, NIH DeepLesion, BraTS-Africa, LIDC-IDRI and OpenBHB) into one analysis-ready corpus with compositional reference statistics.

Methods: The pipeline restores Hounsfield-Unit intensities and applies DICOM or lung windowing for computed tomography, normalizes magnetic resonance slices by percentile, extracts informative axial slices, resamples to 256x256, partitions strictly at the patient level and stores everything in one HDF5 schema; four-class k-means clustering of reference slices yields mean and covariance statistics. It produced 841,739 harmonized images (25.49 GB) from 57 acquisition sites, with realized partition ratios within 2.07 percentage points of the 70/15/15 target for the volumetric sources.

Results: Resampling preserved structural similarity of 0.9233 for LIDC-IDRI and 0.9923 and 0.9917 for the two brain sources, whereas percentile normalization saturated 4.872% and 2.992% of foreground pixels in BraTS-Africa and OpenBHB without moving the reference composition by more than 0.016. The four-cluster partition was reproducible (Adjusted Rand Index 0.988 and 0.910). The scientific contribution is an auditable, regenerable corpus whose limits are measured and reported; its practical value is a ready training substrate that spares other groups the preprocessing.

Conclusion: This study evaluates the corpus only: the effect on generative synthesis is untested and reserved for the companion IntrinsicX-GAN study.

Keywords Medical image harmonization; Data preprocessing pipeline; Generative adversarial networks; Medical image synthesis; multi-site neuroimaging; Reproducible research.

I. Introduction

Deep generative modelling is now central to medical imaging research. Generative adversarial networks (GANs) [1] and the conditional and cycle-consistent translation frameworks derived from them [2], [3] are applied across the field [4], [5]: augmentation for rare pathologies [6], cross-modality translation and contrast synthesis [7], [8], [9], and privacy-preserving surrogate release [10]. Denoising diffusion models [11], [12] have since been brought to the same problems, including whole-brain volumetric generation [13], and support a literature of their own [14]; comparative evaluations find the two families competitive on medical synthesis [15], and both are advanced as routes to shareable surrogate data [16]. In every case, the value of the output turns on a property easy to state and hard to enforce: the image must be not merely realistic in texture but physiologically admissible in composition. A synthetic radiograph whose soft-tissue, bone and air fractions depart from any real anatomy can still pass a discriminator trained on appearance, and is worse than useless once it is used to train or benchmark a clinical

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

model. Recent work therefore constrains synthesis with explicit anatomical or physiological priors rather than adversarial realism alone [17], a need underlined by evidence that strong distributional scores such as the Frechet Inception Distance [18] do not guarantee anatomically usable output [19], [20]. A tractable formulation is a plausibility loss: a penalty on the distance between the compositional statistics of a generated image and a reference distribution estimated from real anatomy of the same modality. It is model-agnostic and cheap to evaluate, but it shifts the burden onto the data. The reference distribution must come from a corpus large enough to be representative, restricted to anatomy free of the pathology to be constrained against, harmonized in intensity and geometry so that the statistic is comparable across images, and partitioned at the patient level so that estimation and downstream evaluation cannot leak into one another.

No existing public resource satisfies these requirements at once. Large single-modality archives are well curated in isolation: NIH ChestX-ray14 [21] for thoracic disease classification, NIH DeepLesion [22] for lesion detection, BraTS-Africa [23] for glioma segmentation in an underrepresented population, LIDC-IDRI [24], distributed through The Cancer Imaging Archive [25], for pulmonary nodule analysis, and OpenBHB [26] for multi-site brain-age prediction. Each was released with a different native format, intensity convention, labelling taxonomy and patient-splitting policy, or none. Aggregating them is not concatenation: it requires per-source decoding, physically correct intensity restoration, geometric resampling, and one partitioning discipline applied uniformly. Aggregated collections such as MedMNIST v2 [27] and RadImageNet [28] serve adjacent purposes, lightweight benchmarking and transfer-learning pretraining, but neither retains the resolution synthesis needs nor supplies reference statistics. Other chest archives, among them CheXpert [29], MIMIC-CXR [30] and PadChest [31], match or exceed ChestX-ray14 in scale but are single-modality and carry no cross-modality harmonization.

Existing harmonization techniques address a different problem. Empirical Bayes batch correction [32] and its neuroimaging adaptations [33], [34] adjust site and scanner effects in derived features and require site labels as input, and learned variants pursue the same goal through site-invariant representations [35], [36]; none of them produces an image corpus. General-purpose frameworks supply operations rather than a specification: nnU-Net configures preprocessing automatically from a dataset fingerprint for segmentation [37], while MONAI [38] and TorchIO [39] provide composable transforms for training pipelines. These tools leave the corpus-level decisions to the user, namely the order in which a storage offset, a display window and a normalization are applied to each archive, one patient-level partitioning discipline across archives, and the estimation of a compositional reference statistic. The IntrinsicX pipeline differs in being a fixed, source-specific and audited specification whose output is a released corpus, and whose partition integrity and fidelity cost are measured and reported rather than assumed.

That engineering is substantial, error-prone, and rarely reported in enough detail to be reproduced. It is also where most defects that later contaminate a generative study are introduced: an unreversed storage offset silently shifts an entire computed tomography cohort, a slice-level split leaks a patient across partitions, and normalization applied per volume rather than per slice destroys the comparability of any compositional statistic computed downstream. Treating the stage as an unreported preliminary is therefore not defensible. Systematic reviews trace much irreproducible or over-optimistic clinical machine learning to exactly these preprocessing and partitioning decisions [40], [41], and leakage across partitions is among the best-characterised failure modes in the wider literature [42].

The research gap can therefore be stated as four conditions that a corpus must meet at once to support plausibility-constrained synthesis. It must keep a spatial resolution at which tissue composition is measurable, which rules out the 28x28 images of MedMNIST v2 [27]. It must place several modalities on one grid and one intensity range, which single-modality archives [21], [26], [29]-[31] cannot do by construction. It must be partitioned at the patient level with ratios that are verified rather than nominal, whereas the source archives apply different partitioning policies or none. And it must ship a reference statistic estimated from reference anatomy, which none of the resources compared in Section IV-A provides. To our knowledge, no public corpus meets all four conditions, and the pathology-oriented archives that dominate the literature [22]-[24] were not designed to.

The objective of this study is therefore to design, implement and empirically validate the IntrinsicX harmonization pipeline, a documented procedure that converts five heterogeneous public archives into one analysis-ready corpus with its plausibility reference statistics, and to measure what that conversion costs in image fidelity. The proposed solution is a modality-aware procedure specified by 21 numbered equations and one algorithm (Section II), and its contributions are stated so that each can be checked against a number: 1) a pipeline specified in full, including the Hounsfield-Unit offset reversal and per-image DICOM windowing applied to computed tomography, the per-slice percentile normalization applied to magnetic resonance imaging, and the occupancy-based informative-slice criterion applied to all volumetric

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

sources; 2) a single HDF5 schema accommodating single-channel and four-channel imaging, per-source label taxonomies, patient identifiers and acquisition-site tags without loss of source-specific information; 3) patient-level partitioning enforced and verified against the written files rather than assumed, with realized image-level ratios reported for every source and a largest deviation of 2.07 percentage points from the 70/15/15 target among the volumetric sources; 4) a procedure for estimating compositional reference statistics by unsupervised four-class intensity clustering, executed for the two modalities in which a reference subset is available, with subject-level bootstrap intervals and an Adjusted Rand Index of 0.988 and 0.910 for the partition; 5) a quantified fidelity cost, namely a structural similarity of 0.9233 to 0.9923 after resampling and saturation of 4.872 and 2.992 per cent of foreground pixels in BraTS-Africa and OpenBHB; and 6) a publicly deposited corpus of 841,739 harmonized images occupying 25.49 GB from 57 acquisition sites, exactly reproducible from the original public sources. The design follows the FAIR principles [43] and published guidance on preparing medical imaging data for machine learning [44].

The scope is deliberately bounded. This article reports the acquisition, harmonization and reference-estimation stage of the IntrinsicX programme and validates it on its own terms: internal consistency, partition integrity, site coverage, and the statistical well-formedness of the derived reference distributions. The plausibility-constrained model that consumes the corpus, IntrinsicX-GAN, with its training protocol and comparative evaluation against unconstrained baselines, is the subject of a companion article in preparation. The separation is methodological rather than a split result: the pipeline and corpus are self-contained, independently reusable by groups with no interest in our generator, and require validation criteria unrelated to synthesis quality. Every result reported here therefore concerns the properties of the corpus, and any statement about synthesis quality is deferred to that companion article.

Section II specifies the pipeline architecture, the source-specific processing paths, the partitioning and storage schema, and the reference-estimation procedure together with the evaluation metrics and the complete algorithm. Section III reports the measured properties of the corpus. Section IV positions it against seven published resources on measurable indicators, discusses its implications and reusability, and states its limitations. Section V concludes and outlines the modelling work that follows.

II. Method

A. Pipeline Architecture and Design Requirements

The pipeline was designed against four requirements derived from the intended downstream use. First, physical correctness: intensity values must be restored to their physically meaningful scale before any windowing or normalization is applied, since an offset or rescale applied out of order is not recoverable afterwards. Second, geometric uniformity: all outputs must share a single spatial grid so that a compositional statistic computed on one image is comparable to the same statistic computed on another. Third, partition integrity: no patient may contribute images to more than one partition, in any source. Fourth, reproducibility: every parameter, seed, and exclusion decision must be recorded so that the corpus can be regenerated from the original public archives without reference to any intermediate artefact held by the authors.

Fig. 1 shows the resulting architecture. Five source archives enter through three native format families, DICOM series, NIfTI volumes, and pre-rendered PNG or NumPy arrays, and through two acquisition regimes, natively two-dimensional projection imaging and volumetric imaging requiring slice extraction. Source-specific decoding and correction paths converge on a common harmonization stage, which resamples to the target grid and writes into the shared HDF5 schema. A separate estimation branch consumes the reference subsets of the harmonized corpus and produces the reference statistics. The complete workflow is given as Algorithm 1, with symbols and fixed parameter values collected in Table 3 and each operation formalized in the equations of Sections II-C to II-H. Stage 1 restores physical intensities, stage 2 extracts informative axial slices from the volumetric sources, stage 3 resamples and quantizes to the common grid, and stage 4 assigns patients to partitions and writes the HDF5 files. A reference-estimation branch and a quality-control branch consume the harmonized corpus and supply the statistics and fidelity measures reported in Section III.

B. Source Archives and Selection Criteria

Five publicly released, previously de-identified archives were selected to span the three modalities of interest while maximizing acquisition diversity. Table 1 summarizes them. Selection criteria were public availability under terms permitting secondary research use, availability of a patient or subject identifier sufficient to enforce patient-level partitioning, and, for at least one source per modality where possible, availability of a reference subset from which reference statistics could be estimated. The last criterion is satisfied for brain magnetic resonance imaging by OpenBHB, whose subjects are uniformly labelled as controls, and for chest radiography by the "No Finding" subset of NIH ChestX-ray14. It is not satisfied for computed tomography, for reasons discussed in Section IV.

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

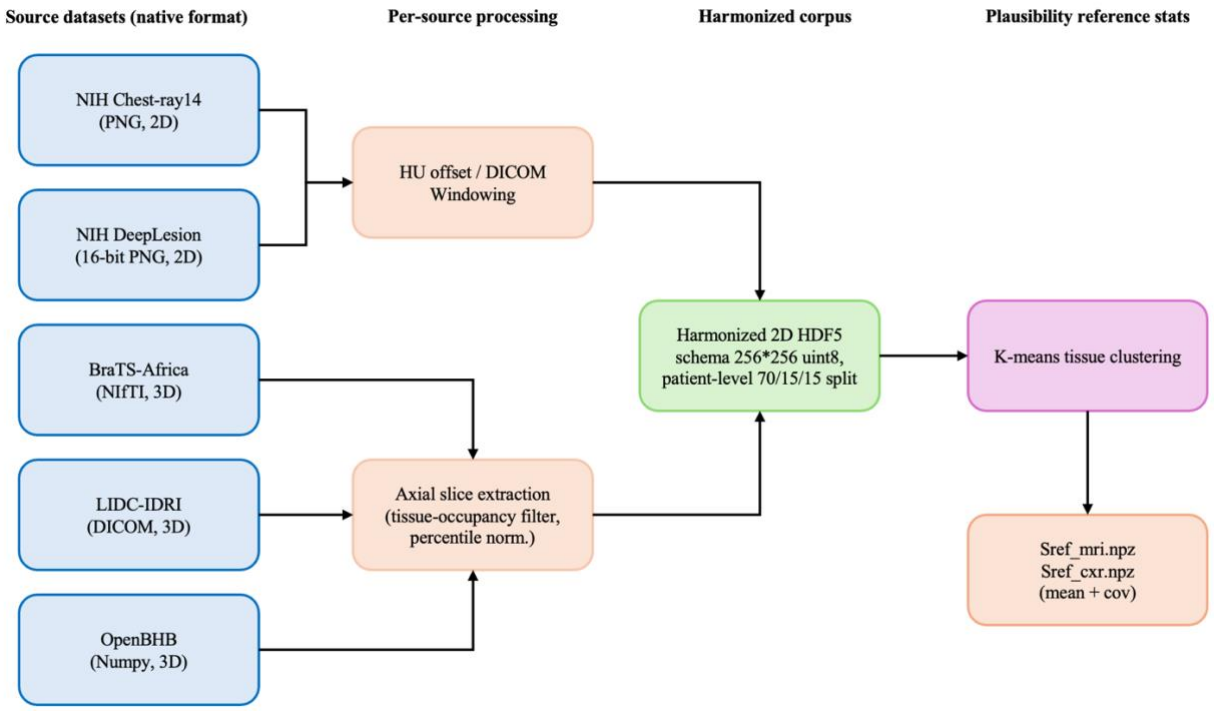


Fig. 1. Architecture of the IntrinsicX multi-source medical image harmonization pipeline. Five public imaging archives containing chest X-ray (CXR), computed tomography (CT), and brain magnetic resonance imaging (MRI) undergo source-specific intensity correction, informative-slice extraction where applicable, spatial resampling, and patient-level partitioning. The processed images are stored in a unified HDF5 corpus, while reference subsets are used to estimate intensity-based compositional statistics.

C. Source-Specific Decoding and Intensity Correction

1. NIH ChestX-ray14

A manifest was constructed by joining the released Data_Entry_2017.csv metadata file, which supplies the patient identifier, filename, and fourteen multi-label thoracic pathology annotations, against every PNG image present in the source release. Images were converted to single-channel grayscale, resampled to 256x256 (Eq. (5)), and stored as unsigned 8-bit integers. The "No Finding" subset, comprising 42,408 training, 9,178 validation, and 8,775 test images, was retained as a separate index and used as the chest radiography reference for the estimation described in Section II-G. The chest radiography reference subset therefore comprises 42,408 of the 78,934 training-split images, or 53.7 per cent.

2. NIH DeepLesion

Source images are 16-bit PNGs in which raw Hounsfield-Unit values are stored with a fixed additive offset, such that the stored pixel value equals the Hounsfield value plus 32,768. The offset was reversed before any further processing, restoring true Hounsfield values. The image-specific display window recorded in the released DL_info.csv metadata was then applied on a per-image basis rather than imposing a single fixed window across the archive, preserving the diagnostic intent of each original acquisition. Windowed images were resampled to 256x256 and stored as unsigned 8-bit integers. Writing P(x,y) for the stored pixel value and h(x,y) for the restored Hounsfield value [45], the offset reversal, Eq. (1) [22], [45], is

$$h(x,y) = P(x,y) - 32768 \quad (1)$$

and the image-specific window of centre c and width w recorded in DL_info.csv is applied as the linear VOI function of DICOM [46] in its continuous form, Eq. (2) [46], with clip(a, 0, 1) = min(1, max(0, a)),

$$W(h;c,w) = \text{clip}\left(\frac{h-c+w/2}{w}, 0, 1\right) \quad (2)$$

Windowed values u = W(h; c, w) in [0, 1] are converted to unsigned 8-bit integers by truncating 255u to its integer part [47]. For DeepLesion, c and w vary from image to image; for LIDC-IDRI, they are fixed (Section II-D).

3. BraTS-Africa

The source release provides five NIfTI volumes per patient, comprising four structural sequences (t1n, t1c, t2f, t2w) and a segmentation mask, for 146 patients across two tumour categories. Volumes were loaded with nibabel and retained at native resolution without resampling as 32-bit floating-point arrays, one HDF5 group per patient, with the tumour category preserved as a group attribute. These volumes served as the source of truth for the slice-extraction stage described in Section II-D.

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Algorithm 1. IntrinsicX medical image harmonization and compositional reference-estimation procedure. The algorithm describes the complete workflow for integrating five public medical imaging archives, including source-specific intensity correction, informative axial-slice selection, spatial resampling to 256 × 256 pixels, 8-bit quantization, patient-level training/validation/test partitioning, HDF5 storage, four-cluster k-means reference estimation, and image quality assessment. Inputs are the source archives and fixed processing parameters; outputs are the harmonized HDF5 files, modality-specific reference statistics, and quality-control report. HU: Hounsfield unit; MRI: magnetic resonance imaging; CXR: chest X-ray; HDF5: Hierarchical Data Format version 5.

(1) **Input:** Archives A1,…,A5; fixed parameters of Table 3
(2) **Output:** HDF5 files; Sref_mri.npz; Sref_cxr.npz; quality report
(3) **FOR** each archive a **DO**
(4) Build the manifest joining patient_id, site_tag and labels to files
(5) **IF** a = NIH ChestX-ray14 **THEN** convert to grayscale
(6) **IF** a = DeepLesion **THEN** restore HU by Eq. (1); apply the per-image window by Eq. (2)
(7) **IF** a = LIDC-IDRI **THEN** order slices by z; restore HU with the DICOM slope and intercept; keep the 512×512 volume
(8) **IF** a = BraTS-Africa or OpenBHB **THEN** keep the native float32 volume
(9) **IF** a is volumetric **THEN** compute slice occupancy and keep slices (Eq. (3)); LIDC-IDRI: lung window (Eq. (2)); MRI: 1st to 99th percentile normalization per slice (Eq. (4))
(10) Resample every image to 256×256 (Eq. (5)); quantize to uint8
(11) **END FOR**
(12) Draw the patient-level partition with seed 42 (Eq. (6)); verify disjoint patient sets on the written files; report realized ratios (Eqs. (7), (8))
(13) Write HDF5 groups with the schema of Table 2
(14) **FOR** each modality m ∈ {MRI, CXR} **DO**
(15) Sample N = 5,000 reference images
(16) Per image: sample 2,000 pixels above 5; fit k-means with K = 4 (Eq. (9)); order the centroids; compute s_i (Eq. (10))
(17) Compute mean and covariance (Eqs. (11), (12)); check singularity; save Sref
(18) **END FOR**
(19) Quality control on 2,000 images per source; bootstrap and sensitivity analyses (Eqs. (15)–(21))

4. LIDC-IDRI

Series-level computed tomography volumes were retrieved through the TCIA nbia.downloadSeries interface. Image data only were retrieved; the nodule-annotation package was not, with the consequence stated in Section IV-D. Of 15,116 retrieved series folders, filtering to series of at least 30 slices carrying Modality = "CT", which excludes scout and localizer acquisitions, yielded 1,018 valid volumes from 1,010 subjects; the discrepancy reflects subjects contributing more than one qualifying series. Within each volume, slices were ordered by the third component of ImagePositionPatient and converted to true Hounsfield values using the per-slice RescaleSlope and RescaleIntercept, then retained at native 512x512 in-plane resolution as three-dimensional source-of-truth data. One further series failed with an input/output error and was excluded. For LIDC-IDRI, each slice z is converted to Hounsfield Units as $h_z = s_z P_z + b_z$, where P_z is the stored array and s_z and b_z are its RescaleSlope and RescaleIntercept [46].

5. OpenBHB

Quasi-raw T1-weighted volumes in NumPy format, with participant metadata recording acquisition site, age, sex, diagnosis and originating study, were retrieved from the HuggingFace Hub release. It comprises 3,984 subjects in three partitions defined by their originators: 3,227 training subjects across 57 acquisition sites, an internal test partition of 362 subjects across 55 sites, and an external test partition of 395 subjects across 6 sites withheld for cross-site generalization. Quality-control filtering on the released quasiraw correlation metric at 0.5 retained all 3,984 subjects, the minimum observed being 0.502, so the distributed data had already been quality-filtered by its originators. Only the training partition was ingested, giving 3,227 subjects uniformly labelled as controls across 57 distinct real sites, each retained as a per-subject HDF5 group in 32-bit floating point tagged with its site. The ingested cohort is 1,669 males and 1,558 females, mean age 25.2 years (standard deviation 14.6, range 6 to 86),

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Table 1. Characteristics and harmonized image yields of the five public medical imaging archives. Source datasets are summarized by imaging modality, original data distribution, total number of harmonized two-dimensional images, and storage footprint. All processed images have a spatial resolution of 256 × 256 pixels. CT: computed tomography; MRI: magnetic resonance imaging; TCIA: The Cancer Imaging Archive; GB: gigabytes (1 GB = 1,024 MB).

Archive	Modality	Origin / distribution	2D images (all splits)	Harmonized 2D size
NIH ChestX-ray14	Chest X-ray (native 2D)	NIH public release	112,120	4.89 GB
NIH DeepLesion	CT (native 2D)	NIH public release	32,735	906.20 MB
BraTS-Africa	Brain MRI (4-channel)	TCIA (PKG-BraTS-Africa)	18,270	932.77 MB
LIDC-IDRI	CT (volumetric)	TCIA (nbia.downloadSeries)	243,449	9.61 GB
OpenBHB	Brain MRI, 57 sites	HuggingFace Hub	435,165	9.19 GB
Total	3 modalities	5 archives	841,739	25.49 GB

from ten contributing studies; the largest single site contributes 26.8 per cent and none fewer than five.

D. Informative Axial Slice Extraction

The three volumetric sources share one slice-extraction procedure. A reference channel was designated per volume: t1n for BraTS-Africa and the sole available channel for LIDC-IDRI and OpenBHB. An axial slice was retained if more than 5 per cent of its pixels exceeded a near-zero occupancy threshold τ (Eq. (3)), discarding the empty and near-empty slices at the extremes of every volume without imposing an anatomical landmark that would differ across modalities. For BraTS-Africa, all four sequences were extracted at the retained indices and stacked into a four-channel image, preserving inter-sequence registration. For LIDC-IDRI, a lung window of centre -600 HU and width 1500 HU was applied (Eq. (2); displayed range -1,350 to +150 HU). Magnetic resonance intensities were then normalized to the first-to-ninety-ninth percentile range per slice (Eq. (4)) rather than per volume, so that compositional statistics computed downstream are comparable across slices of differing anatomical content, and resampled to 256x256 unsigned 8-bit integers (Eq. (5)).

Let V(x,y,z) be the intensity of a volume on the pixel grid Ω of size H x W, and let 1[·] denote the indicator function. The occupancy of axial slice z, Eq. (3) [47], is

$$\rho_z = \frac{1}{|\Omega|}\sum_{(x,y)\in\Omega} \mathbf{1}\left[V(x,y,z) > \tau\right] \quad (3)$$

where ρ_z is the occupancy of slice z, the fraction of its pixels above τ, |Ω| = H x W is the number of pixels in the slice, τ is the near-zero intensity threshold (Table 3), and the slice is retained when ρ_z exceeds ρ_{min} = 0.05. A retained magnetic resonance slice with values v is clipped to its own 1st and 99th percentiles p_1 and p_{99} [48] and rescaled by Eq. (4) [48], with clip as defined above,

$$\tilde{u}(x,y) = \mathrm{clip}\left(\frac{v(x,y)-p_1}{p_{99}-p_1}, 0,1\right) \quad (4)$$

where ũ(x, y) is the normalized intensity of pixel (x, y), and every retained slice is resampled to the 256x256 grid by interpolation with the Lanczos kernel κ(t) = sinc(t) sinc(t/3) for |t| < 3 and zero elsewhere, as in Eq. (5) [47],

$$\hat{I}(u,v) = \sum_{i,j} \kappa\left(\xi_u - i,\ \eta_v - j\right) I(i,j) \quad (5)$$

where I is the slice before resampling, with input pixels indexed by (i, j), Î(u, v) is the resampled value at output pixel (u, v), $\xi_u = (u + 1/2)/s_x - 1/2$ and $\eta_v = (v + 1/2)/s_y - 1/2$ map output to input coordinates, and $s_x = 256/W$ and $s_y = 256/H$ are the scale factors. When an image is reduced, the implementation stretches the support of the kernel by the reduction factor, so that Eq. (5) gives the form of the interpolation and not every detail of its numerical evaluation. Output values are converted to unsigned 8-bit integers by truncation of 255u.

E. Patient-Level Partitioning

Partitioning into training, validation, and test sets targeted a 70/15/15 ratio and was performed at the level of the patient or subject identifier, never at the level of the individual image or slice, using a fixed random seed of 42 for every source. For the volumetric sources, this means that all slices derived from a given subject are confined to a single partition. For LIDC-IDRI, where a small number of subjects contribute multiple series, assignment was made at the series level within the constraint that no subject spans partitions. Because patients contribute unequal numbers of images, the realized image-level ratios necessarily deviate from the nominal target; the magnitude of that deviation is measured and reported in Section III-B rather than assumed.

With N patients (or subjects) in a source and nominal shares r_s^* of 70, 15 and 15 per cent for the training (tr),

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Table 2. Unified HDF5 data schema for the five harmonized imaging archives. The table specifies image dimensions and channel counts, patient identifiers, acquisition-site labels, diagnostic or categorical annotations, and slice-index availability for each source. HDF5: Hierarchical Data Format version 5; uint8: 8-bit unsigned integer; ch: image channels; 2D: two-dimensional.

Field	NIH ChestX-ray14	DeepLesion	LIDC-IDRI (2D)	BraTS-Africa (2D)	OpenBHB (2D)
images	uint8, 256x256, 1 ch	uint8, 256x256, 1 ch	uint8, 256x256, 1 ch	uint8, 256x256, 4 ch	uint8, 256x256, 1 ch
patient_id	present	present	present	present	present
site_tag	1 site	1 site	1 site	1 site	57 sites
label / category	14-way multi-label	lesion type	none (see IV-D)	2 tumour categories	diagnosis = control
slice_index	not applicable	not applicable	present	present	present

validation (va) and test (te) partitions, the training and validation sizes by Eq. (6) [40], [42] were

$$n_s = \left\lfloor \frac{r_s^* N}{100} \right\rfloor, \quad s \in \{\mathrm{tr}, \mathrm{va}\} \tag{6}$$

where n_s is the number of patients assigned to partition s and $\lfloor \cdot \rfloor$ rounds down to the nearest integer, with the test partition receiving the remainder, $n_{te} = N - n_{tr} - n_{va}$. This gives subject counts of 102, 21 and 23 for BraTS-Africa and 2,258, 484 and 485 for OpenBHB, patients being assigned by a random permutation seeded with 42. The three patient sets were required to be pairwise disjoint, a condition verified on the written files [42], and the realized image-level share of each partition and its largest deviation from the nominal target were computed by Eqs. (7) and (8) [40], [42] as

$$r_s = \frac{\sum_{p \in \mathcal{P}_s} m_p}{\sum_p m_p} \tag{7}$$

$$\Delta = \max_s |\, 100\, r_s - r_s^* \,| \tag{8}$$

where m_p is the number of images contributed by patient p, $\mathcal{P}_s$ is the set of patients assigned to partition s, r_s is the realized image-level share of partition s, and Δ is the largest absolute deviation from the nominal share, in percentage points.

Because the partition was drawn at subject level within a fixed set of sites, the site tags present in the three partitions number 57, 53 and 54, so essentially every site appears in all three: the released split measures generalization to unseen subjects and not to unseen acquisition sites. To support the latter, the deposit adds a site-held-out fold assignment, placing the 57 sites in five folds with no site spanning a boundary; the images are unchanged, and only the assignment is new. Because the largest single site contributes 26.8 per cent of subjects, more than a balanced fifth would hold, the folds are necessarily unequal, and that site forms one-fold alone; the per-fold subject and site counts accompany the assignment in the deposit. OpenBHB itself distributes an external test partition of 395 subjects across six withheld sites for exactly this purpose, which we did not ingest and identify as the natural extension for a future version.

F. Harmonized Storage Schema

All five sources are written into the single field schema shown in Table 2. Images are stored as fixed-size unsigned 8-bit arrays at 256x256, single-channel for chest radiography, computed tomography, and OpenBHB, and four-channel for BraTS-Africa. A patient_id field is carried through from the source metadata and permits independent verification of partition integrity by any downstream user. A site_tag field records the originating acquisition site; it is genuinely informative only for OpenBHB, which spans 57 sites, and carries a fixed single-site label for the remaining four sources, but is present uniformly so that site-aware training or evaluation code need not special-case its input. The label field is deliberately not unified across sources, retaining each archive's native taxonomy; the rationale is given in Section IV-D.

G. Estimation of Compositional Reference Statistics

The plausibility reference statistic is a compositional summary of anatomy at the level of intensity-derived clusters. For each modality with a reference subset, 5,000 images were sampled: 5,000 of 304,491 available OpenBHB training slices for magnetic resonance imaging and 5,000 of 42,408 available NIH "No Finding" images for chest radiography. Within each image, k-means was fitted to 2,000-pixel intensities drawn from pixels above an intensity of 5, so that the large background mass does not dominate centroid placement, using k-means++ initialization [49] with three restarts, a maximum of 300 iterations, a tolerance of 1 x 10-4 and a fixed random state of 42. The fitted model was applied to every pixel, including the background, and the four clusters were relabeled in ascending order of centroid intensity so that component k denotes the same intensity class in every image. The fraction assigned to each ordered cluster gives a four-element compositional vector summing to unity by

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Table 3. Mathematical notation and fixed parameters of the IntrinsicX harmonization pipeline. Symbols and parameter values define intensity restoration, DICOM windowing, axial occupancy selection, image resampling, quantization, patient-level partitioning, k-means clustering, reference-statistic estimation, bootstrapping, and quality control. HU: Hounsfield unit; CT: computed tomography; MRI: magnetic resonance imaging; CXR: chest X-ray; uint8: 8-bit unsigned integer.

Symbol	Definition	Value
P, h	Stored pixel value; restored Hounsfield value	offset 32,768 (DeepLesion)
s_z, b_z	RescaleSlope and RescaleIntercept of slice z	per DICOM slice (LIDC-IDRI)
c, w	Window centre and width (HU)	per image (DeepLesion); -600 and 1500 (LIDC-IDRI)
τ	Near-zero intensity threshold for slice occupancy	0.001 (native intensity units)
ρ_{min}	Minimum occupied fraction of a retained slice	0.05
p_1, p_{99}	Percentile bounds of per-slice normalization (MRI)	1st and 99th
H x W, κ	Native grid; interpolation kernel	Lanczos, a = 3
Q	Conversion of unit-range values to 8 bits	uint8, truncation, 0 to 255
r*	Nominal train/validation/test ratio	70/15/15 per cent
seed	Random seed (partitioning; k-means state)	42
K	Number of intensity clusters	4
m	Pixels sampled per image for k-means (intensity > 5)	2,000
N	Reference images per modality	5,000
k-means	Initialization; restarts; iterations; tolerance	k-means++; 3; 300; 1e-4
B, α	Bootstrap resamples; significance level	1,000; 0.05
m_d, initializations	Diagnostic analyses: pixels per image; k-means initializations	4,096; 1
F	Foreground threshold for quality control	intensity > 10
n_{QC}	Images sampled per source for quality control	2,000

construction. The mean vector and full four-by-four covariance matrix were serialized as Sref_mri.npz and Sref_cxr.npz, each recording the sample size and the eligible pool, and are the direct targets of the plausibility loss in the companion modelling work. No statistic was estimated for computed tomography, nor was any source containing a lesion-free reference subset. The clusters derive from pixel intensity alone and carry no validated anatomical interpretation; no comparison against segmentation ground truth or expert annotation was performed, and any physiological reading of the individual components would require it. Formally, let 2,000 foreground pixels (intensity above 5) be sampled from image I_i, $c_i(\cdot)$ be the cluster assignment, applied afterwards to every pixel of the image, and μ_k the centroids. The fit minimizes the within-cluster inertia of Eq. (9) [50],

$$J = \sum_{k=1}^{K} \sum_{x \in \mathcal{C}_k} (x - \mu_k)^2 \quad (9)$$

where K = 4 is the number of clusters, x is the intensity of a sampled pixel and $\mathcal{C}_k$ is the set of sampled pixels assigned to cluster k, after which the compositional vector of each image, Eqs. (10) to (12) [51], and its mean and covariance over the N = 5,000 reference images are

$$s_{i,k} = \frac{1}{|\Omega|} \sum_{x \in I_i} \mathbf{1}\,[c_i(x) = k], \qquad \sum_{k=1}^{K} s_{i,k} = 1 \quad (10)$$

$$\bar{\mathbf{s}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{s}_i \quad (11)$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{s}_i - \bar{\mathbf{s}})\,(\mathbf{s}_i - \bar{\mathbf{s}})^{\top} \quad (12)$$

where $s_{i,k}$ is the fraction of the |Ω| pixels of image I_i assigned to cluster k, $\mathbf{s}_i = (s_{i,1}, \ldots, s_{i,K})$ is the compositional vector of image i, $\bar{\mathbf{s}}$ is the mean compositional vector over the N reference images, $\hat{\boldsymbol{\Sigma}}$ is their K x K sample covariance matrix, and the superscript ⊤ denotes transposition. Because $\mathbf{s}_i$ lies on the simplex, the covariance is singular by construction (rank at most K - 1 = 3), so a Mahalanobis distance from the reference composition, as would be needed to penalize implausible tissue ratios, must use the Moore-Penrose pseudo-inverse, Eq. (13) [52],

$$D^2(\mathbf{s}) = (\mathbf{s} - \bar{\mathbf{s}})^{\top} \hat{\boldsymbol{\Sigma}}^{+} (\mathbf{s} - \bar{\mathbf{s}}) \quad (13)$$

where s is the compositional vector of an image to be assessed, $D^2(\mathbf{s})$ is its squared Mahalanobis distance from the reference mean $\bar{\mathbf{s}}$, and $\hat{\boldsymbol{\Sigma}}^{+}$ is the Moore-

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Penrose pseudo-inverse of Σ, or be evaluated after an isometric log-ratio transformation, Eq. (14) [53],

$$\mathbf{z} = \mathbf{\Psi}^{\top} \ln \mathbf{s}, \qquad \mathbf{\Psi}^{\top}\mathbf{\Psi} = \mathbf{I}_{K-1},\ \mathbf{\Psi}^{\top}\mathbf{1} = \mathbf{0} \tag{14}$$

in which z is the vector of K - 1 log-ratio coordinates of s, the logarithm is taken component-wise, Ψ is a K x (K - 1) orthonormal contrast matrix, I_{K-1} is the identity matrix of size K - 1, and 1 and 0 are vectors of ones and zeros, and all fractions must be strictly positive. Eqs. (13) and (14) state how the reference statistics can be used by a plausibility penalty; no value reported in this study was computed with them. The image-level standard error of each mean fraction is its sample standard deviation divided by the square root of N, and the inflation factor reported in Section III is the ratio of the subject-level bootstrap standard error (Eq. (15)) to this image-level value. The reference statistics of Eqs. (10) to (12) were estimated with 2,000 sampled pixels, three k-means++ initializations and random state 42. The diagnostic analyses of Section III (cluster-number sweep, seed reproducibility and parameter sensitivity) were repeated many times and therefore used a lighter setting of the same algorithm, with 4,096 pixels sampled from each image without the intensity filter and a single initialization (at most 300 iterations, tolerance 10^-4); their results describe the stability of the procedure and are not the reference values themselves.

H. Evaluation Metrics and Statistical Procedures

Uncertainty of the reference statistics was quantified by a subject-level bootstrap [54], [55]: subjects were resampled with replacement B = 1,000 times, the mean composition was recomputed for each resample, and the (1 - α) percentile interval with α = 0.05 was taken, Eq. (15) [54], [55],

$$\mathrm{CI}_{1-\alpha} = \left[\bar{s}^{*}_{(\alpha/2)},\ \bar{s}^{*}_{(1-\alpha/2)}\right] \tag{15}$$

where $\bar{s}^{*}_{(q)}$ is the q-th quantile of the bootstrap means, computed separately for each cluster fraction, and $\mathrm{CI}_{1-\alpha}$ is the resulting percentile confidence interval. Partitions of the same images obtained under different initializations were compared with the Adjusted Rand Index, Eq. (16) [56],

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathbb{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathbb{E}[\mathrm{RI}]} \tag{16}$$

in which RI is the Rand index, the fraction of pixel pairs on which the two partitions agree, $\mathbb{E}[\mathrm{RI}]$ is its expected value for random partitions with the same cluster sizes, and max(RI) is its largest attainable value. The number of clusters was examined with the silhouette coefficient, Eq. (17) [57], where s(i) is the silhouette of pixel i, a(i) is the mean distance of pixel i to its own cluster and b(i) is the mean distance to the nearest other cluster,

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{17}$$

the Calinski-Harabasz index [58] and the Davies-Bouldin index [59], both used in their standard form. Sensitivity to a processing parameter was expressed as the displacement of the mean composition from its value at the adopted setting, taken as the largest absolute change in any of the four cluster fractions. Fidelity of the delivered images was assessed on the foreground F_i of each image, defined as pixels above intensity 10, through the background fraction (the share of pixels outside F_i) and the saturation rate, the share of foreground pixels at the maximum code 255, Eq. (18) [47],

$$S_i = \frac{|\{x \in F_i : I_i(x) = 255\}|}{|F_i|} \tag{18}$$

where S_i is the saturation rate of image i, $I_i(x)$ is the intensity of pixel x, and $|\cdot|$ denotes the number of pixels in a set, and through the Shannon entropy, Eq. (19) [60], of the foreground histogram $p_i(g)$, the fraction of foreground pixels of image i with intensity g, and the corresponding effective number of intensity levels L_i,

$$H_i = -\sum_{g=0}^{255} p_i(g) \log_2 p_i(g), \qquad L_i = 2^{H_i} \tag{19}$$

where H_i is the entropy in bits and g runs over the 256 intensity codes. The cost of resampling was measured by a round trip, resampling each image to 256x256 by Eq. (5) and back to its native grid of size H x W, and by comparing the round-trip image with the original through the structural similarity index [61], with local means μ, variances σ² and covariance σ_{ab} and the constants $C_1 = (0.01 \times 255)^2$ and $C_2 = (0.03 \times 255)^2$ of the original formulation, evaluated with the scikit-image implementation (uniform 7 x 7 window, data range 255), Eq. (20) [61],

$$\mathrm{SSIM}(a, b) = \frac{(2\mu_a\mu_b + C_1)(2\sigma_{ab} + C_2)}{(\mu_a^2 + \mu_b^2 + C_1)(\sigma_a^2 + \sigma_b^2 + C_2)} \tag{20}$$

where a and b are the original and round-trip images, the local statistics are computed in each 7 x 7 window, and SSIM(a, b) is their mean over all windows, and through the peak signal-to-noise ratio, where MSE is the mean squared difference between the two images over the grid Ω, Eq. (21) [47],

$$\mathrm{PSNR} = 10 \log_{10} \frac{255^2}{\mathrm{MSE}} \tag{21}$$

where PSNR is expressed in decibels, and 255 is the largest intensity code. Processing time of each source was projected as the product of its mean measured time per unit and its number of units, and storage per image as the stored size divided by the number of images.

I. Implementation and Computing Environment

The pipeline is implemented in Python and was executed in Google Colab, with source archives, intermediate volumetric data, and final harmonized files staged on Google Drive. Format handling used pydicom [62] for DICOM series, nibabel for NIfTI volumes, h5py for the harmonized output, and scikit-learn [63] for the k-means clustering underlying the reference statistics. Every processing parameter, every exclusion decision, and every defect encountered during development, together with its resolution, is recorded in a pipeline log distributed with the public deposit, so that the harmonization can be re-executed

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

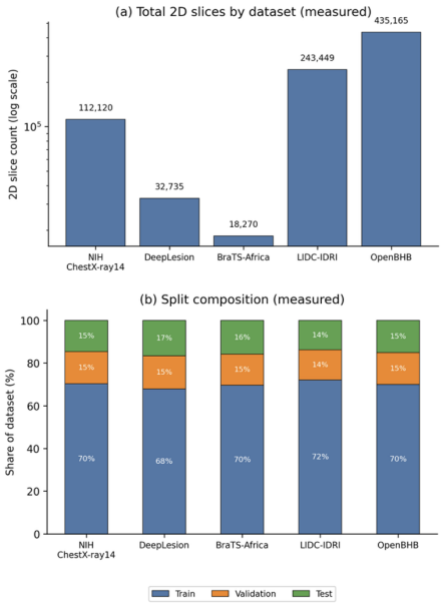


Fig. 2. Distribution and partitioning of the harmonized medical image corpus. (a) Total number of two-dimensional images from NIH ChestX-ray14, DeepLesion, BraTS-Africa, LIDC-IDRI, and OpenBHB, displayed on a logarithmic scale. (b) Realized training, validation, and test proportions, calculated from the stored images after patient-level partitioning. Percentages represent image-level proportions rather than patient-level proportions.

from the original archives independently of any artefact retained by the authors. All fixed parameters are listed in Table 3, and the processing order is given by Algorithm 1; together with the pinned package versions in the repository (Code Availability) they define a reproducible workflow.

III. Result

A. Composition of the Harmonized Corpus

The pipeline produced 841,739 harmonized two-dimensional images. Fig. 2(a) shows the per-source counts on a logarithmic axis, spanning more than an order of magnitude from DeepLesion at 32,735 to OpenBHB at 435,165. NIH ChestX-ray14 and DeepLesion are natively two-dimensional and their counts equal the released counts after manifest joining; the other three derive from slice extraction under the occupancy criterion of Section II-D, yielding 18,270 images from 146 BraTS-Africa patients, 243,449 from 1,018 LIDC-IDRI volumes and 435,165 from 3,227 OpenBHB subjects, or roughly 125, 239 and 135 informative slices per subject. These yields track the differing axial extents of the three protocols, whole-brain for the magnetic resonance sources and whole-thorax at finer slice spacing for LIDC-IDRI, confirming indirectly that the occupancy criterion neither truncated nor over-retained any one source. The five sources therefore contribute 13.3, 3.9, 2.2, 28.9 and 51.7 per cent of the corpus in the order of Table 1, so that OpenBHB alone holds more than half of all images.

B. Partition Integrity and Realized Ratios

Partition composition was measured from the group attributes of the written files rather than inferred from the nominal target. For the three volumetric sources the exact image-level counts are: BraTS-Africa 12,731 training, 2,650 validation, and 2,889 test; LIDC-IDRI 175,459, 34,524, and 33,466; and OpenBHB 304,491, 65,293, and 65,381. Expressed as proportions (Eq. (7)) these are 69.68, 14.50 and 15.81 per cent for BraTS-Africa, 72.07, 14.18 and 13.75 per cent for LIDC-IDRI and 69.97, 15.00 and 15.02 per cent for OpenBHB, so the largest absolute deviation among the three volumetric sources (Eq. (8)) is 2.07 percentage points, in the LIDC-IDRI training share, against 0.81 for BraTS-Africa and 0.03 for OpenBHB. The BraTS-Africa deviation reflects the small cohort, since partitioning 146 patients as 102, 21 and 23 cannot resolve the target ratio finely. For the two natively two-dimensional sources the measured proportions are 70/15/15 for NIH ChestX-ray14 and 68/15/17 for DeepLesion; the NIH ChestX-ray14 training partition holds 78,934 of 112,120 images (70.40 per cent); exact per-partition counts for these two files are recorded in their respective HDF5 group attributes. Fig. 2(b) presents all five measured proportions.

The DeepLesion deviation, about two percentage points in both the training and test shares and thus comparable to that of LIDC-IDRI, is the largest of the natively two-dimensional sources and is instructive. It arises directly from the patient-level constraint: because the number of lesion images contributed per patient varies substantially in that archive, a partition that is balanced in patients is not balanced in images. This is the expected and correct behaviour of a leakage-free partitioning scheme, and the deviation was retained rather than corrected, since correcting it at the image level would reintroduce the very leakage the constraint exists to prevent. Reporting the measured rather than the intended ratio is the minimum standard for any released corpus intended for model training.

C. Storage Footprint

Fig. 3 reports the harmonized two-dimensional footprint per source. The total is 25.49 GB, of which the validation and test partitions together with the two reference statistic files account for 7.42 GB and the training partitions for 18.07 GB. Footprint does not track image count uniformly across sources: BraTS-Africa contributes 932.77 MB from only 18,270 images, a per-image cost of 52.3 KiB against 22.1 to 45.7 KiB for the single-channel sources, or 1.1 to 2.4 times as much, a smaller ratio than its fourfold increase in channel count would suggest because stored size depends on image

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

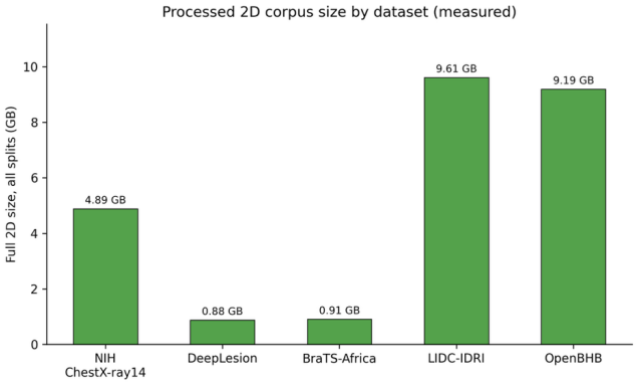


Fig. 3. Storage requirements of the harmonized medical image corpus by source dataset. Storage footprints are reported for the processed two-dimensional images from the five source archives, including all training, validation, and test partitions. Total harmonized storage is 25.49 GB, where 1 GB equals 1,024 MB.

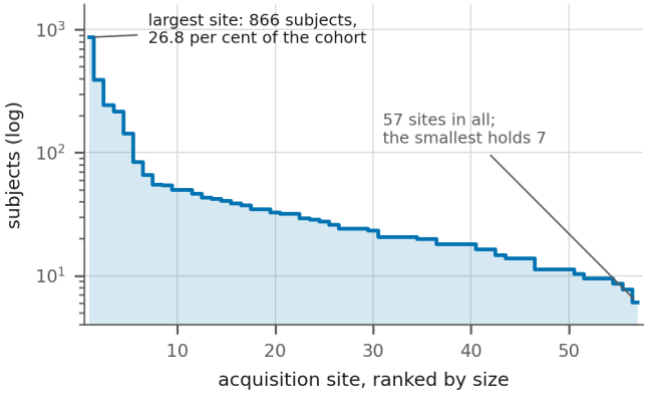


Fig. 4. Distribution of OpenBHB participants across acquisition sites. Subject counts for 3,227 participants are shown across 57 acquisition sites, ordered by decreasing cohort size. The largest site contributes 26.8% of subjects. The vertical axis represents the number of subjects on a logarithmic scale.

content as well as on channel count. Across the corpus the mean is 31.8 KiB per image; sizes are expressed in binary units (1 GB = 1,024 MB). Retaining the native volumetric data for the three volumetric sources as source of truth brings the complete corpus to 132.29 GB.

The storage required to regenerate the corpus is substantially larger than the corpus itself, and prospective reusers should plan accordingly. Reproduction from the original archives requires downloading and retaining the raw source releases before harmonization, together with transient working space for DICOM series expansion, volume assembly, and axial slice extraction; peak storage occupancy during a full end-to-end run is on the order of 2 TB. This storage requirement is a one-time cost and does not apply to users who obtain the harmonized corpus directly from the public deposit, which requires only 25.49 GB of storage. It is incurred only by users who wish to regenerate the corpus independently, either to audit the harmonization against the original acquisitions or to alter a processing decision, and retaining the native volumetric data as source of truth is what makes that verification possible.

D. Multi-Site Coverage

Acquisition-site diversity is the property that most distinguishes this corpus from a single-institution collection, and it is contributed almost entirely by OpenBHB. Fig. 4 shows the distribution of the 3,227 retained subjects across the 57 sites, sorted by descending subject count. The distribution is strongly right-skewed: a small number of high-volume contributing sites account for a large share of subjects, the single largest contributing 26.8 per cent, followed by a long tail of low-count sites. This shape is characteristic of aggregated multi-cohort neuroimaging releases and has a practical consequence for any user of the corpus: a model trained without site-aware sampling will be dominated by the high-volume sites, and any evaluation of site generalization must be stratified rather than pooled. The site_tag field is carried through the harmonized schema precisely so that such stratification is possible without returning to the source metadata.

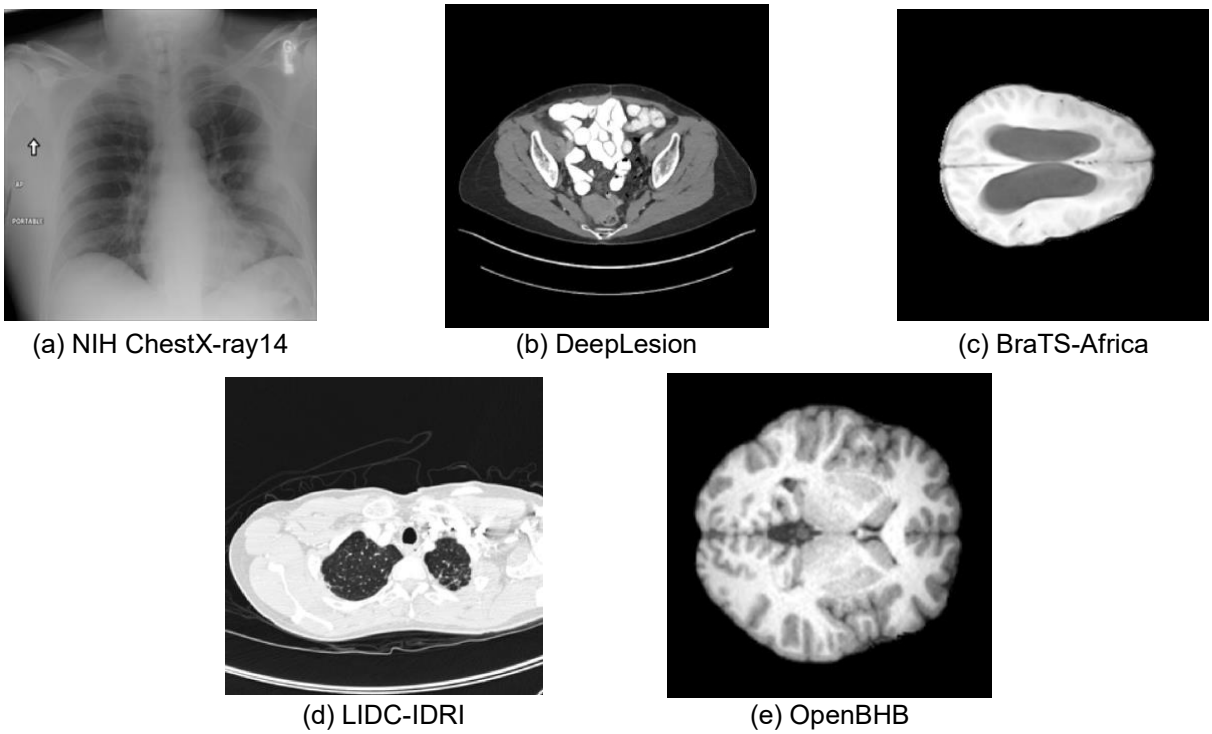
E. Cross-Modality Visual Consistency

Fig. 5 shows one representative test-partition slice from each source at the native harmonized resolution. The panel confirms visually what the numerical results establish indirectly: that projection radiography, cross-sectional computed tomography under two different windowing regimes, and single- and multi-channel brain magnetic resonance imaging have been brought onto a common grid and a common intensity range without any source appearing saturated, inverted, or truncated. The two computed tomography panels are of particular interest, since they follow different windowing paths, the per-image DICOM window for DeepLesion and the fixed lung window for LIDC-IDRI, and yet occupy comparable dynamic ranges after harmonization.

F. Compositional Reference Statistics

Table 4 reports the mean compositional vectors and Table 5 the corresponding covariance matrices, each estimated from 5,000 images of the reference subsets; Fig. 6 presents the two mean vectors graphically and Fig. 7 the two covariance matrices. Three properties confirm that the estimation is well formed. First, both mean vectors sum to unity, as required by construction. Second, no cluster collapsed toward zero or unity in either modality. Third, and most stringently, every row of both covariance matrices (Table 5) sums to zero to within machine precision, and both matrices have rank three rather than four. This is the algebraic signature of a covariance estimated on a simplex-constrained

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).



(a) NIH ChestX-ray14 (b) DeepLesion (c) BraTS-Africa

(d) LIDC-IDRI (e) OpenBHB

Fig. 5. Representative harmonized medical images from the five source archives. Images from (a) NIH ChestX-ray14 (chest radiography), (b) DeepLesion (abdominal/pelvic CT), (c) BraTS-Africa (brain MRI), (d) LIDC-IDRI (thoracic CT), and (e) OpenBHB (brain MRI) are displayed at 256 × 256 pixels after harmonization. All examples are from the test partition and represented as 8-bit unsigned integers (uint8; intensity range 0–255). For BraTS-Africa, one of four MRI channels is displayed.

vector and could not arise from a miscomputed statistic. The estimator fits its centroids to 2,000-pixel intensities sampled per image, and the resulting sampling variability is negligible at the deposited sample size: across five pixel-sampling seeds the mean composition moves by 0.00007 for magnetic resonance imaging and 0.00048 for chest radiography, one to two orders of magnitude below the sampling error of the images themselves. The estimate converges as the inverse square root of the number of images across the range 100 to 5,000, departing from that law by at most 5 per cent, and reaches a standard error of 0.0019 for magnetic resonance imaging and 0.0014 for chest radiography at the 5,000 images used here, so the deposited values are not in a regime where adding images would move them materially. Resampling whole subjects rather than individual images inflates that error by a factor of 1.02 for magnetic resonance imaging but 1.13 for chest radiography; axial slices taken at different levels through one brain differ substantially from one another, whereas repeat radiographs of one patient are close to duplicates, so the chest reference carries slightly less independent information than its image count suggests. For brain MRI the mean fractions of clusters 1 to 4 are 0.7009, 0.0881, 0.1023 and 0.1088, with 95 per cent subject-level bootstrap intervals (Eq. (15); Table 4) of 0.6970 to 0.7051, 0.0866 to 0.0895, 0.1007 to 0.1036 and 0.1070 to 0.1104, that is half-widths of at most 0.0041 or 0.6 to 1.7 per cent of the estimate. For chest radiography the mean fractions are 0.1952, 0.2341, 0.2920 and 0.2787 with an image-level standard error of 0.0014, which the subject-level inflation factor of 1.13 raises to about 0.0016.

The two modalities differ markedly in both composition and dispersion. The magnetic resonance reference places 0.7009 of the image in the darkest cluster, which is consistent with the acquisition geometry: an axial brain slice occupies a minority of a square field of view, and the surrounding region is air. The chest radiography reference is far more evenly distributed, spanning 0.1952 to 0.2920, as expected for a projection radiograph that fills the detector. Magnetic resonance cluster variances lie between 0.0021 and 0.0182, and chest radiography variances between 0.0020 and 0.0096. The covariance structure differs between the two modalities in a way that is interpretable. In the chest radiography reference every off-diagonal entry is negative (Table 5; Fig. 7(b)), the behaviour expected of a compositional statistic in which an increase in one share necessarily reduces the others. In the brain MRI reference the three brighter clusters are negatively correlated with the dominant

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Table 4. Mean fractional composition and bootstrap confidence intervals of the reference imaging distributions. Mean fractions for four intensity-derived clusters are presented for OpenBHB brain MRI and NIH ChestX-ray14 "No Finding" radiographs, each using 5,000 images. Clusters are ordered by increasing centroid intensity. The table includes 95% subject-level bootstrap confidence intervals for brain MRI. CI: confidence interval; MRI: magnetic resonance imaging.

Cluster (ascending intensity)	Brain MRI (OpenBHB)	Chest X-ray (No Finding)	95% CI, subject bootstrap (MRI)
Cluster 1	0.7009	0.1952	0.6970 - 0.7051
Cluster 2	0.0881	0.2341	0.0866 - 0.0895
Cluster 3	0.1023	0.2920	0.1007 - 0.1036
Cluster 4	0.1088	0.2787	0.1070 - 0.1104
Sum	1.0000	1.0000	

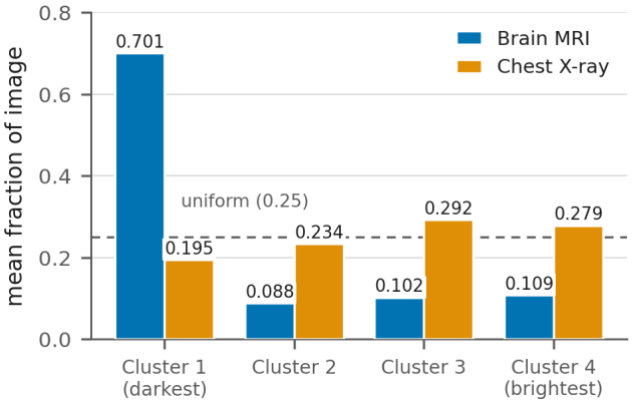


Fig. 6. Mean intensity-cluster composition of brain MRI and chest X-ray reference images. Mean fractional contributions of four intensity-derived clusters are shown for OpenBHB brain MRI and NIH ChestX-ray14 images labelled "No Finding," using 5,000 reference images per modality. Clusters are ordered by increasing centroid intensity. Fractions are dimensionless and sum to unity; the dashed line indicates a uniform fraction of 0.25.

dark cluster but positively correlated with one another, because all three compete against background rather than against each other (Table 5; Fig. 7(a)). In both cases the rows sum to zero and the matrix is singular (Table 5), which constrains how a plausibility penalty may be formulated and is discussed in Section IV-B.

G. Selection of the Number of Clusters

The number of intensity-derived clusters is a free parameter of the reference statistic, and we evaluated k from 2 to 8 on 1,000 images per modality. For each value we recorded the within-image inertia, the silhouette coefficient, the Calinski-Harabasz and Davies-Bouldin indices, the Adjusted Rand Index between partitions obtained under five different random initializations, and the smallest mean cluster fraction. Fig. 8 summarizes the three quantities that bear on the choice, namely the Adjusted Rand Index (Fig. 8(a)), the silhouette coefficient (Fig. 8(b)) and the smallest mean cluster fraction (Fig. 8(c)); the remaining indices are reported with the deposited analysis code.

No internal validity index selects k = 4. The silhouette coefficient decreases monotonically and favours k = 2 in both modalities (Fig. 8(b)), and the inertia curve shows no sharp elbow. This is expected for one-dimensional intensity data, in which pixel values form a continuum rather than well-separated groups, and we do not claim otherwise. The choice of k is therefore not a claim about latent structure but a trade-off between representational richness and reproducibility. We select k = 4 as the largest value at which the partition remains highly reproducible across random initializations in both modalities, with an Adjusted Rand Index of 0.988 for magnetic resonance imaging and 0.910 for chest radiography (Fig. 8(a)), and at which no cluster degenerates, the smallest mean fraction being 0.077 and 0.139 (Fig. 8(c); the smallest mean cluster fraction at K = 4 in the sweep of 1,000 images per modality, obtained with the diagnostic setting of Section II-G; the reference estimator of Table 4 gives 0.0881 and 0.1952). At k = 5 the chest radiography index (Fig. 8(a)) falls to 0.861 and declines to 0.725 at k = 8, so the partition ceases to be a reliable basis for a reference statistic. The mean composition shifts by at most 0.006 across the five initializations. Several hundred slices near the extremes of the volumetric sources contain fewer than four distinct intensity values and cannot support a four-cluster decomposition at all, an independent justification for the occupancy filter of Section II-D.

H. Sensitivity to the Occupancy Threshold

The occupancy threshold of 5 per cent governs which axial slices enter the corpus from the three volumetric sources, and we evaluated thresholds of 1, 2, 5, 10 and 20 per cent on all three. Fig. 9(a) reports the retained slice fraction and Fig. 9(b) the resulting fraction assigned to the darkest intensity-derived cluster. The threshold is a plateau rather than a knife-edge. LIDC-IDRI is entirely insensitive between 1 and 10 per cent (Fig. 9(a)), returning identical slice counts and an identical composition vector to four decimal places, because thoracic computed tomography slices always contain tissue and the criterion never binds. The two brain sources respond smoothly and monotonically: moving from 1 to 5 per cent costs BraTS-Africa six percentage points of slices (Fig. 9(a)) and shifts the composition vector by 0.016 (Fig. 9(b)), while 5 to 10 per cent costs a further twelve points and shifts it 0.019. There is no discontinuity in the neighbourhood of the chosen value. Fig. 10 shows the transition across the

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

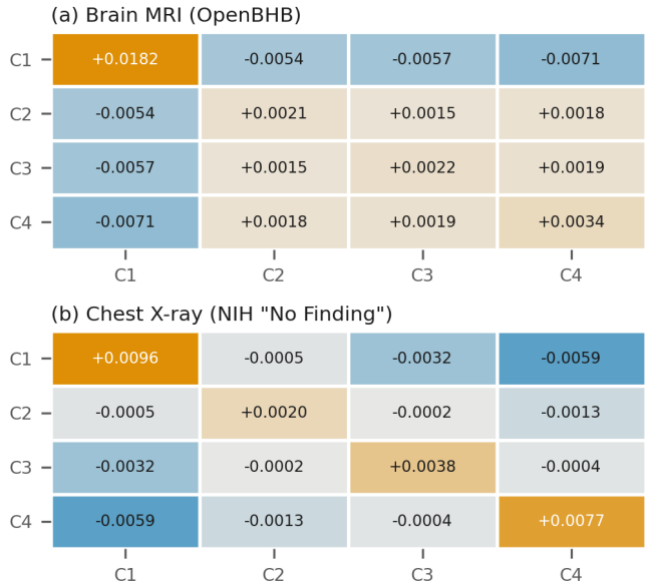


Fig. 7. Covariance matrices of intensity-derived cluster fractions for two imaging modalities. (a) Brain MRI reference distribution from OpenBHB and (b) chest radiography reference distribution from NIH ChestX-ray14. C1–C4 denote clusters ordered by increasing centroid intensity. Matrix entries represent covariances of dimensionless cluster fractions, estimated from 5,000 images per modality. Both covariance matrices have rank three; separate color scales are used for the two panels.

superior boundary slice by slice, for one volume from each magnetic resonance source.

The retained axial coverage confirms that the criterion removes what it is intended to remove: at 5 per cent, BraTS-Africa retains normalized axial positions 0.076 to 0.908, discarding approximately the top nine and bottom eight per cent of each volume, which correspond to the empty vertex and the skull base. Descending through the superior boundary in the volumes of Fig. 10, the discarded slices carry an empty frame or a fragment of scalp, while the first retained slice already shows cortex.

I. Normalization Strategy

Intensities are normalized to the 1st-to-99th percentile range of each slice, computed over the whole slice. We compared this against three alternatives on the two magnetic resonance sources, measuring saturation within the foreground, foreground entropy and the resulting composition. Computing the percentiles over the foreground alone reduces foreground saturation from 11.33 to 1.04 per cent for BraTS-Africa and from 10.86 to 1.05 per cent for OpenBHB while raising entropy. These baseline rates were measured on 300 slices taken at evenly spaced positions of the raw volumes without the occupancy filter, which is why they exceed the 4.872 and 2.992 per cent measured on the

Table 5. Covariance matrices of four intensity-derived cluster fractions for brain MRI and chest radiography. Covariance matrices were estimated from 5,000 reference images per modality. C1–C4 indicate clusters ordered by increasing centroid intensity. Matrix entries represent covariances of dimensionless fractional compositions; each matrix has rank three and zero row sums.

Modality	Cluster	C1	C2	C3	C4
Brain MRI (OpenBHB)	C1	0.0182	-0.0054	-0.0057	-0.0071
	C2	-0.0054	0.0021	0.0015	0.0018
	C3	-0.0057	0.0015	0.0022	0.0019
	C4	-0.0071	0.0018	0.0019	0.0034
Chest X-ray (No Finding)	C1	0.0096	-0.0005	-0.0032	-0.0059
	C2	-0.0005	0.0020	-0.0002	-0.0013
	C3	-0.0032	-0.0002	0.0038	-0.0004
	C4	-0.0059	-0.0013	-0.0004	0.0077

delivered slices in Table 6; the comparison between strategies is internally consistent because all four were applied to the same slices. The implemented choice therefore does discard information: when background dominates a slice, as it does for 78.6 per cent of a BraTS-Africa slice and 66.3 per cent of an OpenBHB slice, the 99th percentile of the full slice falls near the 95th percentile of the tissue distribution and a measurable fraction of anatomy saturates, 4.872 per cent of foreground pixels in BraTS-Africa and 2.992 per cent in OpenBHB in the delivered corpus (Table 6). The compositional statistic is nonetheless substantially robust to the choice: despite a tenfold difference in saturation the mean composition moves by at most 0.016 for BraTS-Africa and 0.004 for OpenBHB, so the released reference statistics do not hinge on this decision even though image fidelity does.

These two saturation rates are what whole-slice percentile clipping predicts. If background pixels all lie below the 99th percentile and tissue occupies a fraction 1 - β of the slice, a fraction of about 0.01/(1 - β) of tissue pixels exceeds it, that is 4.67 per cent for BraTS-Africa (β = 0.786) and 2.97 per cent for OpenBHB (β = 0.663), against 4.872 and 2.992 per cent measured (Table 6). Saturation is therefore a property of the normalization rule and not an acquisition artefact. For image fidelity it means that tissue contrast above the clipping level is lost in the delivered images, so that a model trained on them cannot reproduce graded intensity within the brightest 4.9 and 3.0 per cent of tissue, although the

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

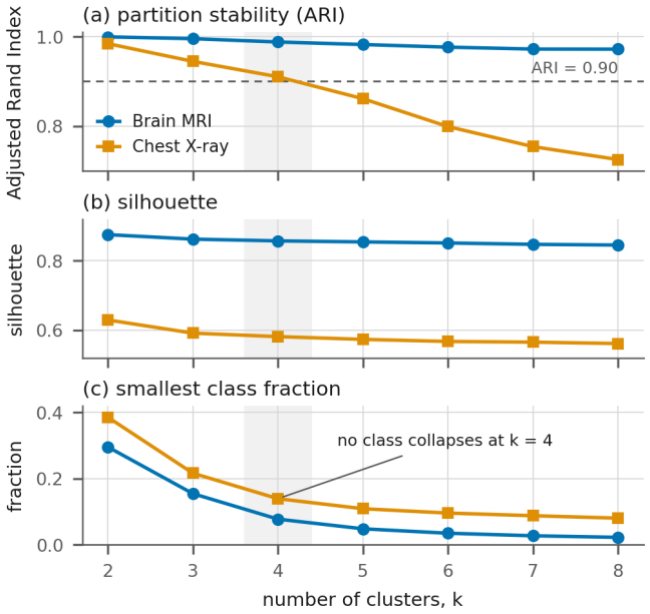


Fig. 8. Evaluation of cluster-number selection for compositional reference estimation. (a) Adjusted Rand Index (ARI) across five random initializations, (b) silhouette coefficient, and (c) minimum mean cluster fraction for cluster numbers k = 2 to 8. Metrics were evaluated using 1,000 images per modality. The dashed line in (a) indicates ARI = 0.90, and the shaded region identifies the selected value, k = 4.

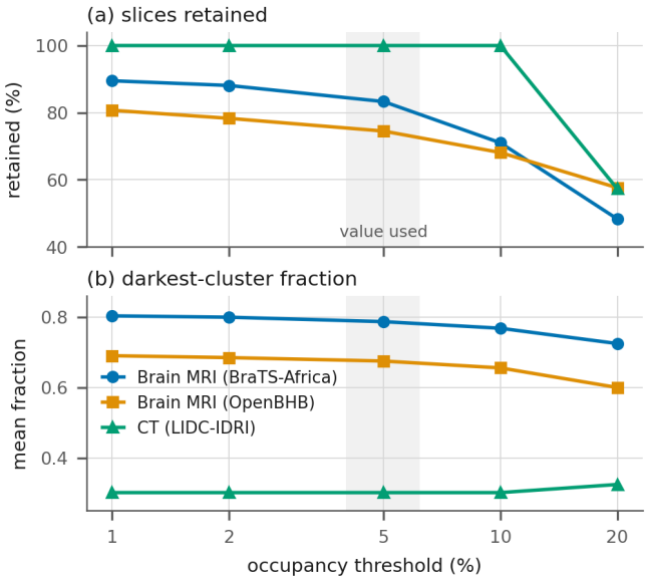


Fig. 9. Sensitivity of volumetric image selection and cluster composition to the axial occupancy threshold. (a) Proportion of axial slices retained from BraTS-Africa, LIDC-IDRI, and OpenBHB and (b) mean fraction assigned to the lowest-intensity cluster. Occupancy thresholds of 1%, 2%, 5%, 10%, and 20% were evaluated using the proportion of pixels exceeding the near-zero intensity threshold. The shaded region indicates the selected 5% threshold.

foreground still uses 152 and 211 effective intensity levels (Table 6). For compositional stability the effect is small, because the mean composition moves by at most 0.016 and 0.004 when the percentiles are computed over the foreground instead. The trade-off between intensity preservation and compositional stability was resolved in favour of whole-slice normalization because it needs no foreground threshold, avoids a new deposit version and leaves the reference statistics stable to within 0.016; foreground-only normalization, which reduces saturation roughly tenfold, is the recommended change for the next version of the corpus (Section IV-D).

For computed tomography the operative choice is a clinical display window rather than percentile normalization, and we compared the implemented lung window against mediastinal, soft-tissue, bone and wide windows and against percentile normalization with and without a body mask. The lung window crushes the fewest pixels to black, 8.5 per cent against roughly 55 per cent for every narrow window, and is the only setting producing a non-degenerate four-cluster decomposition, with all clusters between 0.19 and 0.32 where every alternative push one cluster to between 0.02 and 0.06. The 1st percentile of the sampled Hounsfield distribution sits at -3024, so at least one per cent of each slice is out-of-field padding rather than air, which makes percentile normalization unsuitable for computed tomography on its own terms. The first-cluster fraction ranges from 0.30 to 0.62 depending purely on the display convention adopted, a further reason that no computed tomography reference statistic is released.

J. Quality Control and Resampling Fidelity

Visual inspection of representative slices is not sufficient evidence that harmonization preserved diagnostically meaningful anatomy, and we therefore report quantitative quality-control measures computed on 2,000 sampled images per source. All measures are conditioned on foreground, defined as pixels above an intensity of 10, since a brain or a thorax occupies only part of its frame and zero-valued pixels outside the anatomy are air rather than lost signal. Table 6 reports the background fraction as a descriptive property alongside saturation measured within the foreground, foreground entropy, and the effective number of intensity levels used.

Foreground entropy lies between 6.77 and 7.77 bits of a possible 8 for every source, establishing that harmonization preserved intensity diversity within the anatomy itself. The effective number of intensity levels (Eq. (19)) ranges from 109 for LIDC-IDRI to 218 for NIH ChestX-ray14, and the mean saturation rates (Eq. (18)) are 0.010, 1.934, 4.872, 2.809 and 2.992 per cent for NIH ChestX-ray14, DeepLesion, BraTS-Africa, LIDC-

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

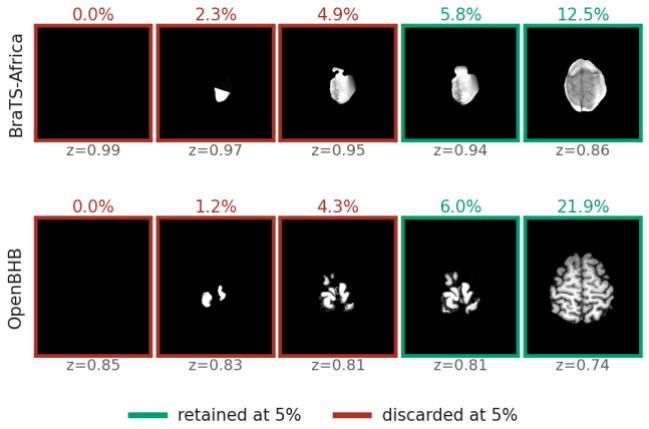


Fig. 10. Representative axial brain MRI slices illustrating the 5% occupancy selection criterion. Images from BraTS-Africa (subject BraTS-SSA-00007-000) and OpenBHB (subject 100053248969) illustrate the transition from excluded to retained slices near the superior anatomical boundary. Red borders indicate excluded slices and green borders indicate retained slices. Percentages denote occupied pixel fractions, and *z* represents normalized axial position (0–1). The selected criterion retains 131 of 155 BraTS-Africa slices and 134 of 182 OpenBHB slices.

IDRI and OpenBHB. Saturation is highest for BraTS-Africa and OpenBHB, where it follows from the interaction between whole-slice percentile normalization and a dominant background, and for LIDC-IDRI, where the lung window clips everything above approximately +150 Hounsfield Units by design. Both effects are discussed in Section IV-D.

The cost of resampling to the common grid was measured directly by round-tripping native-resolution slices through 256x256 and comparing against the original. LIDC-IDRI, natively 512x512, yields a structural similarity index (Eq. (20)) of 0.9233 with a minimum of 0.8283 and a peak signal-to-noise ratio of 32.1 dB (Eq. (21)). BraTS-Africa and OpenBHB are natively 240x240 and 182x218 respectively, so the common grid interpolates them upward rather than downward, and both exceed 0.99 structural similarity, at 0.9923 and 0.9917, with peak signal-to-noise ratios of 34.7 and 39.5 dB. Among the volumetric sources only LIDC-IDRI is genuinely downsampled, from 512x512 to 256x256 (four to one in pixel count), and it is the only one falling below 0.99; its minimum single-slice value of 0.8283 shows that fine detail is lost in the least favourable slices. Round-trip fidelity was not measured for the two natively two-dimensional sources (Section IV-D).

K. Computational Profile

The corpus was built in Google Colab on a two-vCPU instance with one physical core, 13.6 GB of random-access memory and no graphics processing unit, reading from and writing to Google Drive. Table 7 reports the measured per-unit processing cost of each source, timed on a representative sample and projected to the full cohort, together with the number of files read.

Total build time is 40.33 hours and peak resident memory never exceeded 0.72 GB, so the pipeline is not memory-constrained and will run on modest hardware. Storage is 25.49 GB for the harmonized two-dimensional corpus, 132.29 GB including the retained native volumes, and of order 2 TB peak during a full rebuild from the original archives. One observation is of practical value for reproduction: runtime is dominated by storage access rather than computation, and by directory structure rather than data volume. NIH ChestX-ray14 accounts for 76 per cent of the total while holding roughly a third of the source data, because its 112,120 images reside in a single directory where per-file lookup is expensive on a remote mounted filesystem, giving an effective throughput of 0.4 MB/s. Sources whose files are grouped into per-subject or per-series folders are bandwidth-bound and run at 11 to 15 MB/s, a thirty-five-fold difference. Reproduction on local storage would be bound by processor rather than input and output.

IV. Discussion

A. Relation to Prior Harmonization Work and to Existing Resources

Harmonization of multi-site medical imaging data has been approached from three directions, and separating them clarifies what the present work does and does not contribute. The first is statistical harmonization of derived measurements: empirical Bayes batch correction [32] and its neuroimaging adaptations [33], [34] remove site and scanner effects from features such as cortical thickness after segmentation, and learned variants pursue the same goal through site-invariant representations [35], [36]. These are methods rather than resources; they require site labels as input, operate downstream of feature extraction, and produce no image corpus. The second is general-purpose preprocessing software such as nnU-Net [37], MONAI [38] and TorchIO [39], which supplies the individual operations but leaves corpus assembly, partitioning and reference statistics to the user. The third, to which this work belongs, is the release of assembled multi-source corpora.

Against this background we distinguish engineering integration from methodological contribution. The individual operations applied here are standard: reversal of a documented storage offset, application of a recorded display window, percentile normalization, resampling to a common grid, and partitioning on a subject identifier. We claim no novelty for any of them. The contribution lies in three places. First, the corpus ships precomputed compositional reference statistics

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Table 6. Foreground-based image quality metrics for the harmonized medical imaging datasets. Quality measures were calculated using 2,000 sampled images per dataset and include background pixel fraction, foreground saturation at intensity 255, percentage of images with saturation exceeding 1%, foreground Shannon entropy, and effective intensity levels (2^{*H*}). Foreground pixels are defined by intensities greater than 10, where *H* is foreground entropy in bits.

Source	Backgr.	Sat. 255	Above 1%	Entropy	Levels
NIH ChestX-ray14	6.2%	0.010%	0.0%	7.77	218
DeepLesion	49.4%	1.934%	60.8%	7.42	171
BraTS-Africa	78.6%	4.872%	100.0%	7.25	152
LIDC-IDRI	14.3%	2.809%	93.5%	6.77	109
OpenBHB	66.3%	2.992%	100.0%	7.72	211

derived from reference anatomy and intended directly as regularization targets for constrained generative models. Second, partition integrity is verified against the written files rather than assumed from the nominal target, with realized ratios reported including where they deviate. Third, acquisition-site identity is carried through the harmonized schema for all 57 OpenBHB sites, so that site-stratified evaluation is possible without returning to the source metadata. Site effects remain present in the harmonized corpus by design; the tags are preserved precisely so that downstream users may model or remove them, which complements the correction methods above rather than substituting for them. Table 8 positions the harmonized corpus against seven published resources on indicators that can be read from their source publications: scale, resolution, number of sites, patient-level splits and the availability of reference statistics. MedMNIST v2 [27] aggregates a wider range of biomedical imaging tasks but at a uniform 28x28 resolution, a choice appropriate to its stated purpose of lightweight classification benchmarking and incompatible with generative synthesis, where the object of interest is precisely the spatial detail that resolution discards. RadImageNet [28] offers a substantially larger pool of radiologic images with patient-stratified splits and is well suited to transfer-learning pretraining, but does not include chest radiography and, like MedMNIST v2, supplies no reference statistics. The chest radiography archives are single-modality collections of comparable or larger scale than NIH ChestX-ray14 [21]: CheXpert [29] holds 224,316 radiographs from 65,240 patients, MIMIC-CXR [30] 377,110 images from 65,379 patients, and PadChest [31] more than 160,000 images from more than 67,000 patients, each from a single institution and none with a reference statistic. OpenBHB [26] is the closest neuroimaging counterpart; it is the source of 435,165 of our 841,739 images and of all 57 sites, but is distributed as three-dimensional quasi-raw volumes without a two-dimensional corpus or a composition statistic.

The corpus reported here is larger than MedMNIST v2 and smaller only than RadImageNet in raw image count, and it does not seek to displace either. Its distinguishing properties are that it retains the full 256x256 spatial resolution required for synthesis, that it spans chest radiography in addition to computed tomography and magnetic resonance imaging, and that it ships precomputed compositional reference statistics for two of its three modalities. It is purpose-built rather than general-purpose, and its value should be judged against the specific downstream task of plausibility-constrained generation rather than against benchmark breadth.

1. Resolution and harmonization quality.

At 256x256 the corpus carries 83.6 times as many pixels per image as the 28x28 images of MedMNIST v2 [27], a 9.1-fold gain in linear resolution, and the cost of reaching that grid is measured: structural similarity after a round trip is 0.9923 and 0.9917 for the two brain sources, which are interpolated upward from 240x240 and 182x218, and 0.9233 (minimum 0.8283, peak signal-to-noise ratio 32.1 dB) for LIDC-IDRI, which is reduced from 512x512. For scale, a 28x28 representation would reduce a 512x512 computed tomography slice by 18.3 to 1 per axis against 2 to 1 here; to our reading the source publication of MedMNIST v2 does not report a round-trip fidelity for that reduction, and none was measured in this work. Statistical harmonization methods [33], [34] are judged by the removal of site-related variance from derived features, an indicator that does not apply to intensity restoration and resampling; the corresponding indicators here are fidelity (structural similarity 0.9233 to 0.9923; peak signal-to-noise ratio 32.1 to 39.5 dB), residual saturation (0.010 to 4.872 per cent of foreground pixels), foreground entropy (6.77 to 7.77 bits) and the reproducibility of the derived statistic (Adjusted Rand Index 0.988 and 0.910 at K = 4, falling to 0.861 and 0.725 for chest radiography at K = 5 and K = 8). A structural similarity above 0.99 means that interpolation to the common grid leaves brain slices essentially unchanged, whereas 0.9233 for computed

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Table 7. Measured per-unit processing costs and projected total execution times for the IntrinsicX pipeline. Processing performance is summarized by dataset, processing unit, measured execution time per unit (seconds), number of processed units, number of files read, and projected total processing time (hours). Estimates are based on representative samples processed in the reported Google Colab environment; total projected execution time is 40.33 hours.

Source	Unit	s / unit	Units	Files	Hours
NIH ChestX-ray14	image	0.982	112,120	112,120	30.57
DeepLesion	image	0.625	32,735	32,735	5.68
LIDC-IDRI	series	7.100	1,018	198,510	2.01
OpenBHB	subject	1.954	3,227	3,227	1.75
BraTS-Africa	patient	7.878	146	730	0.32
Total				347,322	40.33

tomography quantifies a real but moderate loss of fine structure that matters for lesion-scale tasks and not for global composition.

2. Clustering stability

An Adjusted Rand Index of 0.988 for magnetic resonance imaging means that partitions of the same images obtained from different random initializations agree almost perfectly (1 denotes identical partitions, 0 chance agreement), and 0.910 for chest radiography still indicates strong agreement; both exceed the 0.90 criterion used to adopt K = 4. The fall to 0.861 at K = 5 and 0.725 at K = 8 for chest radiography shows that finer partitions of a one-dimensional intensity continuum are not reproducible. Because no internal index selects K = 4 (the silhouette favours K = 2), what is claimed is reproducibility and non-degeneracy of the reference statistic, not cluster validity.

Site coverage and bias. Fifty-seven acquisition sites distinguish the corpus from the single-institution chest archives [21], [29]-[31] in Table 8, but the coverage is uneven: one site contributes 26.8 per cent of OpenBHB subjects, and site tags span the training, validation and test partitions (57, 53 and 54 sites), so the released split measures generalization to unseen subjects and not to unseen sites, which the five site-held-out folds are meant to address.

3. Processing efficiency, storage and reproducibility.

A full build took a projected 40.33 hours on a two-vCPU machine with no graphics processing unit and a peak resident memory of 0.72 GB, of which 30.57 hours (75.8 per cent) were spent on one directory of 112,120 files read at 0.4 MB/s against 11 to 15 MB/s for per-subject or per-series folders. To our reading, the source publications of the resources in Table 8 report no per-source processing time, peak memory or fidelity measure for the preprocessing that produced their released images, so these figures are offered as a baseline for later comparison and not as a ranking. Storage follows the same logic: the deposited two-dimensional corpus needs 25.49 GB for 841,739 images (31.8 KiB per image), 132.29 GB with the retained native volumes, and about 2 TB at peak for an independent rebuild. Reproducibility is the indicator on which the corpus is most clearly distinguished, because every parameter, seed and exclusion decision is recorded, the code is archived under an open licence, partition integrity is verified on the written files, and the entire corpus can be regenerated from the original archives.

B. Implications for Plausibility-Constrained Synthesis

The results of Section III-F bear directly on the feasibility of the plausibility loss motivating this work. The requirement that the reference statistic be non-degenerate is not automatically satisfied: had the four-class clustering collapsed, for example by assigning the overwhelming majority of pixels in every chest radiograph to a single background class, the resulting mean vector would have been near-constant across images, and the corresponding penalty would have carried essentially no gradient. The measured distributions show this did not occur in either modality, and the negative off-diagonal covariance structure observed in both further indicates that the classes vary jointly in an interpretable way rather than independently.

The form the penalty may take is constrained by a property that the measured statistics make explicit. Because the compositional vector is confined to the simplex, both covariance matrices are singular by construction, a property directly visible in their zero row sums, and neither therefore admits an inverse. A penalty formulated as a conventional Mahalanobis distance against the raw covariance is consequently undefined, and must instead be constructed either through the Moore-Penrose pseudo-inverse or after an isometric log-ratio transform of the four fractions to three unconstrained coordinates. Either route preserves the anti-correlated structure that motivates using the covariance at all, which a per-class Euclidean penalty would discard. The marked difference in dispersion between the two modalities, roughly an order of magnitude in variance, further implies that a single penalty weight applied across modalities would

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Table 8. Comparison of the IntrinsicX corpus with previously published biomedical imaging datasets. Datasets are compared in terms of imaging modalities, two-dimensional image count, spatial resolution, acquisition-site coverage, patient-level partitioning, and availability of compositional reference statistics. CXR: chest X-ray; CT: computed tomography; MRI: magnetic resonance imaging; US: ultrasound; OCT: optical coherence tomography; EM: electron microscopy; n.r.: not reported. A dash indicates that the property was not assessed.

Resource, year [ref]	Content and resolution	Limitations relative to this work	Reference statistics
MedMNIST v2, 2023 [27]	12 modalities incl. CXR, CT, US, OCT, EM; 708,069 images at 28×28	Uniform low resolution suited to lightweight classification benchmarking; the spatial detail needed for synthesis is discarded	No
RadImageNet, 2022 [28]	CT, MRI, ultrasound; approx. 1.35M images at native (variable) resolution; patient-stratified splits	Designed for transfer-learning pretraining; no chest radiography	No
NIH ChestX-ray14, 2017 [21]	CXR; 112,120 images at 1,024×1,024; 1 site	Single modality and single institution; patient-level splits not reported	No
CheXpert, 2019 [29]	CXR; 224,316 images at native resolution; 1 site	Single modality and single institution; no harmonized multi-modal corpus	No
MIMIC-CXR, 2019 [30]	CXR; 377,110 images at native resolution; 1 site	Single modality and single institution; no harmonized multi-modal corpus	No
PadChest, 2020 [31]	CXR; more than 160,000 images at native resolution; 1 site	Single modality and single institution; no harmonized multi-modal corpus	No
OpenBHB, 2022 [26]	Brain MRI; 3D volumes of 182×218×182; 57 sites (training partition)	Distributed as three-dimensional quasi-raw volumes; no two-dimensional corpus	No
IntrinsicX (this work)	**CXR, CT, brain MRI from 5 archives; 841,739 images at 256×256; 57 OpenBHB sites plus 4 single-site sources**	**Patient-level 70/15/15 partition, measured on the written files**	**Yes (MRI, CXR)**

not be appropriate, and that the weight must be calibrated per modality against the reference reported in Table 4.

The absence of a computed tomography reference is a genuine constraint rather than an omission, and it shapes what the companion modelling work can attempt. A plausibility-constrained generator can be trained for brain magnetic resonance imaging and chest radiography with a reference-based penalty, but computed tomography synthesis from this corpus must either proceed unconstrained, borrow a constraint from an external lesion-free computed tomography cohort, or derive one from a lesion-free subset identified post hoc, each of which carries its own assumptions. A single-scanner cohort of 100 clinically healthy adults has recently been deposited [64], but it provides radiomic features rather than images and could therefore inform, but not replace, an image-level reference.

C. Reusability Beyond Generative Modelling

Several properties have value independent of the synthesis application. The harmonization spares reusers the re-implementation of per-source DICOM and NIfTI parsing, Hounsfield-Unit correction, windowing and normalization, which is the majority of the engineering in any multi-source imaging study and the stage at which silent errors are most easily introduced. The verified patient-level partitions support leakage-free benchmarking; the preserved site tags support site-stratified or site-held-out evaluation; and the retained per-source label taxonomies keep the corpus usable for classification and detection work unrelated to generative modelling.

For downstream applications other than synthesis, three practical consequences follow from the measured properties. First, pooled training is dominated by OpenBHB, which holds 51.7 per cent of the images, and within it by one site with 26.8 per cent of the subjects, so site-aware sampling and stratified evaluation are needed. Second, the saturation of 4.872 and 2.992 per cent of foreground pixels in BraTS-Africa and OpenBHB means that intensity-sensitive quantification should use the native volumes retained as the source of truth and not the delivered 8-bit slices. Third, the four-to-one pixel reduction of LIDC-IDRI (structural similarity 0.9233) suits models of global appearance and composition but not tasks that depend on fine structure, for which the native 512x512 data are preferable.

D. Limitations and Sources of Bias

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Ten limitations should be stated explicitly. First, LIDC-IDRI is included without nodule or malignancy labels, because only the image series and not the separate XML annotation package were retrieved; the LIDC-IDRI images in this corpus are therefore usable for adversarial or plausibility-constrained training but not for training an auxiliary lesion classifier. Second, no computed tomography reference statistic is provided, since neither computed tomography source contains a lesion-free subset, both being lesion- or nodule-focused by design. Third, the label taxonomy is not unified across sources, and a user requiring a single cross-modality taxonomy must construct it. Fourth, distribution is partitioned by size: the validation and test partitions and both reference statistic files are hosted directly in the public repository, the training partitions are distributed through a permanent link included in the deposit, and the native volumetric data are not distributed but remain exactly reproducible from the original public archives using the documented pipeline. Fifth, the magnetic resonance reference statistic aggregates slices across the full axial extent of the head without conditioning on slice position, and a reference stratified by normalized axial position would be a stronger, though considerably more elaborate formulation.

Sixth, whole-slice percentile normalization saturates 4.872 per cent of foreground pixels in BraTS-Africa and 2.992 per cent in OpenBHB, as reported in Section III-I; computing the percentiles over the foreground alone would reduce this by an order of magnitude and is the recommended change for a future version of the corpus, which we have not applied here because it would alter every reported count and require a new deposit version. Seventh, the lung window applied to LIDC-IDRI saturates 2.81 per cent of foreground pixels at the upper bound, so bone and dense soft tissue are not recoverable from those slices, which suit pulmonary parenchyma and not mediastinal or skeletal analysis. Eighth, the reference subsets are labelled rather than verified: OpenBHB subjects carry a cohort-level control designation assigned at recruitment, and NIH "No Finding" labels were derived by natural-language processing of radiology reports rather than by expert review of the images, so both describe anatomy labelled as unremarkable rather than anatomy confirmed as normal. The magnetic resonance reference cohort is also young, with a mean age of 25.2 years, and should not be assumed representative of older anatomy. Ninth, we did not obtain expert radiological assessment of the harmonized corpus; the quality-control measures of Section III-J are quantitative substitutes for such review, not replacements for it. Tenth, resampling fidelity was quantified only for the three volumetric sources; the round trip of Section III-J was not run for NIH ChestX-ray14 and DeepLesion, so the cost of reducing those natively two-dimensional images to 256x256 is not reported.

Four sources of bias bear on how the corpus should be used. Selection bias: the reference anatomy comes from a young control cohort (OpenBHB, mean age 25.2 years) and from images labelled by report mining (NIH ChestX-ray14), and the archives span different populations and scanners, so no single demographic or protocol is represented, and the reference statistics describe these cohorts only. Sampling bias: OpenBHB provides 51.7 per cent of the images and one site 26.8 per cent of its subjects, so unweighted pooling favours that cohort. Label bias: reference subsets are labelled rather than verified, so unrecognized pathology would shift the reference composition by an amount that this study cannot quantify. Procedural bias: intensity-only clusters conflate tissue classes with acquisition contrast, whole-slice normalization saturates the brightest tissue by construction (4.872 and 2.992 per cent), and the magnetic resonance statistic pools all axial levels. Site-stratified evaluation, the site-held-out folds, foreground-based normalization in a later version and expert review of the clusters are the corresponding mitigations.

V. Conclusion

The objective of this study was to design, implement and validate the data-engineering stage required for plausibility-constrained medical image synthesis: a documented pipeline that converts five independently governed public archives (NIH ChestX-ray14, NIH DeepLesion, BraTS-Africa, LIDC-IDRI and OpenBHB) into one harmonized, leakage-free corpus with compositional reference statistics, and to measure the fidelity cost of doing so. The pipeline is specified by 21 numbered equations and one algorithm and is regenerable from the original archives.

It produced 841,739 harmonized 256x256 images occupying 25.49 GB (132.29 GB with the retained native volumes) from 57 acquisition sites, in a projected 40.33 hours on a two-vCPU machine without a graphics processing unit. Partitions were verified at the patient level, and the realized image-level ratios lie within 2.07 percentage points of the 70/15/15 target for the volumetric sources, the largest deviation being 72.07 per cent training share for LIDC-IDRI. Resampling preserved structural similarity of 0.9233 for LIDC-IDRI and 0.9923 and 0.9917 for BraTS-Africa and OpenBHB, with peak signal-to-noise ratios of 32.1 to 39.5 dB and foreground entropy of 6.77 to 7.77 bits; the price of whole-slice percentile normalization is saturation of 4.872 per cent of BraTS-Africa and 2.992 per cent of OpenBHB foreground pixels, which moves the reference composition by at most 0.016. The reference distributions are well formed: the mean vectors sum to unity, no cluster collapsed, the covariance rows sum to zero to numerical precision,

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

the four-class partition is reproducible (Adjusted Rand Index 0.988 for magnetic resonance imaging and 0.910 for chest radiography), and the subject-level 95 per cent intervals of the magnetic resonance fractions have half-widths of at most 0.0041. A reference-based penalty can therefore be formulated on this corpus through the pseudo-inverse or log-ratio route that the singular covariance requires, but this study does not show that such a penalty improves generative synthesis.

The scientific contribution is a fully specified, auditable and regenerable harmonization procedure whose limits are measured and openly reported; its practical significance is that other groups can reuse the deposited corpus, 25.49 GB in two dimensions, or regenerate it, without re-engineering the preprocessing. All claims concern the corpus: evaluation of IntrinsicX-GAN against unconstrained baselines is reserved for the companion article in preparation. Future work must validate the four intensity clusters against anatomical segmentation or expert annotation, supply a computed tomography reference statistic from an external or post hoc identified lesion-free cohort, adopt foreground-based normalization in a new deposit version, extend evaluation to the withheld OpenBHB external sites, obtain expert radiological review, and finally evaluate generative models trained on the corpus for fidelity and anatomical admissibility.

Declarations

Ethics Approval and Consent to Participate

Ethical approval and informed consent were not required for this study because it involved exclusively secondary computational analysis of previously de-identified, publicly available medical imaging datasets, with no new human participant recruitment, intervention, or access to personally identifiable information. The original datasets were released under their respective ethical approvals and data-use conditions [21]-[24], [26].

Consent for Publication

Not applicable, as the manuscript contains no identifiable personal information, patient images, or individual case details requiring publication consent.

Funding

No direct financial funding was received for this study.

Competing Interests

The authors declare that they have no financial or non-financial competing interests that could have influenced the research, analysis, interpretation, or publication of this manuscript.

Data Availability

The harmonized medical imaging corpus and associated reference statistics are available through Mendeley Data at https://doi.org/10.17632/6m4kb3vffg.3 [65]. The validation and test partitions and reference-statistic files are deposited directly, while the training partitions are accessible through the link provided in the repository record. The original imaging archives remain subject to their respective access conditions.

Code Availability

The source code, preprocessing procedures, experimental parameters, and reproducibility documentation are available at https://github.com/LAHARRAF/IntrinsicX-GAN and archived at https://doi.org/10.5281/zenodo.22262186 under the MIT License.

Authors' Contributions

Said Laharraf contributed to conceptualization, methodology, software development, data curation, formal analysis, validation, visualization, and original manuscript preparation. Khaoula Addakiri contributed to supervision and manuscript review and editing. Both authors reviewed and approved the final manuscript and accepted responsibility for the integrity and accuracy of the work.

Acknowledgments

The authors acknowledge the National Center for Scientific and Technical Research (CNRST), Morocco, for institutional support through the PhD-Associate Scholarship (PASS) programme and thank the original dataset contributors for providing publicly accessible medical imaging resources.

Declaration of Artificial Intelligence (AI) Use

Generative artificial intelligence (AI) tools were used solely for language editing, grammatical correction, and stylistic refinement of the manuscript; no AI tools were used to generate or fabricate research data, experimental results, figures, or scientific interpretations, and the authors independently reviewed and verified all AI-assisted content and accept full responsibility for the accuracy, originality, and integrity of the published work.

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 27, 2014, pp. 2672-2680.

[2] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 1125-1134, doi: 10.1109/CVPR.2017.632.

[3] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2223-2232, doi: 10.1109/ICCV.2017.244.

[4] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

review," Medical Image Analysis, vol. 58, Art. no. 101552, 2019, doi: 10.1016/j.media.2019.101552.

[5] S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, and A. Mukhopadhyay, "GANs for medical image analysis," Artificial Intelligence in Medicine, vol. 109, Art. no. 101938, 2020, doi: 10.1016/j.artmed.2020.101938.

[6] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," Neurocomputing, vol. 321, pp. 321-331, 2018, doi: 10.1016/j.neucom.2018.09.013.

[7] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, and B. Yang, "MedGAN: Medical image translation using GANs," Computerized Medical Imaging and Graphics, vol. 79, Art. no. 101684, 2020, doi: 10.1016/j.compmedimag.2019.101684.

[8] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with deep convolutional adversarial networks," IEEE Transactions on Biomedical Engineering, vol. 65, no. 12, pp. 2720-2730, 2018, doi: 10.1109/TBME.2018.2814538.

[9] S. U. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur, "Image synthesis in multi-contrast MRI with conditional generative adversarial networks," IEEE Transactions on Medical Imaging, vol. 38, no. 10, pp. 2375-2388, 2019, doi: 10.1109/TMI.2019.2901750.

[10] H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in Simulation and Synthesis in Medical Imaging (SASHIMI), Lecture Notes in Computer Science, vol. 11037, Cham, Switzerland: Springer, 2018, pp. 1-11, doi: 10.1007/978-3-030-00536-8_1.

[11] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 6840-6851.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 10684-10695, doi: 10.1109/CVPR52688.2022.01042.

[13] W. H. L. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. da Costa, V. Fernandez, P. Nachev, S. Ourselin, and M. J. Cardoso, "Brain imaging generation with latent diffusion models," in Deep Generative Models (DGM4MICCAI), Lecture Notes in Computer Science, vol. 13609, Cham, Switzerland: Springer, 2022, pp. 117-126, doi: 10.1007/978-3-031-18576-2_12.

[14] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof, "Diffusion models in medical imaging: A comprehensive survey," Medical Image Analysis, vol. 88, Art. no. 102846, 2023, doi: 10.1016/j.media.2023.102846.

[15] G. Müller-Franzes, J. M. Niehues, F. Khader, et al., "A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis," Scientific Reports, vol. 13, Art. no. 12098, 2023, doi: 10.1038/s41598-023-39278-0.

[16] A. DuMont Schütte, J. Hetzel, S. Gatidis, T. Hepp, B. Dietz, S. Bauer, and P. Schwab, "Overcoming barriers to data sharing with medical image generation: A comprehensive evaluation," npj Digital Medicine, vol. 4, Art. no. 141, 2021, doi: 10.1038/s41746-021-00507-3.

[17] Y. Chu, C. Yang, G. Luo, Z. Qiu, and X. Gao, "Anatomic-constrained medical image synthesis via physiological density sampling," in Medical Image Computing and Computer Assisted Intervention (MICCAI), Lecture Notes in Computer Science, vol. 15011, Cham, Switzerland: Springer, 2024, pp. 69-79, doi: 10.1007/978-3-031-72120-5_7.

[18] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017, pp. 6626-6637.

[19] Y. Skandarani, P.-M. Jodoin, and A. Lalande, "GANs for medical image synthesis: An empirical study," Journal of Imaging, vol. 9, no. 3, Art. no. 69, 2023, doi: 10.3390/jimaging9030069.

[20] Y. Chen, X.-H. Yang, Z. Wei, A. A. Heidari, N. Zheng, Z. Li, H. Chen, H. Hu, Q. Zhou, and Q. Guan, "Generative adversarial networks in medical image augmentation: A review," Computers in Biology and Medicine, vol. 144, Art. no. 105382, 2022, doi: 10.1016/j.compbiomed.2022.105382.

[21] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, 2017, pp. 3462-3471, doi: 10.1109/CVPR.2017.369.

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

[22] K. Yan, X. Wang, L. Lu, and R. M. Summers, "DeepLesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning," Journal of Medical Imaging, vol. 5, no. 3, Art. no. 036501, 2018, doi: 10.1117/1.JMI.5.3.036501.

[23] M. Adewole, J. D. Rudie, A. Gbadamosi, et al., "Brain tumor segmentation in Sub-Saharan Africa patient population: The BraTS-Africa challenge," Neuro-Oncology Advances, vol. 8, no. 1, Art. no. vdag082, 2026, doi: 10.1093/noajnl/vdag082.

[24] S. G. Armato III, G. McLennan, L. Bidaut, et al., "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans," Medical Physics, vol. 38, no. 2, pp. 915-931, 2011, doi: 10.1118/1.3528204.

[25] K. Clark, B. Vendt, K. Smith, et al., "The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository," Journal of Digital Imaging, vol. 26, no. 6, pp. 1045-1057, 2013, doi: 10.1007/s10278-013-9622-7.

[26] B. Dufumier, A. Grigis, J. Victor, C. Ambroise, V. Frouin, and E. Duchesnay, "OpenBHB: A large-scale multi-site brain MRI data-set for age prediction and debiasing," NeuroImage, vol. 263, Art. no. 119637, 2022, doi: 10.1016/j.neuroimage.2022.119637.

[27] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification," Scientific Data, vol. 10, Art. no. 41, 2023, doi: 10.1038/s41597-022-01721-8.

[28] X. Mei, Z. Liu, P. M. Robson, et al., "RadImageNet: An open radiologic deep learning research dataset for effective transfer learning," Radiology: Artificial Intelligence, vol. 4, no. 5, Art. no. e210315, 2022, doi: 10.1148/ryai.210315.

[29] J. Irvin, P. Rajpurkar, M. Ko, et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in Proc. AAAI Conf. Artificial Intelligence, vol. 33, no. 1, 2019, pp. 590-597, doi: 10.1609/aaai.v33i01.3301590.

[30] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," Scientific Data, vol. 6, Art. no. 317, 2019, doi: 10.1038/s41597-019-0322-0.

[31] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vaya, "PadChest: A large chest x-ray image dataset with multi-label annotated reports," Medical Image Analysis, vol. 66, Art. no. 101797, 2020, doi: 10.1016/j.media.2020.101797.

[32] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," Biostatistics, vol. 8, no. 1, pp. 118-127, 2007, doi: 10.1093/biostatistics/kxj037.

[33] J.-P. Fortin, N. Cullen, Y. I. Sheline, et al., "Harmonization of cortical thickness measurements across scanners and sites," NeuroImage, vol. 167, pp. 104-120, 2018, doi: 10.1016/j.neuroimage.2017.11.024.

[34] R. Pomponio, G. Erus, M. Habes, et al., "Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan," NeuroImage, vol. 208, Art. no. 116450, 2020, doi: 10.1016/j.neuroimage.2019.116450.

[35] N. K. Dinsdale, M. Jenkinson, and A. I. L. Namburete, "Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal," NeuroImage, vol. 228, Art. no. 117689, 2021, doi: 10.1016/j.neuroimage.2020.117689.

[36] B. Glocker, R. Robinson, D. C. Castro, Q. Dou, and E. Konukoglu, "Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects," arXiv:1910.04597, 2019, doi: 10.48550/arXiv.1910.04597.

[37] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," Nature Methods, vol. 18, no. 2, pp. 203-211, 2021, doi: 10.1038/s41592-020-01008-z.

[38] M. J. Cardoso, W. Li, R. Brown, et al., "MONAI: An open-source framework for deep learning in healthcare," arXiv:2211.02701, 2022, doi: 10.48550/arXiv.2211.02701.

[39] F. Pérez-García, R. Sparks, and S. Ourselin, "TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," Computer Methods and Programs in Biomedicine, vol. 208, Art. no. 106236, 2021, doi: 10.1016/j.cmpb.2021.106236.

[40] M. Roberts, D. Driggs, M. Thorpe, et al., "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," Nature Machine Intelligence, vol. 3, pp. 199-217, 2021, doi: 10.1038/s42256-021-00307-0.

[41] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: Methodological failures and recommendations for the future," npj Digital Medicine, vol. 5, Art. no. 48, 2022, doi: 10.1038/s41746-022-00592-y.

[42] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," ACM Transactions on Knowledge Discovery from Data, vol. 6, no. 4, Art. no. 15, 2012, doi: 10.1145/2382577.2382579.

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

[43] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, et al., "The FAIR Guiding Principles for scientific data management and stewardship," Scientific Data, vol. 3, Art. no. 160018, 2016, doi: 10.1038/sdata.2016.18.

[44] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren, "Preparing medical imaging data for machine learning," Radiology, vol. 295, no. 1, pp. 4-15, 2020, doi: 10.1148/radiol.2020192224.

[45] G. N. Hounsfield, "Computerized transverse axial scanning (tomography): Part 1. Description of system," British Journal of Radiology, vol. 46, no. 552, pp. 1016-1022, 1973, doi: 10.1259/0007-1285-46-552-1016.

[46] W. D. Bidgood Jr., S. C. Horii, F. W. Prior, and D. E. Van Syckle, "Understanding and using DICOM, the data interchange standard for biomedical imaging," J. Am. Med. Inform. Assoc., vol. 4, no. 3, pp. 199-212, 1997, doi: 10.1136/jamia.1997.0040199.

[47] R. C. Gonzalez and R. E. Woods, Digital Image Processing, 4th ed. New York, NY, USA: Pearson, 2018.

[48] L. G. Nyúl and J. K. Udupa, "On standardizing the MR image intensity scale," Magnetic Resonance in Medicine, vol. 42, no. 6, pp. 1072-1081, 1999, doi: 10.1002/(SICI)1522-2594(199912)42:6<1072::AID-MRM11>3.0.CO;2-M.

[49] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA), 2007, pp. 1027-1035.

[50] S. P. Lloyd, "Least squares quantization in PCM," IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129-137, 1982, doi: 10.1109/TIT.1982.1056489.

[51] J. Aitchison, The Statistical Analysis of Compositional Data. London, U.K.: Chapman and Hall, 1986.

[52] R. Penrose, "A generalized inverse for matrices," Mathematical Proceedings of the Cambridge Philosophical Society, vol. 51, no. 3, pp. 406-413, 1955, doi: 10.1017/S0305004100030401.

[53] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal, "Isometric logratio transformations for compositional data analysis," Mathematical Geology, vol. 35, no. 3, pp. 279-300, 2003, doi: 10.1023/A:1023818214614.

[54] B. Efron, "Bootstrap methods: Another look at the jackknife," The Annals of Statistics, vol. 7, no. 1, pp. 1-26, 1979, doi: 10.1214/aos/1176344552.

[55] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. New York, NY, USA: Chapman and Hall, 1993.

[56] L. Hubert and P. Arabie, "Comparing partitions," Journal of Classification, vol. 2, no. 1, pp. 193-218, 1985, doi: 10.1007/BF01908075.

[57] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53-65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[58] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," Communications in Statistics, vol. 3, no. 1, pp. 1-27, 1974, doi: 10.1080/03610927408827101.

[59] D. L. Davies and D. W. Bouldin, "A cluster separation measure," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224-227, 1979, doi: 10.1109/TPAMI.1979.4766909.

[60] C. E. Shannon, "A mathematical theory of communication," Bell System Technical Journal, vol. 27, no. 3, pp. 379-423, 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.

[61] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, 2004, doi: 10.1109/TIP.2003.819861.

[62] D. Mason, "SU-E-T-33: Pydicom: An open source DICOM library," Medical Physics, vol. 38, no. 6, p. 3493, 2011, doi: 10.1118/1.3611983.

[63] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011. [Online]. Available: https://jmlr.org/papers/v12/pedregosa11a.html

[64] V. Bedei, M. Ostrovskyy, N. Dey, T. Kotyk, and R. S. Sherratt, "Radiomics dataset from chest CT of clinically healthy adults," Data in Brief, vol. 65, Art. no. 112556, 2026, doi: 10.1016/j.dib.2026.112556.

[65] S. Laharraf and K. Addakiri, "Harmonized multi-modal medical imaging train/validation/test corpus and healthy-tissue reference statistics (chest X-ray, CT, brain MRI)," Mendeley Data, ver. 3, 2026, doi: 10.17632/6m4kb3vffg.3.

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Author Biography



Said Laharraf is a doctoral researcher in biomedical engineering at the Higher Institute of Health Sciences (ISSS), Hassan I University, Settat, Morocco, supported by the CNRST PhD-Associate Scholarship (PASS) programme. His doctoral project addresses machine learning for the early diagnosis of chronic diseases, with particular attention to model explainability. His current research concerns generative adversarial networks for medical image synthesis, with an emphasis on intrinsically explainable and anatomically constrained generation, multi-source harmonization of radiography, computed tomography and magnetic resonance imaging data, and reproducible imaging pipelines.



Khaoula Addakiri is a Professor at the Higher Institute of Health Sciences (ISSS), Hassan I University, Settat, Morocco. She teaches and supervises doctoral students in health sciences and medical informatics. Her research interests include applying artificial intelligence in healthcare, especially explainable machine learning for clinical decision support, generative models for medical imaging, and reliable, reusable health data resources. She is currently supervising PhD research on explainable generative adversarial networks for medical image synthesis. In this study, she supervised the research and helped review and edit the manuscript.

Manuscript received 3 September 2026; Accepted 4 November 2026; Available online 27 November 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v9i1.2070
Copyright © 2026 by the authors. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).