The data files are available at https://github.com/fufu1209/pjp-crc-resource-aware-pathology.

File	Contents
supplementary_data/revision1_nct224_manifest_sha256.csv	SHA-256 digest, byte size, class and split for **all 9,900** native-resolution 224 × 224 images
supplementary_data/revision1_nct224_provenance.json	Machine-readable provenance record
supplementary_data/supplementary_TableS1_class_distribution.csv	PathMNIST per-class counts for all three splits (Supplementary Table S1)

Native-resolution 224 × 224 subset

Field	Value
Kaggle mirror	gpreda/nct-crc-he-100k-subset
URL	https://www.kaggle.com/datasets/gpreda/nct-crc-he-100k-subset
Licence reported by the Kaggle CLI	CC0-1.0
Access date	4 July 2026 (manifest regenerated 27 August 2026)
Original source family	NCT-CRC-HE-100K / CRC-VAL-HE-7K, Kather et al., Zenodo DOI 10.5281/zenodo.1214456
Training images	9,000 (1,000 per class, all nine classes)
Validation images	900 (100 per class, all nine classes)
Split construction	train/ and valid/ folders were **supplied by the mirror and used exactly as distributed**. The authors performed no re-partitioning. Labels were read from the filename prefix.
Splitting level	**Image level.** The mirror distributes no patient or slide identifiers, so slide- or patient-level separation between the two folders **cannot be verified**.
Exact-duplicate pairs within the subset	**0**
Exact-duplicate pairs across train/ and valid/	**0**
Independence from PathMNIST	**Not independent.** This subset is drawn from the same source family from which PathMNIST was derived, so source images represented here may also be represented, in downsampled form, in PathMNIST.

Because patient-level independence cannot be confirmed, the 224 × 224 result is reported throughout the manuscript as **image-level validation performance on a public subset**, and explicitly not as independent external validation. This wording appears in the Abstract, Methods, Results and Discussion.

PathMNIST

Field	Value
Kaggle mirror	dongquan/pathmnist-colon-pathology-dataset
URL	https://www.kaggle.com/datasets/dongquan/pathmnist-colon-pathology-dataset
File	pathmnist.npz, 205,615,438 bytes
Access date	4 July 2026 (re-verified 27 August 2026)
Benchmark	MedMNIST v2 (Yang et al., *Sci Data* 2023;10:41)
Splits	89,996 train / 10,004 validation / 7,180 test, 28 × 28 × 3 uint8
Test split origin	CRC-VAL-derived component of the source family, 50 patients

Duplicate screening was performed by SHA-256 digest comparison over every file. The verification is reproducible from the manifest without re-downloading the images.