Training-data sources for littlEye AI

A methodical guide to finding, assessing, downloading and processing histology images with diagnoses, so that they yield a model you can trust — and one you may legally share with colleagues and students.

1 · Why the licence decides the outcome, not the image count

There is an endless supply of histology images on the internet. Only a small fraction is usable, and for a single reason: a model trained on images you may not use must not be distributed further. Its accuracy does not help. If the source has no clear licence, the result stays on your disk forever.

The practical consequence for littlEye AI: models derived from the University of Michigan collection are licensed CC BY-NC-SA, i.e. non-commercial — which is why they are not, and must not be, part of the paid application. The app ships with a single sample model trained on synthetic patterns; real models are trained by the user or exchanged as standalone files.

How this went for us — and what follows from it

We distributed the Michigan-derived models free of charge, as separate files and under the same licence. The author of the collection pointed out that he was unsure this matched the rights holder's view, and put us in touch with their Tech Transfer Office. The answer was unambiguous: our use is commercial and the NonCommercial term does not cover it, even where the images are not part of the app. They offered a paid licence tiered by sales, starting at USD 250 per year. We declined it and deleted every model and image patch derived from that collection.

The lesson for anyone building something similar: if your project generates any revenue at all, do not rely on your own reading of the NonCommercial term. Giving the model away free is not enough, nor is keeping the images out of the product — the rights holder judges the purpose, not the technique. Ask first, in writing; the answer arrives faster than you can train a model.

The rule for choosing a source

A usable source has an explicitly stated licence that permits derivative works (the model) and their further distribution. “Freely available on the internet”, “free for students” or “for educational purposes” is not a licence by itself — those are conditions of access, not permission to process and redistribute.

2 · Checklist — what a source must satisfy

A source is suitable for littlEye AI if it meets all seven points. If it fails the first, there is no need to verify the rest.

1. The licence is stated and permits derivative works. Ideally CC0, CC BY or CC BY-SA. CC BY-NC means non-commercial use only — usable, but the model must then not go into a paid application.
2. The images carry a diagnosis established by a pathologist, not a guess from the picture caption.
3. The magnification is stated, or the pixel size (µm/px). Without it you cannot know whether the images match the scale the model runs at.
4. The stain is H&E;, or it is clearly stated which one. A model trained on H&E; cannot assess immunohistochemistry and vice versa.

5. The images come from multiple patients and it is known how many. This is the most commonly overlooked point and the most common cause of falsely good results.
6. The resolution is at least 360 px on the shorter side — below this threshold detail is merely interpolated (details in chapter 4).
7. The data are anonymized — no names, birth numbers or identifiable details.

3 · Recommended sources

The list is ordered by how well each source fits littlEye AI. Licences were verified in August 2026 directly on the providers' pages — verify them again before downloading, terms change.

Group A — the licence is clear, download right away

Source	What it contains	Licence	Why it fits
NCT-CRC-HE-100K (Kather, Zenodo)	100,000 patches 224×224, colorectum. 9 classes: adenocarcinoma, tumour stroma, lymphocytes, mucus, smooth muscle, normal mucosa, debris, adipose tissue, background. Plus CRC-VAL-HE-7K — 7,180 images from 50 different patients.	CC BY 4.0	0.5 µm/px = 20× objective — exactly the scale littlEye AI runs at. A separate test set from different patients is exactly what honest metrics require.
TCGA-UT (Univ. of Tokyo, Zenodo)	1.6 million patches 256×256 from 32 types of solid cancer, from 7,175 patients and 8,736 diagnostic slides. Six scale levels (0.5–1.0 µm/px). Downloaded per cancer — from 121 MB (cholangiocarcinoma) to 3.2 GB, 35.3 GB in total.	CC BY-NC-SA 4.0 (non-commercial, same as Michigan)	The greatest patient diversity of all — with 7,175 people the “slide-hue” trap is practically impossible. But: patches are 256 px, below the 360 px threshold. For serious work, take the original slides from GDC and tile them to 480 px.
BreakHis (UFPR, Brazil)	9,109 images 700×460, breast, 82 patients. Benign: adenosis, fibroadenoma, phyllodes tumour, tubular adenoma. Malignant: ductal, lobular, mucinous and papillary carcinoma.	CC BY 4.0 stated; the same page, however, says “non-commercial research only” and asks for registration	The only large set with the magnification directly in the label — 40×, 100×, 200× and 400×. Eight real histological diagnoses.
PatchCamelyon (PCam)	327,680 patches 96×96 from lymph nodes, labelled “metastasis yes/no”.	CC0 — public domain	The cleanest licence of all. But: 96 px is far below the usability threshold (chapter 4) — better take the original Camelyon16 slides and tile them yourself.
Human Protein Atlas	44 normal tissues and the 20 most common cancers, high resolution.	CC BY 4.0	Mostly immunohistochemistry, not H&E; — marginal for littlEye AI for now. Useful if you later add an IHC module.
LC25000	25,000 images 768×768: colon adenocarcinoma, normal colon, lung adenocarcinoma, lung squamous cell carcinoma, normal lung.	“freely available for research” — a formal licence is missing	Beware: the 25,000 images were augmented from 750 originals. There are orders of magnitude fewer independent cases than it seems — see trap #2 in chapter 6.

Where the original slides are: the GDC portal (portal.gdc.cancer.gov) provides over 30,000 whole TCGA slides in SVS format. They are not under a Creative Commons licence but under data-use terms — which, among other things, prohibit attempting to identify individual people. For tiling at 480 px it is the richest source in existence.

What you will find nowhere — and why that is your opportunity

A public dataset of forensic histology does not exist. Searching through pathology collections and specialist reviews found not a single open collection of autopsy material with diagnoses — myocardial infarction, drowning, fat embolism, wound age. Everything public is oncological pathology from biopsies and resections.

That means two things. First: from public sources you will never train a model that helps you at the autopsy table — that requires your own material. Second, and more importantly: the material at your institute is professionally unique. When negotiating with a university this is the strongest argument — you are not asking for help with something downloadable, you are building something that does not yet exist.

Group B — ask for permission first

These collections are high-quality and their diagnoses were established by pathologists, but they have no open licence. Downloading them without an agreement is not acceptable even though they are publicly accessible.

Source	What it contains	Status
PEIR Digital Library (Univ. of Alabama)	Over 30,000 curated teaching images from pathology and histology.	Copyright is retained by the university; “non-commercial educational use” with attribution is permitted. Training a model and distributing it requires written consent.
Virtual Pathology, University of Leeds	13,370 slides with established diagnoses and anonymized clinical data.	Intended for teaching and training. In scope and quality the best candidate for collaboration — but strictly by agreement.
NUS PathWeb (Singapore)	Systematic pathology by organ system.	The site has bot protection and forbids indexing; no licence is stated. A permission request is prepared in Spolupraca/ziadost_NUS_pathweb.md.
University of Michigan collection	212 virtual slides of normal histology — already in use.	CC BY-NC-SA 4.0. Any model derived from it must remain non-commercial and under the same licence.

4 · Magnification and resolution — measured values

This is not an estimate. These are measurement results on the littlEye AI baseline model (19 tissues, held-out test of 286 images the model never saw). The test images were modified to imitate different magnifications and worse quality:

Image conditions	Accuracy	Conclusion
20× objective (200× total), 480 px — as the model was trained	78.3 %	reference value
40× objective (400× total)	52.1 %	a quarter of the accuracy is lost
100× oil immersion (1000× total)	10.8 %	with 19 classes chance is 5.3 % — practically guessing
240 px resolution	69.9 %	still usable
160 px resolution	59.4 %	borderline
96 px resolution (phone through the eyepiece)	29.7 %	unusable

Image conditions	Accuracy	Conclusion
64 px resolution	18.5 %	unusable

What this means for choosing a source

The model processes a field of view at 360×360 px and was trained on 480 px patches. Everything below 360 px is interpolated — the detail is physically gone. Therefore: prefer a source with a 20× objective and patches of at least 480 px. A source at a different scale can be used, but it must be rescaled — and above all a separate model must be trained from it, never mixed into an existing one.

The practical consequence for BreakHis: the four magnifications must not be mixed into one model. Either pick the one matching your practice, or train four separate models and switch between them in the app according to the objective you used.

5 · Step-by-step procedure

1. Verify the licence on the source's page and save the text (link and date). It will come in handy when someone asks where your data came from.
2. Download the data and unpack it. With large files (NCT is 11.7 GB) budget for disk space.
3. Find out the number of patients, not the number of images. If the source does not state it, treat the material as risky and do not trust metrics from it.
4. Split by patient, not randomly. All images of one patient must end up either in training or in the test — never in both. Otherwise the model “knows the answer” and the test shows an undeservedly high number.
5. Unify scale and size — crop to a square and rescale to 480 px.
6. Assign diagnoses using the littlEye AI catalogue. Use the catalogue key (e.g. kolon.adenokarcinom), not free text — the key is language-independent and survives translation and renaming.
7. Package as .le3set and import it in the app on the Overview screen with the Import dataset button.
8. Train the model in the Training section. Keep the held-out test switched on — without it you have no way of knowing whether the model is any good.
9. Read the confusion matrix, not the overall accuracy. What matters is the classes with the lowest scores and what the model confuses them with — the confusion is often histologically understandable and shows you what to add to the dataset.
10. Record the provenance. Put the source, licence and download date into the model name and the note. A year from now you will not remember, and without it the model must not be passed on.

6 · Three traps that ruin the result

All three look like success — the model shows high numbers and fails on a real image.

Trap 1 — the model learns the slide hue, not the tissue

If all images of one class come from a single block, the model learns its colour. This genuinely happened to us: a liver trained from a single case classified a redder liver from another lab as kidney. Solution: at least three different cases or blocks per class.

Trap 2 — augmented copies passed off as independent images

Some public sets are inflated by rotating and flipping. LC25000 has 25,000 images, but only 750 originals. If a rotated copy of the same image ends up in the test, the model “knows” it and the test lies. Solution: find out the number of original images and split by those.

Trap 3 — one patient's images in both training and test

The most insidious one, because a random split creates it by itself. The model memorizes the individual tissue features of one particular person. Solution: split by patient — which is why the table in chapter 3 stresses that the NCT-CRC test set comes from 50 different patients.

> **How to tell you have fallen into a trap**
>
> The model scores above 95 % on the test and fails on a real image from your microscope. It is almost always one of these three traps, not an app defect. An honest model on mixed material scores more like 75–85 % — and is more usable than a model with 99 % on material from a single slide.

7 · How to approach a university

The goal is not to obtain images at any cost but to start a collaboration. A university will share data when it can see what will come of it and that the material is handled correctly.

- Come with a working demonstration. An app that already recognizes 19 tissues is a different argument from an idea. Show the model live.
- Say precisely what you are asking for. Not “some images”, but for instance: 50 slides per diagnosis, H&E;, 20× objective, at least 3 different patients per diagnosis, anonymized, as JPEG or TIFF.
- Offer something in return. The department gets the model, students get a practice tool, and the collaborating institute is credited as co-author of the dataset.
- Define the limits clearly. littlEye AI is a research and teaching tool, not a medical device, and does not replace the pathologist. Say this before they ask — it signals that you know what you are doing.
- Agree the licence in writing before any images change hands: who the author is, what they may be used for, whether a model derived from them may be distributed and under what condition.

Natural partners in Slovakia: the Institute of Histology and Embryology and the Institute of Pathological Anatomy of Comenius University in Bratislava (already collaborating), the Institute of Histology and Embryology of UPJŠ in Košice, and the Institute of Histology and Embryology of JLF UK in Martin.

8 · The legal and ethical minimum

- State the provenance of every model you distribute — source, licence and download date.

- Respect the non-commercial condition. A model derived from CC BY-NC material must not be part of a paid application or a paid service.
- Preserve the ShareAlike condition where the source has one — the derived model must stay under the same licence.
- Never process non-anonymized material. An autopsy image is personal data as long as it can be linked to a particular person.
- With your own material, check in advance what your employer and the ethics committee permit — even for teaching.

Verified in August 2026 directly on the providers' pages. Licence terms change — verify them again before downloading.

Ján Šikuta, MD, PhD. · forensika.eu@icloud.com · littlEye AI 2.2.3 · methodology v1.0