

Zdroje trénovacích dát pre littlEye AI

Metodický postup, ako nájsť, posúdiť, stiahnuť a spracovať histologické snímky s diagnózou tak, aby z nich vznikol model, ktorému sa dá veriť — a ktorý sa dá legálne šíriť kolegom a študentom.

1 · Prečo o výsledku rozhoduje licencia, nie počet snímok

Histologických snímok je na internete nekonečne veľa. Použiteľných je z nich malá časť, a to z jediného dôvodu: model natrénovaný na snímkach, ktoré nesmiete použiť, sa nesmie šíriť ďalej. Nepomôže, že je presný. Ak zdroj nemá jasnú licenciu, výsledok zostane navždy len na vašom disku.

Praktický dôsledok pre littlEye AI: modely odvodené zo zbierky University of Michigan sú pod licenciou CC BY-NC-SA, teda nekomerčné — preto nie sú a nesmú byť súčasťou platenej aplikácie. Aplikácia obsahuje jediný ukázkový model natrénovaný na syntetických vzoroch; skutočné modely si používateľ trénuje sám alebo si ich vymieňa ako samostatné súbory.

Ako to dopadlo u nás — a čo z toho plynie

Michiganské modely sme širili bezplatne, ako samostatné súbory a pod tou istou licenciou. Autor zbierky nás upozornil, že si nie je istý, či to zodpovedá predstave držiteľa práv, a spojil nás s ich Tech Transfer Office. Odpoveď bola jednoznačná: naše použitie je komerčné a podmienka NonCommercial ho nepokrýva — a to aj vtedy, keď snímky nie sú súčasťou aplikácie. Ponúkli platenú licenciu odstupňovanú podľa predaja, od 250 USD ročne. Odmietli sme ju a všetky modely aj výrezy odvodené z tejto zbierky sme zmazali.

Poučenie pre kohokoľvek, kto stavia niečo podobné: ak z vášho projektu plynie akýkoľvek príjem, nespoliehajte sa na vlastný výklad podmienky NonCommercial. Nestačí, že model rozdáte zadarmo, ani že snímky v produkte nie sú — držiteľ práv posudzuje účel použitia, nie techniku. Spýtajte sa vopred a písomne; odpoveď príde rýchlejšie, než stihnete model natrénovať.

Pravidlo pri výbere zdroja

Použiteľný je zdroj s výslovne uvedenou licenciou, ktorá dovoľuje odvodené dielo (model) a jeho ďalšie šírenie. „Voľne dostupné na internete“, „zadarmo pre študentov“ ani „na vzdelávacie účely“ samo osebe licencia nie je — sú to podmienky prístupu, nie povolenie na spracovanie a šírenie.

2 · Kontrolný zoznam — čo musí zdroj spĺňať

Zdroj je pre littlEye AI vhodný, ak splní všetkých sedem bodov. Ak zlyhá na prvom, ostatné už netreba overovať.

1. Licencia je uvedená a dovoľuje odvodené dielo. Ideálne CC0, CC BY alebo CC BY-SA. CC BY-NC znamená len nekomerčné použitie — použiteľné, ale model potom nesmie ísť do platenej aplikácie.
2. Snímky majú uvedenú diagnózu stanovenú patológom, nie odhad z popisu obrázka.
3. Je uvedené zväčšenie alebo veľkosť pixela ($\mu\text{m}/\text{px}$). Bez toho neviete, či snímky sedia s mierkou, na ktorej model beží.
4. Farbenie je HE, prípadne je jasne uvedené aké. Model natrénovaný na HE nevie hodnotiť imunohistochémiu a naopak.

5. Snímky pochádzajú od viacerých pacientov a je známe, od koľkých. Toto je najčastejšie prehliadaný bod a najčastejšia príčina falošne dobrých výsledkov.
6. Rozlíšenie je aspoň 360 px na kratšej strane — pod touto hranicou sa detail už len dopočítava (podrobne v kapitole 4).
7. Dáta sú anonymizované — bez mien, rodných čísel a identifikovateľných údajov.

3 · Odporúčané zdroje

Zoznam je zoradený podľa toho, ako dobre sedí na littlEye AI. Licencie sú overené k augustu 2026 priamo na stránkach poskytovateľov — pred stiahnutím ich overte znova, podmienky sa menia.

Skupina A — licencia je jasná, stiahnuť sa dá hneď

Zdroj	Čo obsahuje	Licencia	Prečo sedí
NCT-CRC-HE-100K (Kather, Zenodo)	100 000 výrezov 224×224, kolorektum. 9 tried: adenokarcinóm, stróma nádoru, lymfocyty, hlien, hladká svalovina, normálna sliznica, detritus, tukové tkanivo, pozadie. K tomu CRC-VAL-HE-7K — 7 180 snímok od 50 iných pacientov.	CC BY 4.0	0,5 µm/px = objektív 20× — presne mierka, na ktorej littlEye AI beží. Samostatná testovacia sada od iných pacientov je presne to, čo treba na poctivé metriky.
TCGA-UT (Univ. of Tokyo, Zenodo)	1,6 milióna výrezov 256×256 z 32 typov solídnych nádorov, od 7 175 pacientov z 8 736 diagnostických sklíčok. Šesť úrovní mierky (0,5–1,0 µm/px). Sťahuje sa po nádoroch — od 121 MB (cholangiokarcinóm) po 3,2 GB, spolu 35,3 GB.	CC BY-NC-SA 4.0 (nekomerčne, rovnako ako Michigan)	Najväčšia rozmanitosť pacientov zo všetkých — pri 7 175 ľuďoch pasca „odtieň skla“ prakticky nehrozí. Ale: výrezy majú 256 px, teda pod hranicou 360 px. Na vážnu prácu radšej siahnite po pôvodných sklách v GDC a orežte si ich na 480 px.
BreakHis (UFPR, Brazília)	9 109 snímok 700×460, prsník, 82 pacientok. Nezhubné: adenóza, fibroadenóm, fyloidný nádor, tubulárny adenóm. Zhubné: ductálny, lobulárny, mucinózny a papilárny karcinóm.	CC BY 4.0 uvedená; tá istá stránka však píše „len nekomerčný výskum“ a žiada registráciu	Jediný veľký súbor, kde je zväčšenie priamo v označení — 40×, 100×, 200× a 400×. Osem skutočných histologických diagnóz.
PatchCamelyon (PCam)	327 680 výrezov 96×96 z lymfatických uzlín, označené „metastáza áno/nie“.	CC0 — voľné dielo	Licenčne najčistejší zo všetkých. Ale: 96 px je hlboko pod hranicou použiteľnosti (kapitola 4) — berte radšej pôvodné sklá Camelyon16 a orežte si ich sami.
Human Protein Atlas	44 normálnych tkanív a 20 najčastejších nádorov, vysoké rozlíšenie.	CC BY 4.0	Prevažne imunohistochemia, nie HE — pre littlEye AI zatiaľ okrajové. Užitočné, ak neskôr pridáte IHC modul.
LC25000	25 000 snímok 768×768: adenokarcinóm hrubého čreva, normálne črevo, adenokarcinóm pľúc, skvamocelulárny karcinóm pľúc, normálne pľúca.	„voľne dostupné pre výskum“ — formálna licencia chýba	Pozor: 25 000 snímok vzniklo augmentáciou zo 750 pôvodných. Nezávislých prípadov je teda rádovo menej, než sa zdá — viď pascu č. 2 v kapitole 6.

Kde sú pôvodné sklá: portál GDC (portal.gdc.cancer.gov) sprístupňuje vyše 30 000 celých sklíčok TCGA vo formáte SVS. Nie sú pod licenciou Creative Commons, ale pod podmienkami použitia dát — tie okrem iného zakazujú pokúšať sa identifikovať konkrétnych ľudí. Na tiling na 480 px je to najbohatší zdroj, aký existuje.

Čo nenájdete nikde — a prečo je to vaša príležitosť

Verejný dataset súdnolekárskej histológie neexistuje. Hľadanie cez zbierky patológie aj cez odborné prehľady nenašlo ani jednu otvorenú kolekciu autoptického materiálu s diagnózami — infarkt myokardu, utopenie, tuková embólia, vek poranenia. Všetko verejné je onkologická patológia z biopsií a resekátov.

To znamená dve veci. Po prvé: z verejných zdrojov nikdy nenatrénujete model, ktorý by vám pomohol pri pitve — na to treba váš vlastný materiál. Po druhé, a dôležitejšie: materiál, ktorý máte na ústave, je odborne jedinečný. Pri jednaní s univerzitou je to najsilnejší argument — nežiadate o pomoc s niečím, čo sa dá stiahnuť, ale staviate niečo, čo zatiaľ neexistuje.

Skupina B — najprv treba požiadať o súhlas

Tieto zbierky sú kvalitné a diagnózy v nich stanovili patológovia, ale nemajú otvorenú licenciu. Sťahovať ich bez dohody nie je v poriadku ani vtedy, keď sú verejne prístupné.

Zdroj	Čo obsahuje	Stav
PEIR Digital Library (Univ. of Alabama)	Vyše 30 000 kurátorovaných výučbových snímok z patológie a histológie.	Autorské práva si ponecháva univerzita, povolené je „nekomerčné vzdelávacie použitie“ s uvedením zdroja. Na tréning modelu a jeho šírenie treba písomný súhlas.
Virtual Pathology, University of Leeds	13 370 sklíčok s vopred stanovenou diagnózou a anonymizovanými klinickými údajmi.	Určené na výučbu a školenia. Rozsahom a kvalitou najlepší kandidát na spoluprácu — ale výhradne na základe dohody.
NUS PathWeb (Singapur)	Systematická patológia po orgánových systémoch.	Stránka má ochranu proti robotom a zákaz indexovania, licencia nie je uvedená. Žiadosť o súhlas je pripravená v Spolupraca/ziadost_NUS_pathweb.md.
Zbierka Univerzity v Michigane	212 virtuálnych sklíčok normálnej histológie — už používame.	CC BY-NC-SA 4.0. Model z nich odvodený musí zostať nekomerčný a pod rovnakou licenciou.

4 · Zväčšenie a rozlíšenie — namerané hodnoty

Toto nie je odhad. Sú to výsledky merania na základnom modeli littlEye AI (19 tkanív, nezávislý test 286 snímok, ktoré model nikdy nevidel). Testovacie snímky sa upravovali tak, aby napodobnili iné zväčšenie a horšiu kvalitu:

Podmienky snímky	Presnosť	Záver
objektív 20× (celkovo 200×), 480 px — tak, ako sa model učil	78,3 %	referenčná hodnota
objektív 40× (celkovo 400×)	52,1 %	stráca sa štvrtina presnosti
objektív 100× imerzia (celkovo 1000×)	10,8 %	pri 19 triedach je náhoda 5,3 % — prakticky hádanie
rozlíšenie 240 px	69,9 %	ešte použiteľné
rozlíšenie 160 px	59,4 %	hraničné
rozlíšenie 96 px (mobil cez okulár)	29,7 %	nepoužiteľné
rozlíšenie 64 px	18,5 %	nepoužiteľné

Čo z toho vyplýva pre výber zdroja

Model spracúva zorné pole v rozlíšení 360×360 px a učil sa na výrezoch 480 px. Všetko pod 360 px sa dopočítava — detail tam už fyzicky nie je. Preto: uprednostnite zdroj s objektívom 20× a výrezmi aspoň 480 px. Zdroj s inou mierkou sa dá použiť, ale musí sa preškálovať — a hlavne sa z neho musí trénovať samostatný model, nie miešať do existujúceho.

Praktický dôsledok pre BreakHis: štyri zväčšenia sa nemiešajú do jedného modelu. Buď si vyberte to, ktoré zodpovedá vašej praxi, alebo natrénujte štyri samostatné modely a v aplikácii ich prepínajte podľa toho, akým objektívom ste snímku urobili.

5 · Postup krok za krokom

1. Overte licenciu na stránke zdroja a text si uložte (odkaz aj dátum). Bude sa hodiť, keď sa vás niekto opýta, odkiaľ dáta máte.
2. Stiahnite dáta a rozbaľte ich. Pri veľkých súboroch (NCT má 11,7 GB) počítajte s miestom na disku.
3. Zistite počet pacientov, nie počet snímok. Ak to zdroj neuvádza, považujte materiál za rizikový a metrikám z neho neverte.
4. Rozdeľte podľa pacienta, nie náhodne. Všetky snímky jedného pacienta musia skončiť buď v tréningu, alebo v teste — nikdy v oboch. Inak model „pozná odpoveď“ a test ukáže nezaslúžene vysoké číslo.
5. Zjednoťte mierku a rozmer — orežte na štvorec a preškálujte na 480 px.
6. Priradte diagnózy podľa katalógu littlEye AI. Použite katalógový kľúč (napr. kolon.adenokarcinom), nie voľný text — kľúč je jazykovo nezávislý a prežije preklad aj premenovanie.
7. Zabaľte ako .le3set a naimportujte v aplikácii na obrazovke Prehľad tlačidlom Importovať dataset.
8. Natrénujte model v sekcii Tréning. Nechajte zapnutý nezávislý test — bez neho nemáte ako zistiť, či je model dobrý.
9. Prečítajte konfúznú maticu, nie celkovú presnosť. Zaujímajú vás triedy s najnižšou úspešnosťou a to, s čím si ich model mýli — často je tá zámena histologicky pochopiteľná a ukáže vám, čo do datasetu doplniť.
10. Zapište pôvod. Do názvu modelu a do poznámky uveďte zdroj, licenciu a dátum stiahnutia. O rok si to nepamätáte a bez toho sa model nesmie posunúť ďalej.

6 · Tri pasce, ktoré pokazia výsledok

Všetky tri vyzerajú ako úspech — model má vysoké čísla a na reálnej snímke zlyhá.

Pasca 1 — model sa naučí odtieň skla, nie tkanivo

Ak všetky snímky jednej triedy pochádzajú z jedného bloku, model sa naučí jeho farbu. Toto sa nám reálne stalo: pečeň natrénovaná z jediného prípadu klasifikovala červenšiu pečeň z iného pracoviska ako obličku. Riešenie: aspoň tri rôzne prípady alebo bloky na každú triedu.

Pasca 2 — augmentované kópie vydávané za nezávislé snímky

Niektoré verejné súbory sú nafúknuté otáčaním a preklápaním. LC25000 má 25 000 snímok, ale vzniklo ich len 750 pôvodných. Ak otočená kópia tej istej snímky skončí v teste, model ju „pozná“ a test klame. Riešenie: zistíte počet pôvodných snímok a delíte podľa nich.

Pasca 3 — snímky jedného pacienta v tréningu aj v teste

Najzákernejšia, lebo náhodné rozdelenie ju vytvorí samo. Model si zapamätá individuálne znaky tkaniva konkrétneho človeka. Riešenie: rozdeľujte podľa pacienta — preto je v tabuľke v kapitole 3 pri NCT-CRC zdôraznené, že testovacia sada pochádza od iných 50 pacientov.

Ako spoznáte, že ste do pasce spadli

Model má na teste vyše 95 % a na reálnej snímke z vášho mikroskopu zlyhá. Takmer vždy je to jedna z týchto troch pascí, nie chyba aplikácie. Poctivý model na zmiešanom materiáli má skôr 75–85 % — a je použiteľnejší ako model s 99 % na materiáli z jedného skla.

7 · Ako osloviť univerzitu

Cieľom nie je získať snímky za každú cenu, ale začať spoluprácu. Univerzita dá dáta vtedy, keď vidí, čo z toho vznikne a že sa s materiálom zaobchádza korektne.

- Príďte s hotovou ukážkou. Aplikácia, ktorá už rozpoznáva 19 tkanív, je iný argument než nápad. Ukážte model naživo.
- Povedzte presne, čo žiadate. Nie „nejaké snímky“, ale napríklad: 50 preparátov na diagnózu, HE, objektív 20×, aspoň 3 rôzni pacienti na diagnózu, anonymizované, vo formáte JPEG alebo TIFF.
- Ponúknite protihodnotu. Model dostane pracovisko k dispozícii, študenti dostanú nástroj na precvičovanie a spolupracujúci ústav sa uvádza ako spoluautor datasetu.
- Vymedzte sa jasne. littlEye AI je výskumný a výučbový nástroj, nie zdravotnícka pomôcka a nenahrádza patológa. Toto povedzte skôr, než sa opýtajú — je to znak, že viete, čo robíte.
- Dohodnite licenciu písomne ešte pred odovzdaním snímok: kto je autor, na čo sa smú použiť, či sa smie šíriť model odvodený z nich a pod akou podmienkou.

Prirodzení partneri na Slovensku: Ústav histológie a embryológie a Ústav patologickej anatómie LF UK v Bratislave (už spolupracujú), Ústav histológie a embryológie UPJŠ v Košiciach a Ústav histológie a embryológie JLF UK v Martine.

8 · Právne a etické minimum

- Uveďte pôvod pri každom šírení modelu — zdroj, licenciu a dátum stiahnutia.
- Rešpektujte nekomerčnú podmienku. Model odvodený z materiálu CC BY-NC nesmie byť súčasťou platenej aplikácie ani platenej služby.
- Zachovajte podmienku ShareAlike, ak ju zdroj má — odvodený model musí ostať pod rovnakou licenciou.

- Nikdy nespracúvajte neanonymizovaný materiál. Snímka z pitvy je osobný údaj, pokiaľ sa dá spojiť s konkrétnym človekom.
- Pri vlastnom materiáli si vopred overte, čo dovoľuje váš zamestnávateľ a etická komisia — aj keď ide o výučbu.

Overené k 22. augustu 2026 priamo na stránkach poskytovateľov. Licenčné podmienky sa menia — pred stiahnutím ich overte znova.

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