BIOLOGY · APPLIED COMPUTATIONAL SCIENCE

THE AI DATA PIPELINE

Preprocess · Train · Evaluate

Machine-learning workflows in the biological and biomedical imaging sciences

1 INTRODUCTION: WHY BIOLOGY NEEDED A PIPELINE

Modern biology is no longer limited by how much data it can *collect*, but by how much it can *interpret*. One **whole-slide image (WSI)** of a histology section at 40× magnification holds 10^9–10^{10} pixels. A routine **MRI** study produces several hundred slices per patient. A high-content screen images tens of thousands of wells overnight. No human can read this exhaustively, consistently or quickly.

Machine learning (ML) — the subfield of **artificial intelligence (AI)** in which a model learns a mapping from data rather than from hand-written rules — converts that signal into biological measurement. But a model is not a device you point at data; it is the *output* of a disciplined, reproducible sequence of operations. That sequence is the **data pipeline**, conventionally divided into three stages:

1. **Preprocessing** — cleaning, standardising and structuring raw data so the model sees biological signal rather than measurement artefact.
2. **Training** — fitting model parameters to labelled data by iterative minimisation of a loss function.
3. **Evaluation** — quantifying, on unseen data, how well the model performs, how it fails, and whether the result generalises beyond the laboratory that produced it.

The pipeline is **the scientific method rewritten computationally**. Preprocessing is *specimen preparation* — fixing, sectioning, staining, mounting, so every sample is comparable. Training is *the experiment* — a hypothesis repeatedly adjusted against observation. Evaluation is *peer review and replication* — the claim must survive an independent, unseen test. A pipeline missing any stage is not merely inefficient; it is scientifically invalid, exactly as an unblinded, uncontrolled bench experiment is invalid.

> **THE DEFINITION, STATED PRECISELY**
>
> An **AI data pipeline** is a reproducible, automated sequence of stages — *data acquisition → preprocessing → training → evaluation → deployment and monitoring* — that transforms raw biological measurements into a validated predictive model while controlling for bias, leakage and artefact at every step.
>
> Reproducibility and version control are what distinguish a pipeline from an ad-hoc script.

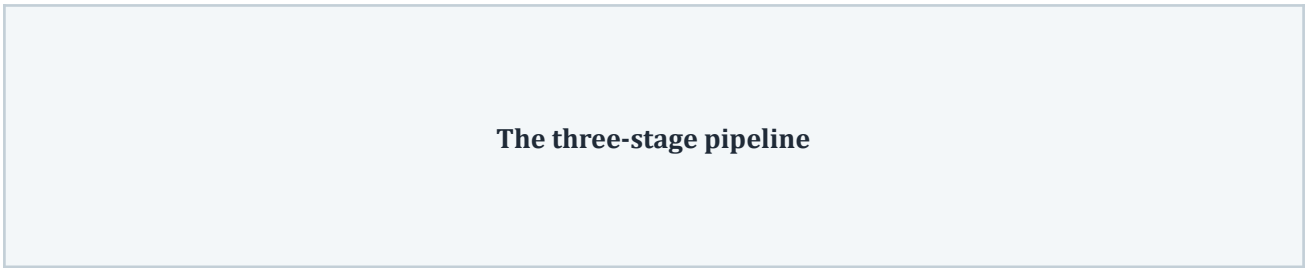

Figure 1. *Overview of the AI data pipeline in biology. Raw specimen data enters at the left, preprocessing standardises it, training fits the model, evaluation tests it on held-out data. The dashed return path is the iterative refinement loop that dominates real practice.*

1.1 The pipeline is a loop, not a line

Textbook diagrams draw an arrow travelling left to right; in practice it is a **cycle**. Poor evaluation results send you back to preprocessing (was normalisation wrong?) or to acquisition (do we need a second hospital?). This feedback path is **iterative refinement**, and honest reporting of how often it was traversed is required by modern reporting standards.

One direction of travel must **never** occur: information from the **test set** must not flow backwards into preprocessing or training. When it does, the result is **data leakage** — the single commonest cause of published biomedical AI results that fail to replicate (§3.6).

2 STAGE 0: PROBLEM FRAMING AND DATA ACQUISITION

Strictly, preprocessing is stage one — but a pipeline that begins there has already made its most consequential decisions silently. Two questions must be answered on paper first.

2.1 Framing the biological question as a learning task

The biological question must be translated into a canonical **learning task**. Choosing the wrong formulation is a design error no amount of training can repair.

Learning task	Model output	Biological example
Binary classification	One of two classes, with a probability.	Does this sentinel lymph node section contain metastatic carcinoma?
Multi-class classification	One of *k* mutually exclusive classes.	Classifying a white blood cell as neutrophil, eosinophil, basophil, lymphocyte or monocyte.
Multi-label classification	Any number of co-occurring labels.	A chest radiograph may show cardiomegaly *and* pleural effusion at once.
Semantic segmentation	A class label for every pixel or voxel.	Delineating tumour core, oedema and necrosis in a brain MRI.
Object detection	Bounding boxes plus classes.	Counting and localising mitotic figures in a histology field.
Regression	A continuous value.	Predicting left-ventricular ejection fraction from a cardiac MRI.
Survival analysis	A risk score or hazard.	Predicting recurrence-free survival from tumour morphology.

2.2 Sources of biological image data

Each modality imposes its own preprocessing obligations, and each artefact must be matched to the modality that produces it.

- **Histopathology (WSI)** — H&E-stained slides, scanned. Problems: enormous file size, **stain variability** between laboratories, tissue folds, out-of-focus regions, pen marks.
- **Radiology (CT, MRI, PET)** — volumetric, stored as **DICOM**. Problems: **anisotropic voxel spacing**, scanner-to-scanner intensity differences, MRI **bias field** (smooth intensity drift from coil inhomogeneity), motion artefact.
- **Projection radiography (X-ray)** — 2-D, high throughput. Problems: variable exposure, positioning, and **burned-in text or laterality markers** that models learn as shortcuts.
- **Optical and fluorescence microscopy** — uneven illumination, photobleaching, channel bleed-through, focal drift.
- **Electron microscopy / cryo-EM** — very low signal-to-noise ratio, section-to-section misalignment.

Figure 2. *The principal biological imaging modalities used to train AI models, each shown with the dominant artefact it contributes. Scale and artefact profile decide which preprocessing operations are mandatory.*

2.3 Ground truth: the quiet weakness of every biological dataset

A supervised model can never exceed the accuracy of its labels. In biology labels come from three sources, in descending reliability:

1. **Hard endpoints** — a genomic sequence, a culture result, documented five-year survival. Objective, but scarce and delayed.
2. **Expert annotation** — a pathologist outlining tumour, a radiologist marking a nodule. Available at scale but subject to **inter-observer variability**, quantified by **Cohen's kappa (κ)** for two raters or **Fleiss' kappa** for more. This figure sets a realistic ceiling on achievable model performance.

3. **Weak or automatic labels** — mined from radiology reports, or a diagnosis code attached to a whole slide. Cheap, plentiful, noisy. Gives rise to **weakly supervised learning** and to **multiple-instance learning (MIL)** (§4.7).

THE LABEL CEILING

A model reported at 95% accuracy against ground truth whose two annotators agree with each other only 88% of the time is not 95% accurate; it has partly learned one expert's idiosyncrasies.

Annotation quality, inter-observer agreement (κ) and label provenance must therefore be reported alongside model performance — a core requirement of the **CLAIM** checklist.

2.4 Ethics, consent and de-identification

Biological data describing identifiable people is regulated. Images must be **de-identified** before computation: DICOM headers carry patient name, date of birth and hospital identifiers, and head scans contain reconstructible facial surface geometry. **Anonymisation** removes identifiers irreversibly; **pseudonymisation** replaces them with a separately held key; **defacing** or skull-stripping removes facial geometry, because a 3-D render of a head scan is biometrically identifying.

The pipeline must also record **provenance** — institution, scanner, protocol, date. This is not administrative overhead: it is the only way to later diagnose **batch effects** and **domain shift** (§5.7).

Anatomy of a de-identification step

Figure 3. *A DICOM object separated into pixel data and header metadata, showing which header fields are stripped, which are retained for provenance, and where defacing acts on the pixel data.*

3 STAGE 1: PREPROCESSING

Preprocessing is the set of transformations that make every sample comparable, suppress artefact, and present biological signal in a learnable form. It routinely occupies 60–80% of the effort of an applied ML project, and it is where most published biomedical models are silently broken.

The governing principle is simple: **preprocessing should remove variation that is technical and preserve variation that is biological.** Justify every operation below against that sentence. Normalising away scanner brightness is correct; normalising away a genuine difference in nuclear density between tumour and normal tissue destroys the signal you wanted.

3.1 Quality control and curation

Quality control (QC) identifies and excludes unusable samples. It must be defined *a priori* and applied uniformly, exactly as exclusion criteria are pre-specified in a clinical trial.

- **Corrupt or incomplete files** — truncated DICOM series, missing slices, unreadable WSI pyramid levels.
- **Artefact rejection** — tissue folds, air bubbles, pen annotation, motion ghosting, metal or beam-hardening artefact in CT.
- **Focus and blur detection** — sharpness metrics such as variance of the Laplacian flag out-of-focus microscopy fields.
- **Tissue detection** — 70–90% of a WSI canvas is empty glass. **Otsu thresholding** in the saturation channel of HSV space cheaply separates tissue from background and avoids enormous wasted computation.
- **Outlier and duplicate detection** — duplicated images are a hidden leakage source; near-duplicate detection by perceptual hashing is standard.

> **KEY POINT**
>
> **QC criteria must be pre-specified and the number of exclusions reported with reasons.** A pipeline that silently drops difficult images inflates its own performance — **selection bias**, which CONSORT-style flow diagrams exist to expose.

3.2 Format conversion and structuring

Raw formats (DICOM, NIfTI, SVS, CZI, TIFF) are built for acquisition and archiving, not training throughput. Preprocessing converts them to efficient representations — **NIfTI** or compressed **HDF5/Zarr** for volumes, **tiles** for whole-slide images, packed binary shards for streaming. The **data schema** is fixed here: array shape, channel order, dtype, and the file-to-patient mapping. Convention is (N, C, H, W) for 2-D batches and (N, C, D, H, W) for volumes.

3.3 Intensity normalisation

Two scans of the same anatomy from two scanners can have entirely different numerical ranges. **Intensity normalisation** puts them on a common scale so the model is not forced to learn scanner identity.

Method	Operation	When used
Min–max scaling	$x' = (x - \min) / (\max - \min) \rightarrow [0, 1]$.	Bounded data; outlier-sensitive, so apply after percentile clipping.
Z-score standardisation	$x' = (x - \mu) / \sigma \rightarrow$ mean 0, SD 1.	The default for MRI and most deep-learning inputs.
Percentile clipping / windowing	Clip to the 1st–99th percentile, or a clinical CT window.	CT, where Hounsfield Units are calibrated and windowing mimics radiologist practice.
Histogram matching	Warp one image's histogram onto a reference.	Harmonising multi-site MRI cohorts.
N4 bias-field correction	Estimates and divides out smooth multiplicative intensity drift.	Essentially mandatory for quantitative MRI.

> **CRITICAL RULE — NORMALISATION STATISTICS COME FROM THE TRAINING SET ONLY**
>
> μ and σ must be computed on the **training split alone**, then applied unchanged to validation and test.

Method	Operation	When used
Computing them over the whole dataset before splitting leaks distributional information about the test set into training. This is the most common form of **data leakage**.		

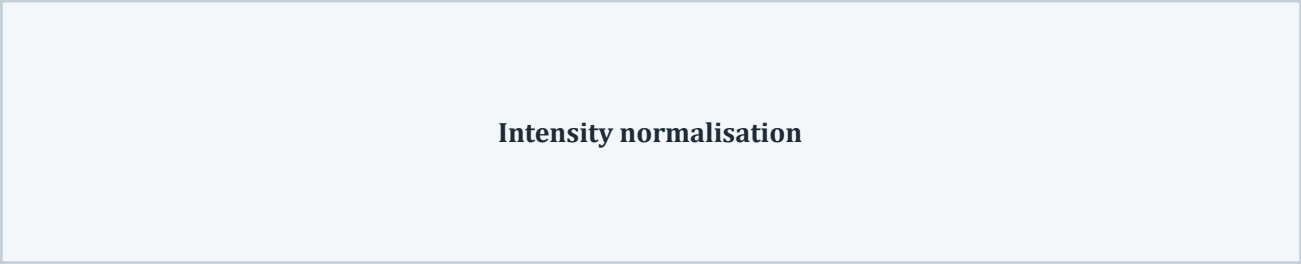

Figure 4. *Intensity histograms of the same anatomical region from two scanners, before and after z-score standardisation. Technical variation collapses; biological difference between tissue classes is preserved.*

Stain normalisation — the histopathology special case

H&E staining varies with reagent batch, section thickness, staining time and scanner colour profile, and a model trained in one laboratory frequently collapses on another's slides purely because of colour. **Stain normalisation** maps every slide's colour distribution onto a common reference.

- **Reinhard** — matches per-channel mean and standard deviation in the perceptually uniform LAB space. Fast, simple, blind to stain chemistry.
- **Macenko** — uses SVD of the optical density image to estimate the two **stain vectors** (haematoxylin, eosin) and re-mixes them to a reference. The standard baseline.
- **Vahadane / sparse NMF** — decomposes optical density into sparse stain concentration maps; better structure preservation, slower.
- **Stain augmentation** — the opposite philosophy: rather than removing colour variation, deliberately amplify it during training so the model becomes invariant to it. Increasingly preferred, because it produces robustness to colours never seen.

The underlying physics is the **Beer–Lambert law**: optical density is linearly proportional to stain concentration. Converting RGB to optical density, $OD = -\log_{10}(I / I_0)$, is what makes linear unmixing of the two stains valid — the physical basis on which the whole method rests.

Figure 5. *Three H&E fields from different laboratories before normalisation, and the same three after Macenko normalisation to a common reference. Colour variance is removed; nuclear and stromal morphology is unchanged.*

3.4 Spatial standardisation

- **Resampling to isotropic spacing** — an MRI at 0.5 × 0.5 × 5.0 mm has voxels ten times longer in one axis. Resampling to 1 × 1 × 1 mm makes a convolutional kernel cover the same physical distance in every direction. Use **B-spline or linear interpolation for intensity images but nearest-neighbour for label masks**, or interpolation invents non-existent class values.
- **Registration** — aligning images so a coordinate refers to the same anatomy across subjects. **Rigid** allows rotation and translation; **affine** adds scaling and shear; **deformable** allows local warping. Multi-modal registration (PET to CT) uses **mutual information** as the similarity metric.

— **ROI cropping** — restricting to the organ or tissue, removing background that would otherwise dominate the loss.

— **Resizing versus padding** — resizing changes apparent object size; padding preserves it. Where **absolute size is diagnostic** (nodule diameter, nuclear area), pad to a fixed canvas rather than resize.

— **Patching / tiling** — a 100 000 × 100 000 px WSI cannot enter a GPU. It is cut into 256 × 256 or 512 × 512 tiles at a defined magnification, usually with **overlap** so objects on a boundary are not lost; predictions are stitched back and overlaps averaged.

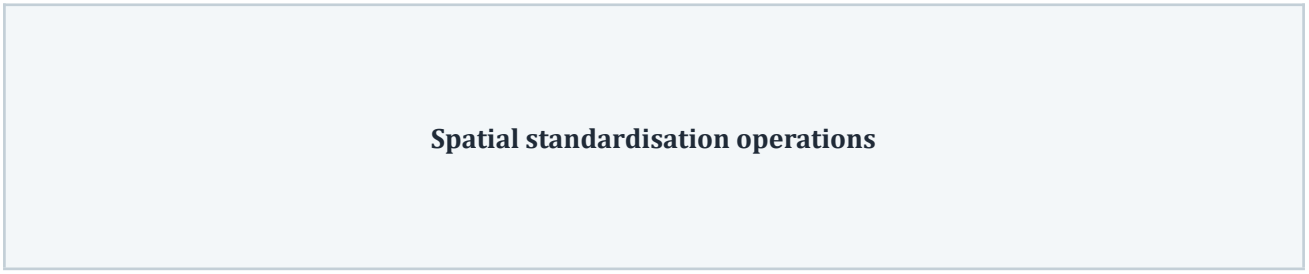

Figure 6. *Left: anisotropic voxels resampled to an isotropic grid. Centre: rigid versus deformable registration. Right: a whole-slide image decomposed into overlapping tiles, with the stitching overlap indicated.*

3.5 Feature engineering versus representation learning

Classical biological ML used **hand-crafted features** — nuclear area, circularity, chromatin texture, Haralick descriptors, **radiomics** features from a segmented tumour. These are interpretable and work with small cohorts, which is why radiomics persists where sample sizes number in the hundreds. **Deep learning** instead performs **representation learning**: convolutional layers discover features directly from pixels. This scales far better but needs more data and yields less interpretable features. Current practice often combines the two — deep features extracted, then fed to a classical classifier when the labelled cohort is small.

3.6 Data splitting — and the leakage traps that ruin biology papers

— **Training set** (60–80%) — parameters are fitted on this.

— **Validation set** (10–20%) — used *during* training to select hyperparameters, decide when to stop, and choose between architectures. Also called the *development* or *tuning* set.

— **Test set** (10–20%) — touched **once**, at the very end, to produce the reported performance. Also called the *hold-out* set.

In biology, *how* you split matters more than the ratio, because biological samples are rarely independent. The correct unit of splitting is almost never the image.

Leakage mechanism	How it happens	Correct control
Patient-level	Two tiles from one slide, or two slices from one volume, land in different splits. The model recognises the patient, not the disease.	Split by **patient identifier** — *grouped* or *patient-wise* splitting.
Site / batch	One hospital contributes most positives; the model learns the scanner signature.	Stratify by site; reserve at least one site entirely for **external validation**.
Temporal	Future data used to predict the past, or a protocol change straddling the split.	Split chronologically for any deployment claim.
Preprocessing	Normalisation statistics or feature selection computed on all data before splitting.	Fit every transform on the training fold; apply to the others.
Augmentation	An augmented copy of a training image appears in validation.	Augment strictly after splitting, training fold only.
Duplicate	The same image present twice under different filenames.	De-duplicate by hashing before splitting.

Stratification must also be applied: each split preserves the class proportions of the whole dataset, and ideally the distribution of age, sex, scanner and site. A rare-disease test set that happens to contain three positives cannot support a meaningful confidence interval.

Figure 7. *Above: naive random splitting of patches distributes tiles from one patient across training and test, inflating performance. Below: grouped patient-level splitting keeps every sample from one individual in a single partition.*

3.7 Data augmentation

Data augmentation enlarges the training set with label-preserving transformations, acting as a powerful **regulariser** and encoding known **invariances** of the domain. It matters most in biology, where labelled cohorts are small and expensive. The test of a valid augmentation is simple: *would this transformed image occur in reality, and would its label still be correct?*

- **Spatial** — rotation, flips, translation, scaling, elastic deformation. Histology and microscopy are **rotationally invariant** (a cell has no canonical orientation), so full 360° rotation is valid. A chest radiograph is **not** left–right invariant: horizontal flipping creates *situs inversus* and can invert a laterality-dependent finding.
- **Photometric** — brightness, contrast, gamma; for histology, **stain jitter** in haematoxylin–eosin concentration space.
- **Noise and blur** — Gaussian noise, Rician noise for MRI, mild defocus; simulates lower-quality acquisitions and improves robustness.
- **Occlusion-based** — random erasing / Cutout, forcing distributed rather than single-region evidence.
- **Mixing-based** — **Mixup** blends two images and their labels; **CutMix** pastes a patch of one into another. Both improve calibration, though anatomically implausible outputs make them contested in clinical imaging.
- **Generative** — GAN- or diffusion-synthesised samples to enrich rare classes. Synthetic data must never enter the test set, and the risk of learning generator artefacts must be reported.

AUGMENTATION DISCIPLINE

Augmentation applies **only to the training set**, and normally *on the fly* (a fresh random transform every epoch) rather than written to disk. Validation and test data get the deterministic preprocessing only.

The single exception is **test-time augmentation (TTA)** — predicting several transformed versions of a test image and averaging to reduce variance. This is a legitimate inference technique, not leakage, but it must be declared.

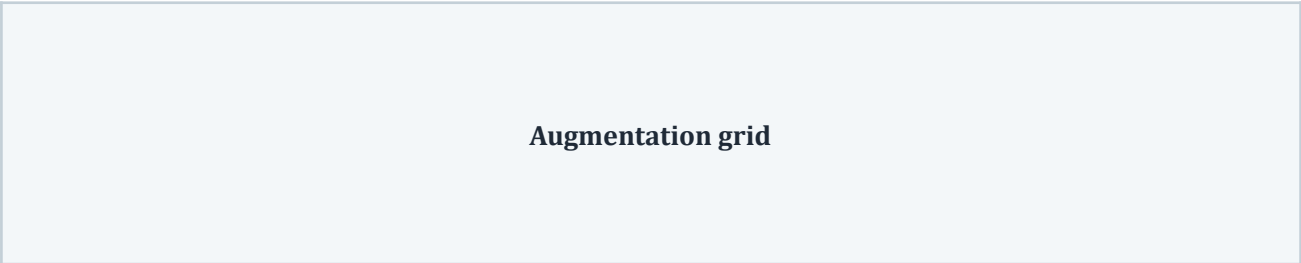

Figure 8. *One specimen field and eight label-preserving transformations: rotation, flip, elastic deformation, scale, brightness and contrast shift, stain jitter, additive noise, random erasing. The label is invariant across all nine panels.*

3.8 Handling class imbalance

Biological datasets are almost always imbalanced — pathology is rare relative to normal tissue, and in screening cohorts positives may be under 1%. A model predicting "normal" for everything then reaches 99% accuracy while being clinically worthless: the **accuracy paradox**.

- **Resampling** — **oversampling** the minority class or **undersampling** the majority. **SMOTE** synthesises minority points by interpolating between neighbours; suited to tabular features, poorly suited to raw images.
- **Class weighting** — weighting each class's loss contribution inversely to its frequency, so a rare positive costs more to get wrong.
- **Focal loss** — down-weights easy, well-classified examples so gradient concentrates on hard cases; now standard in medical segmentation.
- **Hard-negative mining** — retraining on the false positives produced by an earlier round; classic in mitosis and nodule detection.

— **Threshold adjustment** — leave training unchanged and move the decision threshold at inference to reach a required sensitivity, choosing the operating point from the validation ROC curve.

Critically, **imbalance is corrected only in the training set**. Validation and test must retain the true **prevalence** of the population, because metrics such as positive predictive value depend directly on prevalence and become meaningless on an artificially balanced test set.

4 STAGE 2: TRAINING

Training adjusts a model's **parameters** so its predictions on labelled data move closer to ground truth. Everything reduces to one idea: define a measure of wrongness, then move the parameters in the direction that reduces it.

4.1 The learning loop

Almost all supervised deep learning repeats this five-step cycle for every **mini-batch**:

1. **Forward pass** — a batch of preprocessed inputs passes through the network, producing predictions $\hat{y}$.
2. **Loss computation** — a **loss function** $L(\hat{y}, y)$ quantifies the disagreement between prediction and ground truth as a single number.
3. **Backward pass (backpropagation)** — the chain rule is applied backwards through the network to compute $\partial L/\partial w$, the **gradient** of the loss with respect to every parameter.
4. **Parameter update** — an **optimiser** steps each parameter against its gradient: $w \leftarrow w - \eta \cdot \partial L/\partial w$, where **η** is the **learning rate**.
5. **Repeat** — until the training set has been seen once (one **epoch**), then repeat for many epochs.

The distinction matters: an **iteration** is one parameter update on one mini-batch; an **epoch** is a full sweep of the dataset. 10 000 patches at batch size 32 gives 313 iterations per epoch.

4.2 Model families used in biological imaging

Architecture	Core mechanism	Typical biological use
CNN	Learned filters slide across the image, exploiting **translation equivariance** and **local connectivity**; deep stacks build a hierarchy from edges → textures → structures.	Image-level classification: pneumonia on chest X-ray, cell-type identification.
ResNet	**Residual (skip) connections** let gradients bypass layers, permitting very deep networks without vanishing gradients.	The default backbone and transfer-learning starting point.
U-Net	Symmetric **encoder–decoder** whose skip connections reunite high-resolution spatial detail with semantic context.	The dominant architecture for biomedical **segmentation** — tumour, organ, cell, vessel.
Vision Transformer	Image split into patch tokens; **self-attention** models long-range relationships across the whole image at once.	Whole-slide analysis where distant context matters; needs large data or pretraining.
Foundation / SSL models	Pretrained on millions of unlabelled images with a pretext task, then fine-tuned.	Pathology and radiology encoders used as general-purpose feature extractors.

nnU-Net is worth noting here: a *self-configuring* framework that derives its preprocessing, patch size, batch size and architecture automatically from a "dataset fingerprint", it has outperformed most bespoke designs across biomedical segmentation challenges (Isensee et al., *Nature Methods*, 2021). Its lesson is a general one — **most performance in biomedical segmentation comes from correct preprocessing and training configuration, not from novel architecture.**

The U-Net architecture

Figure 9. *Encoder–decoder structure of a U-Net. The contracting path reduces spatial resolution while increasing feature depth; the expanding path restores resolution; skip connections carry fine spatial detail across, which is why the architecture recovers accurate biological boundaries.*

4.3 Loss functions

Loss	Used for	Behaviour to remember
Cross-entropy	Classification.	Penalises confident wrong predictions heavily; the default.

Loss	Used for	Behaviour to remember
Dice loss	Segmentation.	Optimises region overlap directly, so it copes with severe foreground/background imbalance — essential when a tumour is 0.5% of a volume.
Dice + cross-entropy	Segmentation.	The de-facto standard: Dice supplies the overlap objective, cross-entropy stabilises early gradients.
Focal loss	Imbalanced detection.	A $(1-p)^{\gamma}$ factor suppresses loss from easy examples.
Tversky / focal-Tversky	Asymmetric costs.	Weights false negatives more heavily than false positives — appropriate when missing a lesion is worse than over-calling one.
MSE / MAE	Regression.	MSE is outlier-sensitive; MAE is more robust.
Contrastive / InfoNCE	Self-supervised pretraining.	Pulls augmented views of one image together and pushes different images apart, learning representations without labels.

4.4 Optimisers, learning rate and batch size

Pure **gradient descent** computes the gradient over the entire dataset — accurate but impossible at scale. **Stochastic gradient descent (SGD)** estimates it from a mini-batch, adding noise that helps escape poor minima. **Momentum** accumulates velocity to accelerate consistent directions and damp oscillation. **Adam** maintains per-parameter adaptive learning rates from running estimates of the gradient's first and second moments; **AdamW** decouples weight decay from the adaptive step and is the common default.

- **Learning rate (η)** — the single most important hyperparameter. Too high, training diverges; too low, it stalls or takes impractically long. Typical starting values: 1e-3 for Adam, 1e-2 for SGD with momentum.
- **Schedule** — η is not constant. **Warm-up** ramps it up over the first few hundred iterations; **step decay**, **cosine annealing** or **polynomial decay** then reduce it for fine convergence.
- **Batch size** — larger batches give smoother gradients and better hardware use but consume more memory and can generalise slightly worse. 3-D medical imaging often forces batch sizes of 2–8, hence **patch-based training** and **gradient accumulation**.
- **Mixed-precision training** — 16-bit computation with a 32-bit master copy of the weights roughly halves memory and doubles throughput at negligible accuracy cost.

4.5 Regularisation and the bias–variance trade-off

Underfitting (high bias): the model is too simple, and training and validation loss are both high. **Overfitting (high variance)**: the model has memorised the training set including its noise, so training loss keeps falling while validation loss rises. The gap between the two curves is the diagnostic signature.

- **Dropout** — randomly deactivates units each forward pass, preventing co-adaptation of features.
- **Weight decay (L2)** — penalises large weights, favouring smoother functions.
- **Batch / layer / instance normalisation** — normalises activations within a layer, stabilising and accelerating training with mild regularisation. With very small batches use **group normalisation**, since batch statistics become unreliable.
- **Early stopping** — halt when validation loss has not improved for a set **patience** and restore the best checkpoint. Simple and among the most effective controls available.
- **Data augmentation** — the strongest regulariser available in biology, which is why §3.7 sits where it does.

Training dynamics and the overfitting signature

Figure 10. *Training and validation loss against epoch for three regimes: underfitting, good fit and overfitting. The point at which validation loss turns upward defines the early-stopping checkpoint.*

4.6 Transfer learning and self-supervision

Labelled biological data is scarce; general image data is not. **Transfer learning** initialises a network with weights learned on a large source dataset, then adapts it. Two regimes: in **feature extraction** the pretrained backbone is **frozen** and only a small classifier head is trained — appropriate for hundreds of samples; in **fine-tuning** some or all backbone layers are unfrozen and trained at a reduced learning rate, often with **discriminative learning rates** (lower early, higher late) — appropriate for thousands.

Because natural photographs and stained tissue differ profoundly, **domain-specific pretraining** now dominates. Models are pretrained on large unlabelled biomedical corpora by **self-supervised learning (SSL)**, where supervision is manufactured from the data itself — predicting a masked patch (**masked image modelling**) or forcing two augmented views of one tile to produce matching embeddings (**contrastive learning**: SimCLR, DINO). Since unlabelled slides and scans exist in enormous quantity while annotations do not, SSL is arguably the most important recent development for AI in biology.

4.7 Weak supervision and multiple-instance learning

A pathologist reports "metastasis present" for a slide, not for each of its 50 000 tiles. **Multiple-instance learning (MIL)** solves exactly this: the slide is a **bag** of tile **instances**, labelled positive if *at least one* instance is positive. The model scores instances and aggregates them — by max-pooling, or in the now-standard **attention-based MIL** by learned attention weights. The attention map doubles as a built-in interpretability output, showing which regions drove the decision.

4.8 Cross-validation and hyperparameter search

When the cohort is small — as in biology it usually is — a single split wastes data and gives an unstable estimate. **k-fold cross-validation** divides data into *k* folds, trains *k* models each holding out one fold, and averages, yielding both a mean and a variance. **Stratified group k-fold** — stratified by class, grouped by patient — is the correct variant for biomedical data. **Leave-one-out (LOOCV)** is the extreme k = N case, used only for very small cohorts.

Hyperparameters are then searched by **grid search** (exhaustive, exponential cost), **random search** (usually more efficient per unit budget) or **Bayesian optimisation** (a surrogate model proposes promising configurations). Selection is made on the **validation folds only**; if the test set participates in model selection, the result is optimistically biased and must be described as a validation result.

Figure 11. *Five-fold cross-validation with patient-level grouping. Each row is one fold and the highlighted block is the held-out validation partition. All samples from a patient stay within one partition, and class proportions are preserved in every fold.*

> **NESTED CROSS-VALIDATION**
>
> When hyperparameter tuning and performance estimation must both come from one limited dataset, the correct procedure is **nested cross-validation**: an inner loop selects hyperparameters, an outer loop estimates generalisation. Nested CV is what prevents *optimistic bias from model selection* when a cohort is too small to afford a separate hold-out set.

4.9 Reproducibility of the training run

A training run is an experiment and must be reproducible. The record should contain **random seeds**, exact **library and driver versions**, the **dataset version and split file**, the complete **hyperparameter configuration**, the **hardware**, and the **training curves**; MLflow, Weights & Biases and DVC exist to enforce this. Exact bitwise reproducibility on a GPU is not always attainable, since some parallel operations are non-deterministic — the honest standard is that the procedure and seeds are published so results reproduce within a stated tolerance.

5 STAGE 3: EVALUATION

Evaluation asks what training cannot: *does this model work on data it has never seen, and does it fail in ways that matter?* A single headline accuracy answers neither. Rigorous evaluation is multi-metric, reports uncertainty, tests generalisation across sites, and inspects failure modes.

5.1 The confusion matrix — the root of everything

	Actually positive	Actually negative
Predicted positive	**True Positive (TP)** — metastasis present and detected.	**False Positive (FP)** — healthy tissue flagged as tumour. *Type I error.*
Predicted negative	**False Negative (FN)** — metastasis present but missed. *Type II error.*	**True Negative (TN)** — healthy tissue correctly cleared.

Every classification metric is an arithmetic combination of these four counts. In biology FP and FN carry radically different costs, which is why they must never be collapsed into one accuracy figure.

Metric	Definition	Interpretation and caution
Accuracy	(TP + TN) / total	Misleading under imbalance — the *accuracy paradox*. Rarely appropriate alone in biology.
Sensitivity / Recall	TP / (TP + FN)	Of all true cases, how many were caught. The priority in **screening**, where a missed cancer is unacceptable.
Specificity	TN / (TN + FP)	Of all healthy cases, how many were correctly cleared. Low specificity means unnecessary biopsies.
Precision / PPV	TP / (TP + FP)	Of everything flagged, how much was real. **Prevalence-dependent** — the same model has lower PPV in a low-prevalence population.
NPV	TN / (TN + FN)	Of everything cleared, how much was truly clear. Also prevalence-dependent.
F1 score	2 · (P · R) / (P + R)	Harmonic mean of precision and recall; ignores TN entirely.
Balanced accuracy	(Sens + Spec) / 2	A fairer summary than accuracy under imbalance.
MCC	Correlation over all four cells, −1 to +1	**Matthews Correlation Coefficient** — the most informative single summary for imbalanced binary problems.
Cohen's kappa (κ)	Agreement corrected for chance	Compares model against expert, and expert against expert.

5.2 Threshold-free metrics: ROC and precision–recall

A classifier outputs a probability; a **decision threshold** converts it to a class. Because the threshold is a free choice, performance should be summarised across all thresholds.

- **ROC curve** — sensitivity against 1 − specificity as the threshold sweeps. The **AUROC** has an elegant interpretation: *the probability that a randomly chosen positive scores higher than a randomly chosen negative.* 0.5 is chance, 1.0 perfect.
- **Precision–recall curve** — precision against recall; its area is the **AUPRC** or average precision. Because it ignores true negatives it is far more informative when positives are rare — AUROC can look impressive on a heavily imbalanced screening set while the model is clinically unusable.
- **Operating point selection** — the deployed threshold is chosen on the **validation set**, usually by fixing a required sensitivity (e.g. ≥ 95% for malignancy) and reporting the specificity achieved. **Youden's J** (sensitivity + specificity − 1) is a common context-free alternative.

Confusion matrix, ROC and precision–recall

Figure 12. *The 2 × 2 confusion matrix and the two threshold-free curves derived from it. ROC is insensitive to class prevalence; precision–recall is not, which is why it is the more honest summary for rare biological events.*

5.3 Segmentation and detection metrics

— **Dice Similarity Coefficient (DSC)** — 2|A ∩ B| / (|A| + |B|). The standard overlap measure, equivalent to F1 computed over pixels. A Dice of 0.90 for a large organ and 0.90 for a 4 mm lesion are very different achievements, so lesion size must be reported alongside.

— **Intersection over Union (IoU) / Jaccard** — |A ∩ B| / |A ∪ B|. Always numerically lower than Dice for the same segmentation; never compare a Dice figure directly against an IoU figure.

— **Hausdorff distance (HD95)** — maximum boundary-to-boundary distance, reported at the 95th percentile to suppress single-voxel outliers. Captures *boundary* accuracy that overlap metrics hide: a mask can have excellent Dice yet a clinically fatal boundary error.

— **Average symmetric surface distance (ASSD)** — mean distance between predicted and reference surfaces; a smoother companion to HD95.

— **FROC** — sensitivity against average false positives per image. The standard for lesion detection (pulmonary nodules, metastasis foci) because, unlike ROC, it accommodates multiple findings per image.

— **mAP** — mean average precision across recall levels and IoU thresholds; standard in object detection, e.g. mitotic figure counting.

Figure 13. *Predicted mask versus expert reference mask. The intersection defines Dice and IoU; the largest boundary discrepancy defines the Hausdorff distance. Two segmentations with equal Dice can differ substantially in boundary error.*

5.4 Calibration and uncertainty

A model that outputs 0.9 should be correct about 90% of the time when it says 0.9. That is **calibration**, distinct from discrimination (AUROC). Modern deep networks are systematically **over-confident**, and where the score is used to triage cases, miscalibration is a direct safety problem.

— **Reliability diagram** — predicted probability against observed frequency; the diagonal is perfect calibration. **Expected Calibration Error (ECE)** and the **Brier score** summarise it as scalars.

— **Temperature scaling** — a single scalar learned on the validation set that rescales the logits. Simple, effective, and does not change prediction ranking or AUROC.

— **Uncertainty estimation** — **Monte-Carlo dropout** (dropout kept active at inference, sampled repeatedly) and **deep ensembles** both yield a spread usable to abstain on ambiguous cases. Selective prediction — "refer to a human when uncertain" — is increasingly how AI is deployed in diagnostics.

5.5 Statistical rigour

A performance number without an interval is not a scientific result.

— **Confidence intervals** — usually by **bootstrap resampling** of the test set (1 000–10 000 resamples, 2.5th and 97.5th percentiles). Resampling must be at the **patient** level, not the image level.

— **Significance testing** — **DeLong's test** for comparing two AUROCs on the same data; **McNemar's test** for comparing two classifiers' paired error patterns.

— **Multiple-comparison correction** — Bonferroni or Benjamini–Hochberg **false discovery rate** control when many metrics, subgroups or models are compared.

— **Sample size and power** — a 40-case test set cannot distinguish AUROC 0.85 from 0.90. Reporting interval width makes this transparent.

— **Repeat runs** — training is stochastic, so report mean ± SD across several seeds, not one lucky run.

5.6 Interpretability: opening the black box

— **Grad-CAM** — weights the final convolutional feature maps by the gradient of the target class, producing a coarse heat map of evidence regions. The most widely used method in medical imaging.

— **Saliency and Integrated Gradients** — pixel-level attribution from gradients of the output with respect to the input.

— **SHAP** and **LIME** — model-agnostic local attribution, more common for tabular or radiomics features than raw pixels.

— **Attention maps** — in attention-based MIL and transformers the attention weights are themselves an interpretability output.

— **Concept-based methods** — testing whether a model has internally encoded a human-meaningful concept such as "nuclear pleomorphism" or "spiculated margin".

An essential caveat: **saliency maps explain the model, not the biology.** They fail some published sanity checks, and a plausible-looking heat map is not evidence of correct reasoning. Interpretability outputs need expert validation, not presentation as self-evident.

5.7 Generalisation, domain shift and shortcut learning

— **Internal validation** — a held-out test set from the same cohort, sites and scanners. Necessary but weak evidence.

— **External validation** — evaluation on data from an entirely independent institution, scanner or population. Performance almost always drops, and the size of the drop is the real measure of the model's value. A model without external validation is a hypothesis, not a result.

— **Domain shift** — deployment data differs statistically from training data. **Covariate shift** (a new scanner or staining protocol) is commonest in imaging; **prior probability shift** (changed prevalence) appears when moving from a curated cohort to a screening population; **concept drift** is the relationship itself changing over time.

— **Shortcut learning / the Clever Hans effect** — high performance achieved through a spurious correlate rather than the biology. Documented cases include models keying on chest-drain tubes to detect pneumothorax, on burned-in laterality markers, on hospital-specific scanner text, and on the ruler beside melanoma lesions in dermatology photographs. This is a *preprocessing* failure that only *evaluation* can expose — precisely why the three stages are one system.

— **Subgroup / fairness analysis** — performance reported separately by age, sex, ethnicity where available, scanner manufacturer and site. A model performing well on average while failing an under-represented subgroup is unsafe, and this **algorithmic bias** originates in unrepresentative training data, not in the algorithm.

Domain shift and shortcut learning

Figure 14. *Left: feature distributions of two hospitals overlap only partially, illustrating covariate shift and the resulting drop under external validation. Right: a Grad-CAM heat map concentrated on a burned-in image marker rather than the lesion — the signature of shortcut learning.*

5.8 Reporting standards

Two checklists dominate the field. **CLAIM** (Checklist for Artificial Intelligence in Medical Imaging), updated 2024 in *Radiology: Artificial Intelligence*, specifies what an imaging AI study must report. **TRIPOD+AI** (2024) extends the TRIPOD statement for clinical prediction models to machine learning. Related standards include **STARD** for diagnostic accuracy, **SPIRIT-AI** and **CONSORT-AI** for AI trials, and **FUTURE-AI** for trustworthy medical AI. Their shared demand: describe the data, splits, preprocessing, model, metrics, uncertainty and failure analysis in enough detail that another laboratory could repeat the work.

6 WORKED EXAMPLE: DETECTING LYMPH-NODE METASTASIS

The **CAMELYON16** challenge is the canonical end-to-end case study for this topic. The task: detect metastatic breast carcinoma in whole-slide images of **sentinel lymph node** sections. The dataset comprises **400 H&E-stained whole-slide images** from two Dutch centres — Radboud University Medical Center and University Medical Center Utrecht — split into 270 training and 130 test slides.

Pipeline stage	What was done	Principle illustrated
Ground truth	Slides scanned at 40×; pathologists exhaustively outlined metastatic regions, with immunohistochemistry confirming ambiguous cases.	Label quality; a hard confirmatory endpoint anchoring expert annotation.
QC & tissue detection	Pen marks, folds and out-of-focus regions handled; Otsu thresholding in HSV discarded the ~85% empty glass.	§3.1 — artefact rejection and background removal.
Tiling	Each slide cut into 256 × 256 px tiles; one slide yields 10^4–10^5 tiles.	§3.4 — patching a gigapixel image.
Stain handling	Macenko normalisation and aggressive stain augmentation to bridge the two centres.	§3.3 — technical variation removed, morphology preserved.
Splitting	Slide- and patient-level splits; no tile from a test slide ever appears in training.	§3.6 — grouping to prevent leakage.
Imbalance	Tumour tiles are a tiny minority; hard-negative mining retrained on the model's own false positives.	§3.8 — imbalance corrected in training only.
Training	CNN backbones fine-tuned from ImageNet; modern re-implementations use attention-based MIL with slide-level labels.	§4.6, §4.7 — transfer learning and weak supervision.
Inference	Tile probabilities stitched into a slide-level heatmap, connected-component features aggregated to a slide score.	Stitching overlapping predictions.
Evaluation	Slide-level **AUROC** for classification and **FROC** for lesion localisation, benchmarked against a pathologist panel.	§5.2, §5.3 — task-appropriate, multi-metric evaluation.

The headline finding, reported by Ehteshami Bejnordi and colleagues in *JAMA* (2017), is the important one: the best algorithms matched or exceeded a panel of pathologists working **under a simulated time constraint**, while pathologists without time limits remained highly accurate. The defensible reading is not "AI beats pathologists" but that **AI performs best as an assistive triage layer directing limited expert attention** — which is how such systems are actually deployed.

6.1 Deployment and monitoring — the fourth stage

- **Integration** — the model must run inside existing infrastructure (PACS for radiology, LIS for pathology) at acceptable latency.
- **Drift monitoring** — inputs monitored for **covariate shift** (a scanner upgrade, a reagent change) and outputs for prior shift. Silent degradation is the characteristic failure mode of deployed medical AI.
- **Human-in-the-loop design** — a defined escalation path for low-confidence predictions, and audit logging of every inference.
- **Periodic revalidation** — performance re-measured on freshly collected local data, never assumed to persist.
- **Model version control** — a change to weights is a change to a regulated device; regulators handle this through **Predetermined Change Control Plans (PCCP)**, which specify in advance how a model may be updated without a new submission.

7 BIAS, ETHICS AND REGULATION

Bias is not a single defect but something that enters at every stage of the pipeline.

Stage	Bias introduced	Example
Acquisition	Sampling / representation bias	A dermatology model trained overwhelmingly on light skin performs poorly on darker skin.
Annotation	Observer bias	A single annotator's systematic tendencies become the model's ground truth.

Stage	Bias introduced	Example
Preprocessing	Exclusion bias	QC criteria that disproportionately reject scans from older equipment used in under-resourced settings.
Training	Optimisation bias	A loss dominated by the majority class quietly ignores rare but critical presentations.
Evaluation	Reporting bias	Aggregate metrics concealing subgroup failure; publication of only the best of many runs.
Deployment	Automation bias	Clinicians over-trusting model output and under-weighting contradictory judgement.

The regulatory picture shows how mature the field has become: as of the FDA's mid-2026 list update, more than **1,500 AI-enabled medical devices** had US regulatory authorisation, roughly **three-quarters of them radiology devices** — the clearest demonstration that biomedical imaging is where AI has moved from research into routine practice. In the EU such software is regulated as a medical device under the **MDR**, with the **EU AI Act** adding a high-risk classification layer.

Data-governance obligations sit alongside: **GDPR** and **HIPAA** constrain how identifiable biological data may be processed, which has driven interest in **federated learning**, where the model travels to the hospitals holding the data and only weight updates are shared, so raw patient images never leave the institution.

8 SUMMARY OF PRINCIPLES

THE PIPELINE ON ONE LINE

Frame the task → acquire and de-identify → quality control → normalise (intensity + stain) → standardise geometry → split by patient → augment training folds → choose architecture and loss → optimise with a schedule and regularisation → select on validation → evaluate once on held-out and external data with multi-metric, interval-reported results → interpret, monitor, revalidate.

Every stage answers one question: *am I removing technical variation while preserving biological variation, and is the test set genuinely unseen?*

8.1 Eight principles that govern the whole pipeline

1. Preprocessing removes **technical** variation and preserves **biological** variation.
2. Normalisation statistics are computed on the **training split only**.
3. Splitting is done at the **patient level**, with stratification — never at the image or patch level.
4. Augmentation applies to the **training fold only** and must be **biologically plausible** (rotation is valid for histology; horizontal flip is not valid for chest radiographs).
5. Class imbalance is corrected in training, but the test set must retain **true prevalence**.
6. Accuracy alone is meaningless under imbalance — report **sensitivity, specificity, PPV, AUROC and AUPRC** with **confidence intervals**.
7. **External validation** across institutions is the only convincing evidence of generalisation.
8. Performance is capped by **annotation quality**; report inter-observer agreement (κ).

9 CONCLUSION

The AI data pipeline is best understood not as a piece of software engineering that biology has borrowed, but as the scientific method expressed in computational form. Its three stages map cleanly onto practices biologists already recognise: preprocessing is specimen preparation, training is the experiment, and evaluation is independent replication. Read that way, the rules that govern the pipeline stop looking arbitrary. Fitting normalisation statistics on the training split alone is the same discipline as blinding; splitting by patient rather than by image is the same discipline as respecting the true experimental unit; external validation is simply replication in another laboratory.

The evidence points consistently in one direction: **most of the performance, and nearly all of the failure, lives outside the model architecture.** nnU-Net outperformed bespoke designs by configuring preprocessing correctly. CAMELYON16-era systems succeeded because of tiling, stain handling and hard-negative mining. The published failures — models keying on chest drains, laterality markers or dermatology rulers — are without exception preprocessing and dataset failures that only rigorous evaluation exposed. Understanding *why* each stage exists, rather than merely listing the stages, is what the subject actually demands.

For biology specifically, three constraints shape everything. Labels are scarce and imperfect, which drives transfer learning, self-supervision and weak supervision. Samples are not independent, which makes patient-level grouping non-negotiable. And the cost of a false negative rarely equals the cost of a false positive, which is why no single metric can summarise a clinically meaningful result. The field is now mature enough to be regulated — over 1,500 AI-enabled devices authorised in the United States, three-quarters of them in imaging — and that maturity has moved the central question from *can a model achieve high accuracy?* to *can it be shown to work, safely and equitably, on data it has never seen?* Answering that question is what the pipeline exists to do.

10 GLOSSARY OF KEY TERMS

Term	Definition
Annotation	Expert-assigned label or outline used as ground truth.
AUROC / AUPRC	Area under the ROC / precision–recall curve; threshold-free performance summaries.
Augmentation	Label-preserving transformation of training data used as a regulariser.
Backpropagation	Chain-rule computation of loss gradients with respect to every parameter.
Batch effect	Systematic technical variation associated with processing group, site or scanner.
Bias field	Smooth multiplicative intensity drift in MRI; removed by N4 correction.
Calibration	Agreement between predicted probability and observed frequency.
Covariate shift	Change in input distribution between training and deployment.
Cross-validation	Repeated train/validate partitioning for a stable estimate on limited data.
Dice coefficient	Overlap metric $2\|A \cap B\| / (\|A\| + \|B\|)$; the standard segmentation score.
DICOM	Standard medical image format carrying pixel data plus extensive metadata.
Early stopping	Halting training when validation loss stops improving; best checkpoint restored.
Epoch	One complete pass of the training set through the model.
External validation	Evaluation on data from an entirely independent institution or population.
Federated learning	Training across institutions by sharing model updates, not raw data.
Ground truth	The reference standard against which predictions are judged.
Hyperparameter	A configuration value set before training, not learned from data.
Inter-observer variability	Disagreement between experts labelling the same data; measured by kappa.
Leakage	Any flow of information from the test set into training or preprocessing.
Learning rate	Step size of the parameter update; the most influential hyperparameter.
Loss function	Scalar measure of prediction error, minimised during training.
MIL	Multiple-instance learning; bag-level labels supervise instance-level learning.
Overfitting	Memorisation of training data including noise, giving poor generalisation.
Prevalence	Proportion of positive cases in a population; determines PPV and NPV.
Registration	Spatial alignment of images into a common coordinate frame.
Regularisation	Any constraint that reduces variance and improves generalisation.
Resampling (spatial)	Interpolating an image onto a new, typically isotropic, voxel grid.
Segmentation	Assignment of a class label to every pixel or voxel.
Self-supervised learning	Pretraining on unlabelled data using a pretext task derived from the data itself.
Shortcut learning	Reliance on a spurious correlate rather than the intended biological signal.
Stain normalisation	Mapping H&E colour distributions onto a common reference.
Stratification	Preserving class and covariate proportions across data splits.
Transfer learning	Reusing weights learned on a source task to initialise a target task.
Whole-slide image (WSI)	Digitised histology slide, gigapixel-scale, stored as an image pyramid.

11 SELECTED REFERENCES

Isensee, F. *et al.* (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18, 203–211.

Ehteshami Bejnordi, B. *et al.* (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22), 2199–2210.

Litjens, G. *et al.* (2018). 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience*, 7(6), giy065.

Ronneberger, O., Fischer, P. & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI*, 234–241.

Macenko, M. *et al.* (2009). A method for normalizing histology slides for quantitative analysis. *IEEE ISBI*, 1107–1110.

Tustison, N. J. *et al.* (2010). N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6), 1310–1320.

Tejani, A. S. *et al.* (2024). Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiology: Artificial Intelligence*, 6(4), e240300.

Collins, G. S. *et al.* (2024). TRIPOD+AI statement: updated guidance for reporting clinical prediction models. *BMJ*, 385, e078378.

Geirhos, R. *et al.* (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2, 665–673.

US Food and Drug Administration (2026). *Artificial Intelligence-Enabled Medical Devices* — public list, update of 16 June 2026.