

Label-Free Structure Discovery in Breast Cancer Histopathology: K-Means and Fuzzy C-Means on Handcrafted Descriptors under Subtype-Disjoint Cross-Validation

Julius Odili

Covenant University

Omoju John

`john.omojupgs@stu.cu.edu.ng`

Covenant University

Adebisi Olumide

Lagos State University Teaching Hospital

Adewole Martins

Apexibyte

Research Article

Keywords: breast cancer histopathology, unsupervised clustering, handcrafted features, K-Means, Fuzzy C-Means, subtype-disjoint cross-validation, label-efficient benchmarking

Posted Date: August 25th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10796929/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

There will be over 3.2 million new breast cancer cases a year by 2050 and most of these will be in low- and middle-income countries (LMICs) where pathologists are limited. The current state-of-the-art approach to computational pathology is supervised deep learning, which relies on the availability of large collections of annotated images which are hard to be assembled in such setting. First, we explore the ability of hand-crafted morphological and textural descriptors with only unsupervised clustering of the descriptors to distinguish benign from malignant breast histopathology.

Methods

Seven-thousand, nine-hundred and nine images from 82 patients were analysed, with 148 features per image extracted from five feature families (GLCM, LBP, Gabor, morphological, statistical intensity) after five stages of pre-processing (resize to 224×224, Macenko stain normalisation, Gaussian filtering, median filtering, and CLAHE applied to the LAB luminance channel). Features were z-scored and reduced by PCA to 21–22 features at a 95% variance threshold (85.6% reduction), fit on training data only. K-Means and Fuzzy C-Means were compared under a subtype-disjoint 4-fold cross-validation, where each test fold held out one full benign and one full malignant tumour subtype, ensuring patient-level disjointness and assessing cross-subtype generalisation.

Results

K-Means selected $k = 2$ in three folds and $k = 4$ in the fourth; Fuzzy C-Means selected $c = 2$ in all folds. K-Means showed higher internal validity than Fuzzy C-Means (silhouette 0.2197 ± 0.0450 vs. 0.1740 ± 0.0376 ; Davies–Bouldin 1.8113 ± 0.1378 vs. 2.0328 ± 0.1474) and comparable agreement with diagnosis under a post-hoc oracle cluster-to-class mapping (accuracy 0.6871 ± 0.1559 vs. 0.6854 ± 0.1349), against an all-malignant constant-predictor baseline of 0.600. In 3 of 8 algorithm×fold evaluations the post-hoc mapping degenerated to a constant malignant assignment, so aggregate sensitivity figures should not be interpreted as detection performance.

Conclusions

Handcrafted descriptors carry a modest signal aligned with malignancy, but not enough to support diagnostic use. This is a preliminary, label-free benchmark rather than a diagnostic system: clinical validity is not established, the mapping uses test labels, and evaluation is retrospective on a single Brazilian cohort.

1 Introduction

Breast cancer is the most common cancer in the world, accounting for around 2.3 million new cases and 685,000 deaths in 2020 [1]. By 2050, the number of cases is estimated to be over 3.2 million per year [2, 3] and will increasingly occur in low- and middle-income countries. The epidemiological profile in Nigeria and sub-Saharan Africa is distinct from that in Western countries where triple-negative types are more prevalent, the age of presentation is approximately 10 years lower and the 5-year survival is less than 50% [4, 5]. Limited access to computational diagnostic infrastructure is a contributing factor, as is late presentation and severe pathologist shortage [6].

The gold standard for breast cancer diagnosis is histopathological examination of haematoxylin and eosin (H&E)-stained tissue, although the process is labor-intensive and subject to inter-reader variability [7]. This is the motivation to propose the use of computer-aided diagnosis (CAD) to overcome this limitation and supervised deep learning systems are known to reach high accuracy rate on benchmark corpora like BreaKHis [8, 9]. There are two challenges to transferring that work to resource poor environments. First, supervised models are usually data intensive, and they also need properly labelled ground truth data which can be quite time consuming, expensive, and difficult to acquire and manually annotated by pathologists [10]. Second, many published BreaKHis evaluations partition the data at the image level, where images from the same patient can be found in both the training and test sets, leading to inflated evaluation scores [11].

Unsupervised clustering is an alternative exploratory approach without the need for class labels for model fitting. For medical image analysis and cancer analysis, K-Means and Fuzzy C-Means (FCM) have been used to segment image regions depending on the inherent intensity, colour, and texture features of the image, and handcrafted image descriptors like GLCM and other texture features are able to represent these image properties in compact representation [13, 14].

The question that this study attempts to answer is narrow and empirical: At what level of disjointness of the evaluation partition, coarser than the patient level, can the benign/malignant distinction be retrieved from hand-crafted descriptors through clustering alone? No diagnostic system is proposed. We describe what it can do, what it cannot do, and where it cannot do so, so that these methods can be compared to an "absolute minimum" standard of performance against label-efficient and semi-supervised methods.

1.1 Contributions

- A fully specified and reproducible label-free pipeline for breast histopathology, comprising five stages of preprocessing, 148 handcrafted features per image across five descriptor families, z-scoring and PCA, and two clustering algorithms with hyperparameter optimization during training.
- Subtype-disjoint cross validation: Each test fold contains one full benign and one full malignant tumour subtype. Each patient has exactly one subtype, ensuring that patient-level disjointness as a consequence and further ensuring generalisation to morphologies not included in training.
- Clear separation of three phases of evaluation that are often mixed: unsupervised model construction, label-free internal evaluation of model validity, and label-based interpretation of clusters in a post-hoc

fashion. The third phase of this study uses test labels and is therefore reported as upper bound of oracle performance (not predictive performance).

- A description of the failure modes of post-hoc cluster labelling in class imbalance, with three of eight algorithm×fold evaluations resulting in a constant assignment, plus a baseline of constant predictor assignments as a point of comparison for all reported figures.

2 Methods

2.1 Dataset

BreaKHis [15] is a collection of 7,909 H&E histopathological images obtained from 82 patients at four magnifications (40×, 100×, 200×, 400×), stored in 700×460 pixel 3-channel RGB PNG images. There were 2,480 benign images and 5,429 malignant images, with a benign to malignant image ratio of around 31:69, across four subtypes (adenosis, fibroadenoma, phyllodes tumour and tubular adenoma) for benign and four subtypes (ductal, lobular, mucinous and papillary carcinoma) for malignant. Images were distributed across magnifications as 40× (1,995), 100× (2,081), 200× (2,013) and 400× (1,820).

The images were considered to be independent samples. Magnification was measured but not used as a stratification variable, as a fusion axis, nor as a covariate in any analysis; its possible consequences are discussed in § 4.7.

2.2 Image preprocessing

Each image was subjected to five successive operations. (i) Resize to 224×224 size bilinearly. (ii) Macenko stain normalisation [16]: The stain vectors of H&E were estimated by singular value decomposition of the optical density matrix, and the concentration of stains was rescaled to a reference profile. The reference statistics were fitted one time on a randomly drawn image (SOB_M_LC-14-15570C-40-003), the target means were [1.1807, 0.6439] and the target standard deviations were [0.2363, 0.2117] with $\beta = 0.15$, and the robust angle percentile was $\alpha = 1.0$. (iii) Gaussian smoothing using 5×5 kernel and $\sigma = 1$. (iv) 5×5 Median Filter. (v) CL-AHE (contrast limited adaptive histogram equalisation) on L channel in LAB colour space with clip limit = 2.0 and 8×8 tile grid, and then back to BGR.

In writing the result of step (i) must be stated explicitly. The horizontal scale factor of 0.320 and vertical scale factor of 0.487 differ from each other, and have an anisotropy of 1.52, for the resized target aspect ratio of 224×224, when compared to the source image aspect ratio of 700×460. So, throughout the corpus horizontal structure has been compressed in relation to vertical structure. The impact is not evenly distributed across the descriptors of § 2.3, some of the descriptors are computed on distorted shapes, such as contour perimeter, solidity, extent and the circularity proxy are computed on different physical distances, for offsets at 0° and 90°, respectively. The impact is not uniform across each of the descriptors of § 2.3: the contour perimeter, solidity, extent and the circularity proxy are computed on distorted shapes, and offsets at 0° and 90° span different physical distances. The distortion is the same for all images, so it will not cause any confusion when comparing folds or comparing algorithms, but it will also result in the

morphological features not being measures of actual tissue geometry. It would be removed if you resize it to be isotropic (pad to a square first then resize).

2.3 Handcrafted feature extraction

Each image was preprocessed and five descriptor families were computed, resulting in a 148-dimensional vector for each. Every descriptor of texture was computed on the greyscale version, while intensity statistics were computed per colour channel.

GLCM. Grey-level co-occurrence matrices are created for 256 grey levels, symmetric and normalised, at distances, $d = \{1,3,5\}$ and orientations, $\theta = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$. Each of the 12 (d, θ) pairs yielded four properties: contrast, homogeneity, energy and correlation, which were stored separately, rather than averaged over the orientations, leading to a total of $4 \times 12 = 48$ features.

LBP. Uniform local binary patterns with $P = 8r$ sampling points and at the distance $r \in \{1,2,3\}$. A 10-bin density-normalised histogram of the pattern map was kept for each radius, resulting in $3 \times 10 = 30$ features.

Gabor. A filter bank of six orientations (0 to $5\pi/6$ in steps of $\pi/6$) and four spatial frequencies (0.1, 0.2, 0.3, 0.4), with 9×9 kernels, $\sigma = 4.0$, $\gamma = 0.5$ and $\psi = 0$. The mean and standard deviation of each filter response was preserved, resulting in $6 \times 4 \times 2 = 48$ features.

Morphological. The process of retaining the contours with an area exceeding 10 pixels has been performed using Otsu thresholding followed by external contour extraction. The following set of 10 scalars were calculated: contour count, mean, standard deviation and maximum contour area, mean perimeter, mean solidity, mean extent, circularity proxy (mean area/mean perimeter of maximum contour area squared), and standard deviations of extent and solidity.

Statistical intensity. The mean, standard deviation, skewness and kurtosis have been calculated for each of the three BGR channels, which results in 12 features.

Table 1

Handcrafted feature families, parameter settings and dimensionality. All counts verified against the extracted feature matrix, whose shape was $7,909 \times 148$.

Family	Descriptor	Parameters	Features
Texture — GLCM	Grey-level co-occurrence	$d \in \{1,3,5\}$; $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$; 4 properties, not averaged	48
Texture — LBP	Uniform local binary patterns	$r \in \{1,2,3\}$; $P = 8r$; 10-bin density histogram per radius	30
Texture — Gabor	Gabor filter bank	6 orientations $\times$ 4 frequencies; mean and SD of response	48
Morphological	Otsu + external contours	Area > 10 px; 10 aggregate shape statistics	10
Statistical intensity	Per-channel moments	Mean, SD, skewness, kurtosis $\times$ 3 BGR channels	12
Total per image	—	—	148

2.4 Standardisation and dimensionality reduction

Features were z scored to mean 0 and variance 1, within each cross-validation fold, based on statistics in the training partition. After standardisation, the principal component analysis was then fit on the training portion and the number of components fitted was selected to explain 95% of the cumulative variance, and the principal component projection was applied to the test portion without refitting. No test data were given to the scaler or the projection.

2.5 K-Means clustering

The number of initialisations was automatically chosen and the number of iterations was capped at 500, using k-means ++ initialisations and a fixed random seed (42) for K-Means. Eight configurations per fold were obtained by searching the cluster count $k \in \{2, \dots, 9\}$. The optimum value of k , which maximised the silhouette coefficient on the training partition, was preserved and the model that was fitted used to classify the test images into clusters. Selection did not employ any class labels.

2.6 Fuzzy C-Means clustering

Fuzzy C-Means calculates membership vector u , $\sum u = 1$, so that membership diffuseness is controlled by the number m . There were twelve combinations with $c \in \{2,3,4,5\}$ and $m \in \{1.5, 2.0, 2.5\}$ for each fold, and up to 1,000 iterations and tolerance of 5×10^{-3} . Those of the configurations that collapsed to a single occupied cluster were not considered. A configuration which maximises the training silhouette coefficient over hardened labels was kept, and this configuration's centroids were used to calculate test memberships. The argmax of the membership vector was used to obtain hard labels. No class labels were used for selection.

2.7 Subtype-disjoint cross-validation

The grouping units were done at the level of tumour subtype and cross validation was carried out. There are eight types of BreakHis, four of which are benign and four of which are malignant. These eight units were then stratified using a fixed seed of 42 and k-fold partitioning with shuffling of the units. Since the minority class size is only four, there was no option to make five folds of the dataset and the protocol was reduced to four folds, such that each test partition has exactly one benign and one malignant subtype, while the other six are in the training partition.

Two properties follow. First, a straightforward distinction between the two types of partitioning: no image of a patient can be in both of the partitions, hence it is clearly stronger to have the partitions be subtype-disjoint as opposed to patient-disjoint, since image-level leakage and patient-level leakage are both excluded by construction. Second, the protocol assesses the generalisation to tumour morphologies completely lacking from the training partition, a harder and different task than within-subtype patient-level cross-validation. Here, results are reported on a per patient basis only with the above-mentioned difference.

There is also a protocol cost, which must be clearly outlined. Stratification balances the number of subtypes per fold, NOT patients or images. The test partitions then consist of 1,129-3,895 images and 12–42 patients, while the benign image fraction of test partitions ranges from 11.4% to 61.8% compared to a corpus-wide fraction of 31.4%. All summed results in § 3 are thus heavily dependent on the distribution of the subtypes in the folds, and should be viewed in the light of the per-fold distribution of subtypes as given in Table 2.

2.8 Internal cluster validity

The silhouette coefficient [18] (higher the better) indicates a contrast between intra-cluster cohesion and the nearest-cluster separation; the Davies–Bouldin index [19] (lower the better) is the mean worst-case similarity between cluster pairs; and the Calinski–Harabasz index [20] (higher the better) is the ratio of between-cluster dispersion to within-cluster dispersion. The Calinski–Harabasz index is not independent of the number of samples in a fold, so cannot be compared between folds of different size; see § 3.6 for a quantification.

2.9 Post-hoc label-based interpretation

The ground-truth classes of the test images in each cluster were mapped to the majority class, and accuracy, malignant recall (sensitivity), benign recall (specificity) and F1 were calculated based on this mapping for each cluster to relate them to diagnosis. There are three aspects of this procedure that warrant emphasis.

- The mapping is done on test labels. It is an oracle procedure, and the numbers that come out of it are an optimistic upper bound on the number that the same clustering would get if it were based on a mapping estimated from the training data. They are not predictive performance and no classifier is assumed to be deployable.
- When class imbalance exists, the mapping may be degenerate: If all the test images belong to one of the clusters that is associated with malignant, then the sensitivity equals to 1.000 and the specificity

equals to 0.000, regardless of the quality of the cluster. This type of folding is made very clear in § 3.7.

- If the number of clusters chosen is more than two, the mapping reduces multiple clusters to two classes. In this context, the binary metrics of a such a fold are not commensurate with those of a two-cluster fold (see § 4.4).

No class labels were used for any of the clustering algorithms: either when fitting or during hyperparameter optimization. All label dependence occurs in this, the interpretative, stage.

2.10 Reference baseline and aggregation

The test partitions are significantly imbalanced, and imbalanced in different ways, and the suitable benchmark for the accuracy of the post-hoc mapping is a constant predictor which assigns the majority class of each test partition. This baseline achieves 0.6361, 0.8860, 0.4960 and 0.3817 in Folds 1 to 4 respectively, mean 0.600. All accuracy statements therein should be taken with reference to it. Cross-fold summaries are presented as unweighted means $\pm$ standard deviations for four folds. None of the four folds are sufficiently powered to conduct an adequate inferential test and none were conducted, nor is there any claim of statistical significance anywhere in this paper.

3 Results

3.1 Cross-validation partitions

The actual composition of each fold is presented in Table 2. The membership of the fold was recovered by rerun of the original seeded stratified partition and the numbers of the training and test images are exactly reproduced (6,664/1,245; 4,014/3,895; 6,780/1,129; 6,269/1,640), which confirms the reconstruction.

Table 2

Composition of the four subtype-disjoint cross-validation folds. Each test partition holds out one benign and one malignant subtype in full. Patient counts follow from the published BreakHis subtype-to-patient distribution.

Fold	Held-out subtypes (test)	Test patients	Test images	Train images	Test benign %
1	Phyllodes tumour (453) + mucinous carcinoma (792)	12	1,245	6,664	36.4
2	Adenosis (444) + ductal carcinoma (3,451)	42	3,895	4,014	11.4
3	Tubular adenoma (569) + papillary carcinoma (560)	13	1,129	6,780	50.4
4	Fibroadenoma (1,014) + lobular carcinoma (626)	15	1,640	6,269	61.8

3.2 Feature extraction and dimensionality reduction

The feature extraction resulted a 7,909 x 148 matrix. Texture descriptors (GLCM (48), LBP (30) and Gabor (48)) account for 85.1% of the 148 dimensions, morphological descriptors account for 10 dimensions and intensity statistics for 12 dimensions.

On average, $21.25 \pm 0.50\%$ of the variance was retained using PCA at the 95% variance threshold (Table 3); the number of retained components was 21 in three folds and 22 in one. This is equivalent to an 85.6% dimensionality reduction. The small difference does not necessarily imply that the intrinsic dimensionality of the descriptor space remains stable with respect to any partition in the set of disjoint partitions of the data by subtypes, nor does it necessarily imply any statement about overfitting or generalisation.

Table 3
PCA dimensionality reduction per fold at the 95% cumulative variance threshold.

Fold	Components retained	Variance explained (%)	Input dimensionality
1	21	95.07	148
2	21	95.12	148
3	22	95.44	148
4	21	95.05	148
Mean $\pm$ SD	21.25 $\pm$ 0.50	95.17 $\pm$ 0.18	148

3.3 K-Means

K-Means selected $k = 2$ in Folds 1, 2 and 3 and $k = 4$ in Fold 4 (training silhouette 0.2446, 0.2342, 0.2385 and 0.1824 respectively). The mean test silhouette was 0.2197 ± 0.0450 and mean Davies–Bouldin index 1.8113 ± 0.1378 , indicating moderate cohesion and moderate separation. The Calinski–Harabasz index was on average 427.96 ± 296.27 , and the value of 867.42 in Fold 2 seems to be an artefact of test-partition size (see § 3.6).

Under the post-hoc oracle mapping, mean accuracy was 0.6871 ± 0.1559 , sensitivity 0.7439 ± 0.1939 , specificity 0.4084 ± 0.3453 and F1 0.7087 ± 0.1669 . The accuracy is 8.7 percentage points better than the baseline of 0.600 that was achieved with a constant predictor. This high diversity in specificity is due to one degenerate fold that is identified in § 3.7.

Table 4

K-Means results per fold. k^* is the cluster count selected by silhouette maximisation on training data. $\uparrow$ higher is better; $\downarrow$ lower is better. Fold 4 is not commensurate with the others because $k^* = 4$ (§ 4.4); Fold 2 has a degenerate post-hoc mapping (§ 3.7).

Fold	k^*	Silhouette $\uparrow$	DBI $\downarrow$	CHI $\uparrow$	Accuracy	Sensitivity	Specificity	F1
1	2	0.2208	1.8529	290.32	0.6490	0.7184	0.5276	0.7225
2	2	0.2737	1.6758	867.42	0.8860	1.0000	0.0000	0.9396
3	2	0.2208	1.9858	223.61	0.5093	0.7286	0.2935	0.5956
4	4	0.1637	1.7309	330.48	0.7043	0.5288	0.8126	0.5772
Mean $\pm$ SD	—	0.2197 $\pm$ 0.0450	1.8113 $\pm$ 0.1378	427.96 $\pm$ 296.27	0.6871 $\pm$ 0.1559	0.7439 $\pm$ 0.1939	0.4084 $\pm$ 0.3453	0.7087 $\pm$ 0.1669

3.4 Fuzzy C-Means

FCM selected $c = 2$ in all four folds, with $m = 1.5$ in Folds 1 to 3 and $m = 2.0$ in Fold 4. Selection of the smallest admissible fuzzification exponent in three of four folds suggests that the silhouette criterion emphasized relatively sharp partitions, and at least that K-Means is doing as well or better; it is not sufficient to guarantee that the feature-space clusters are well-separated.

Mean test silhouette was 0.1740 ± 0.0376 , Davies–Bouldin 2.0328 ± 0.1474 and Calinski–Harabasz 399.63 ± 301.72 , all inferior to K-Means. Post-hoc accuracy was 0.6854 ± 0.1349 , sensitivity 0.7640 ± 0.2747 , specificity 0.3400 ± 0.3930 and F1 0.6993 ± 0.1954 .

Table 5

Fuzzy C-Means results per fold. c^* and m^* are the cluster count and fuzzification exponent selected on training data. Folds 1 and 2 have degenerate post-hoc mappings (§ 3.7).

Fold	c^*	m^*	Silh. $\uparrow$	DBI $\downarrow$	CHI $\uparrow$	Accuracy	Sensitivity	Specificity	F1
1	2	1.5	0.1419	2.0899	218.76	0.6361	1.0000	0.0000	0.7776
2	2	1.5	0.2284	1.8181	844.47	0.8860	1.0000	0.0000	0.9396
3	2	1.5	0.1638	2.1527	205.75	0.5943	0.4857	0.7012	0.5429
4	2	2.0	0.1619	2.0705	329.53	0.6250	0.5703	0.6588	0.5372
Mean $\pm$ SD	—	—	0.1740 $\pm$ 0.0376	2.0328 $\pm$ 0.1474	399.63 $\pm$ 301.72	0.6854 $\pm$ 0.1349	0.7640 $\pm$ 0.2747	0.3400 $\pm$ 0.3930	0.6993 $\pm$ 0.1954

3.5 Comparative performance

The mean values obtained by K-Means were higher in all three of the internal validity indices, and non-overlapping range in silhouette (0.164–0.274 vs. 0.142–0.228) and Davies–Bouldin (1.676–1.986 vs.

1.818–2.153). Post hoc accuracies vary by 0.0017, which is within a reasonable fold to fold range of ~ 0.14. Both FCM and K-Means had higher mean sensitivity (0.7640 vs. 0.7439) and higher mean specificity (0.4084 vs. 0.3400), with the increases being largely due to the number of degenerate folds that each algorithm generated, and not due to any difference in discriminative behaviour. No formal hypothesis test was performed and none would be adequately powered with four folds and the comparison below is descriptive.

The following comparison, in Table 6, requires special emphasis. Both algorithms have a mean F1 lower than the constant predictor (0.7087 and 0.6993, respectively, when compared with 0.7332). This is not unusual: F1 is an ideal summary of an all-malignant assignment and not a good one under this class prior. It nevertheless emphasizes that F1 should not be used as an indicator of performance in this task and that accuracy in both algorithms do better than the baseline, it is the one of the four post-hoc measures on which neither algorithm beats giving all images to the maximum class.

Table 6
Comparative summary across four folds (unweighted mean $\pm$ SD), with the constant-predictor reference. No significance testing was performed.

Metric	K-Means	FCM	Constant predictor
Silhouette $\uparrow$	0.2197 $\pm$ 0.0450	0.1740 $\pm$ 0.0376	n/a
Davies–Bouldin $\downarrow$	1.8113 $\pm$ 0.1378	2.0328 $\pm$ 0.1474	n/a
Calinski–Harabasz $\uparrow$	427.96 $\pm$ 296.27	399.63 $\pm$ 301.72	n/a
Accuracy	0.6871 $\pm$ 0.1559	0.6854 $\pm$ 0.1349	0.6000 $\pm$ 0.2172
Sensitivity	0.7439 $\pm$ 0.1939	0.7640 $\pm$ 0.2747	1.0000
Specificity	0.4084 $\pm$ 0.3453	0.3400 $\pm$ 0.3930	0.0000
F1	0.7087 $\pm$ 0.1669	0.6993 $\pm$ 0.1954	0.7332 $\pm$ 0.1654
Selected clusters	k = 2, 2, 2, 4	c = 2 in all folds	n/a
Fuzzification	n/a	m = 1.5, 1.5, 1.5, 2.0	n/a
Degenerate mappings	1 of 4 folds	2 of 4 folds	4 of 4 by construction

3.6 The Calinski–Harabasz index tracks test-partition size

The ratio of between- to within-cluster dispersion divided by the test-partition size is 0.2332, 0.2227, 0.1981 and 0.2015 for K-Means and 0.1757, 0.2168, 0.1822 and 0.2009 for FCM. These ratios are almost constant, while the raw index changes by almost 4 times. This large (4.2 $\times$) standard deviation in the Calinski–Harabasz index is therefore purely an artifact of the spread in the size of test-partitions and provides no information about cluster structure. It is mentioned here merely for its completeness and is to be taken as an indicator of validity of the classification and not as a subspecific indicator of a well separated benign subclass, as specificity in this fold is 0.000.

3.7 Degenerate post-hoc mappings

Of the eight algorithm×fold evaluations, the majority-vote mapping assigned the malignant label to all clusters in three of them, such that all test images were labeled as malignant. Table 7 presents the confusion matrix for each evaluation, derived from the sensitivities and specificities reported in the tables, along with the fold composition from Table 2; all elements in these tables are integers within 5×10^{-7} .

Table 7
Confusion matrices per algorithm and fold, recovered from reported metrics and fold composition. Rows marked † have a degenerate mapping in which every cluster received the malignant label; their sensitivity of 1.000 and specificity of 0.000 are properties of the mapping, not of the clustering.

Fold	Algorithm	TP	FN	TN	FP	Note
1	K-Means	569	223	239	214	
1	FCM	792	0	0	453	† all clusters → malignant
2	K-Means	3,451	0	0	444	† all clusters → malignant
2	FCM	3,451	0	0	444	† all clusters → malignant
3	K-Means	408	152	167	402	
3	FCM	272	288	399	170	
4	K-Means	331	295	824	190	
4	FCM	357	269	668	346	

The consequences for aggregate figures are material. Excluding the degenerate fold, K-Means attains mean accuracy 0.6209, sensitivity 0.6586 and specificity 0.5446 over Folds 1, 3 and 4. Excluding both degenerate folds, FCM attains mean accuracy 0.6097, sensitivity 0.5280 and specificity 0.6800 over Folds 3 and 4. The headline mean sensitivities of 0.7439 and 0.7640 are therefore substantially attributable to folds in which no discrimination occurred, and should not be described as detection performance.

3.8 Sensitivity analysis restricted to two-cluster folds

Restricting K-Means to the three folds in which $k = 2$ was selected gives mean accuracy 0.6814 ± 0.1904 , sensitivity 0.8157 ± 0.1597 , specificity 0.2737 ± 0.2644 , F1 0.7526 ± 0.1739 and silhouette 0.2384 ± 0.0306 . The restricted accuracy is only slightly lower than the four-fold mean, and the specificity is significantly lower, being the single fold with the highest specificity (0.8126) and the largest benign fraction (61.8%) being excluded. The logged output no longer allows for any further sensitivity analysis (per cluster assignments not saved).

4 Discussion

4.1 What the clustering structure shows, and what it does not

Observation. The obtained silhouette scores for K-Means range from 0.16 to 0.27, which is a weak-to-moderate level of cluster structure. The silhouette scores for FCM vary from 0.14 to 0.23, which also suggests a weak-to-moderate cluster structure. Under an oracle mapping, post-hoc accuracy of 0.687 is 8.7 percentage points greater than a constant predictor baseline accuracy of 0.600.

Interpretation. Handcrafted descriptors have some signal that falls along the benign/malignant axis, but this is not the dominant axis of variation in the feature space. The range of silhouette values are similar to the presence of substantial overlap between the emergent clusters. The margin is the actual margin over a constant predictor, but it is measured with the benefit of a mapping fitted to test labels; a mapping fitted at training time would only do as well, and likely not as well.

The former interpretation that this means that the benign/malignant axis is the most important axis of variation in the feature space is removed. That claim is not supported by the evidence: it would need, at least, an agreement statistic without mapping, a training-fitted mapping, and folds that didn't degenerate in the mapping.

4.2 K-Means versus Fuzzy C-Means

K-Means achieved the best mean internal validity values across all three indices and its improvements are observed in each fold on the silhouette and Davies–Bouldin indices. The choice of the smallest fuzzification exponent in three of four folds is in the same direction as FCM's choice of the smallest m in three of four folds: the criterion preferred crisper partitions, which one would expect when a hard partitioning method is already satisfactory. Post-hoc accuracy is virtually the same (0.6871 compared to 0.6854, with an average gap of ~ 0.14). The higher mean of FCM's sensitivity should not be interpreted as a benefit because it is due to the fact that FCM is a result of two folds of degeneration instead of one. There is no evidence on this basis that either algorithm is better for this purpose; K-Means has the advantages of internal validity and on computational cost.

4.3 Class imbalance and the limits of post-hoc mapping

The image ratio of 31/69 benign-to-malignant is highly correlated with the majority vote cluster labelling scheme. If a cluster consists of a mixture of classes in roughly corpus proportion then the majority of the classes in the cluster is malignant, otherwise the mapping collapses if all the clusters are in the same. Here, it happened in three of eight evaluations and the fold in which it happened for both algorithms (11.4% benign test fraction) is the same fold in which it was the most imbalanced. The sensitivity and specificity obtained by such a mapping are not measures of the clustering but are measures of the class prior. This is a general property of post-hoc cluster labelling under imbalance and does not seem to be peculiar to this dataset, and thus presents good arguments for reporting mapping-free agreement statistics like the adjusted Rand index along with any binary metrics. These were not calculated here (§ 4.8).

4.4-Fold 4 and the selection of four clusters

Observation. K-Means selected $k = 4$ in Fold 4, against $k = 2$ in the other three folds. Fold 4's test partition is fibroadenoma plus lobular carcinoma, with a benign image fraction of 61.8%, the highest of any fold and roughly double the corpus rate.

Interpretation. The reason for the selection cannot be established from the logged outputs. Silhouette was maximised on the training partition, not the test partition, so any explanation appealing to the composition of the test set is unsupported. Determining whether the four clusters correspond to subtype structure, magnification, staining variation or something else would require inspecting the cluster assignments against those covariates, which was not done.

The methodological consequence is definite regardless of cause. Collapsing four clusters onto two classes by majority vote is a different operation from mapping two clusters onto two classes, and the resulting binary metrics are not commensurate. Averaging Fold 4's clinical metrics with the other three is therefore not a well-defined operation, and the four-fold means in Tables 4 and 6 should be treated as descriptive summaries rather than as estimates of a single quantity. The restricted analysis in § 3.8 is reported for this reason. Internal validity indices are unaffected by this concern, since they are computed on the cluster labels themselves.

4.5 Leakage prevention and the cost of the partitioning scheme

Partitioning at subtype level excludes both image-level and patient-level leakage by construction, since patients' nest within subtypes. Standardisation and PCA were fitted on training partitions only, and both algorithms' hyperparameters were selected on training data without reference to class labels. On the specific question of leakage, this protocol is more conservative than the patient-level splitting recommended in the literature [11].

The cost is that the design is not a patient-level cross-validation and its results are not comparable to one. Each test partition contains morphologies never seen in training, which makes the task harder; and the eight grouping units admit only four folds, with test partitions varying 3.4-fold in size and 5.4-fold in benign fraction. A patient-level protocol over the 82 available patients would support five or ten folds with far more homogeneous partitions, and would answer the question most readers of a BreakHis paper expect to see answered. Reporting both protocols would be stronger than either alone; only the subtype-disjoint protocol was executed here.

4.6 Relation to prior work

On BreakHis, the accuracy of supervised systems ranges from 85–98% [8, 9]. They are also not comparable to us, because they have full label supervision, many have image-level partitioning to allow for patient level leakage [11] and none of which we are aware of testing under subtype-disjoint partitioning. This work can be compared to the constant-predictor baseline of § 2.10 and other future label-free or label-efficient methods tested in the same protocol.

Histological image feature clustering analyses have reported silhouette scores below about 0.25, which are indicative of weak or scant evidence of cluster formation [12]. Salloum [21] observed that the K-Means classification of the histogram-of-oriented-gradient features is able to separate benign from malignant tissue, and in a systematic comparison of several texture descriptors, Öztürk and Akdemir [22] showed that these descriptors are very suitable for image classification in histopathology. Likewise, our feature set is dominated by texture descriptors which represent 85.1% of the dimensions (126 out of 148 features). But the contribution in dimension alone is not enough to establish discriminative contribution and we did not quantify the latter.

4.7 Limitations

- Post-hoc cluster to class mapping is fitted on test labels. All accuracy, sensitivity, specificity and F1 values are best case scenarios and not representative of a viable classifier.
- Three of eight algorithm×fold combinations gave degenerate mappings, leading to overestimated aggregate sensitivity figures.
- Very limited statistical resolution, 4 folds over eight grouping units. There is no hypothesis test performed and no hypothesis test reported.
- The sizes of the test partitions vary by 3.4-fold and benign fractions by 5.4-fold, causing the unweighted cross-fold means to be dominated by fold composition.
- Magnification was recorded but never used. Images at 40× and 400× of the same slide were analysed as independent samples, were not fused, not stratified or covariate adjusted and no magnification-specific analysis was performed. The effect of magnification on the structure recovered, if any, is not measured and not accounted for.
- Images were downsampled from a 700 × 460 to 224 × 224 to exclude fine nuclear detail that would be captured by morphological and GLCM descriptors.
- Reference statistics were fitted using Macenko method to a single randomly selected image. The pipeline's sensitivity to the choice was not evaluated.
- BreakHis is a single institution Brazilian cohort. It cannot draw general conclusions about Nigerian or other African populations, the clinical rationale behind such work. This is not meant to amount to clinical validation and no conclusion regarding the diagnostic performance in any population can be drawn.
- Handcrafted descriptors have limitations in terms of the level of representation achieved compared to learned or self-supervised ones.

4.8 Recommended additional analyses

The following were not executed, and are not results, but future work. Cross validation at patient level was performed on 82 subjects (in addition to the cross validation performed at the protocol level which was performed on a protocol disjoint set). (iii) A cluster-to-class mapping that was estimated on the training portion and reused on the test portion, providing true predictive, not oracle numbers. (iii) Agreement statistics, like the adjusted Rand index and normalised mutual information, which are not affected by the

degeneracy problem outlined in § 3.7. (iv) Size-weighted cross-fold aggregation and/or median cross-fold aggregation. (v) Analysis by magnification – explicit multi-magnification fusion. (vi) Abolish feature families to quantify instead of assume the discriminative contribution of the texture descriptors.

5 Conclusion

We investigated unsupervised clustering of handcrafted descriptors for the recovery of the benign/malignant class distinction in the context of breast histopathology, using no training labels. PCA reduced the number of features from an image from 148 features, across five descriptor families, to 21–22, an 85.6% reduction at 95% variance. In the case of 4-fold cross validation, done in a manner such that the 4 folds are mutually exclusive, K-Means selected two clusters in 3 folds and four in the fourth, and Fuzzy C-Means selected two clusters in each fold. On all three indices (silhouette 0.2197 vs. 0.1740; Davies–Bouldin 1.8113 vs. 2.0328; and under a post-hoc oracle mapping, accuracy 0.6871 vs. 0.6854), the two algorithms showed no significant differences when compared to the constant-predictor baseline of 0.600.

Fairly, descriptors that are handcrafted have a small, genuine signal in the direction of malignancy, about nine points above the simple predictor, using the benefit of a mapping fitted on test labels, but this signal is not of sufficient strength to support any diagnostic purpose. Three out of eight evaluations gave degenerate mappings where there was no discrimination at all. What this work provides, is a well-defined and clearly reported protocol for a floor, a leakage-free protocol whose limitations are specified, and against which label-efficient and semi-supervised methods can be compared. Any conclusions on clinical utility would run counter to the data available in Nigeria and would need to be evaluated locally, without the use of test labels, and be prospective. These are all conditions that do not apply here.

Implementation and Code Availability

The analysis code, experiment configuration and materials for the reproducibility are available online on GitHub and on Zenodo. The source repository can be obtained from <https://github.com/JohnnyBrains/breakhis-label-free-clustering> and the version used in this study has been permanently archived as release v1.0.0. The corresponding Zenodo archive is available at <https://doi.org/10.5281/zenodo.21962011> (DOI: 10.5281/zenodo.21962011). Experiments were run in Google Colab, with Python 3.12. The initialisation for Fuzzy C-Means was not seeded and cannot be reproduced, the only fixed random seed used was for the Macenko reference image selection, the cross-validation partition and the initialisation of the K-Means. The Figs. 2–6 were reprojected at the resolution of this publication directly from the per-fold results logged. No value is plotted that differs from the result of the analysis. Figure 1 is copied verbatim from the analysis output. The repository includes the analysis notebook and related code and data files needed to replicate the reported analyses and figures.

Table 8
Software environment, as resolved at runtime in the execution environment.

Component	Version	Role
Python	3.12	Core environment (Google Colab)
NumPy	2.0.2	Array computation and linear algebra
pandas	2.2.2	Results tabulation
OpenCV (opencv-python)	4.13.0.92	Image I/O, resizing, filtering, CLAHE, Gabor kernels
scikit-image	0.25.2	GLCM and local binary pattern extraction
scikit-learn	1.6.1	StandardScaler, PCA, K-Means, cross-validation, metrics
scikit-fuzzy	0.5.0	Fuzzy C-Means clustering and prediction
SciPy	1.16.3	Skewness, kurtosis, mode
Matplotlib	3.10.0	Figure generation
tqdm	4.67.3	Progress reporting

Macenko stain normalisation was implemented directly from the optical density decomposition rather than through a third-party package. Seaborn was not used. The staintools and spams packages appear in the environment setup as an intended fallback but were not imported, and the spams build did not complete; neither is a dependency of the analysis as executed.

Declarations

Disclosure of Interest

The authors do not have any competing interests with regard to the present work.

Author Contribution

J.O. conceived and supervised the study. O.J. designed the methodology, developed and executed the analysis pipeline, performed the experiments and statistical analysis, and drafted the manuscript. A.O. contributed clinical and pathological context and critically reviewed the manuscript. A.M. contributed to software implementation and critically reviewed the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors are grateful to Covenant University for providing institutional support. The BreakHis data was made public for use with the aforementioned terms.

Data Availability

The dataset analysed in this study, BreakHis, is publicly available from the Laboratório Visão Robótica e Imagem (VRI), Universidade Federal do Paraná, at <https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/>. The analysis code, experiment configuration, and materials needed to reproduce the reported results are available on GitHub at <https://github.com/JohnnyBrains/breakhis-label-free-clustering>, and the version used in this study has been permanently archived on Zenodo at <https://doi.org/10.5281/zenodo.21962011>.

References

1. Sung H, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209–49. <https://doi.org/10.3322/caac.21660>.
2. Bray F, et al. Global cancer statistics 2022: GLOBOCAN estimates for 36 cancers in 185 countries. *CA Cancer J Clin*. 2024;74(3):229–63. <https://doi.org/10.3322/caac.21834>.
3. Freihat O, Sipos D, Kovacs A. Global burden and projections of breast cancer incidence and mortality to 2050: a comprehensive analysis of GLOBOCAN data. *Front Public Health*. 2025;13:1622954. <https://doi.org/10.3389/fpubh.2025.1622954>.
4. Gbaa LZ, et al. Breast cancer in Makurdi, North Central Nigeria: 10-year histopathological retrospective survey. *West Afr J Med Biomed Sci*. 2025;4(1):45–58.
5. Jedy-Agba E, et al. Cancer incidence in Nigeria: Report from the Nigerian National System of Cancer Registries. *Cancer Epidemiol*. 2020;64:101647. <https://doi.org/10.1016/j.canep.2019.101647>.
6. Wilson ML, Fleming KA, Kuti MA. The global pathologist shortage. *Lancet Glob Health*. 2021;9(8):e1045–6. [https://doi.org/10.1016/S2214-109X\(21\)00261-0](https://doi.org/10.1016/S2214-109X(21)00261-0).
7. Kang HJ, Kwon SY, Kim A, Kim WG, Kim EK, Kim AR, Kim C, Min SK, Park SY, Sung SH, Yoon HK, Lee A, Lee JS, Lee HI, Lee HC, Lim SC, Jun SY, Jung MJ, Jung CW, Cho SY, Cho EY, Choi HJ, Park SY, Kim JY, Park IA, Kwon Y. A multicenter study of interobserver variability in pathologic diagnosis of papillary breast lesions on core needle biopsy with WHO classification. *J Pathol Transl Med*. 2021;55(6):380–7. <https://doi.org/10.4132/jptm.2021.07.29>.
8. Litjens G, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60–88. <https://doi.org/10.1016/j.media.2017.07.005>.
9. Campanella G, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019;25(8):1301–9. <https://doi.org/10.1038/s41591-019-0508-1>.
10. Abels E, Pantanowitz L, Aeffner F, Zarella MD, van der Laak J, Bui MM, Vemuri VNP, Parwani AV, Gibbs J, Agosto-Arroyo E, Beck AH, Kozlowski C. Computational pathology definitions, best practices, and

- recommendations for regulatory guidance: a white paper from the Digital Pathology Association. *J Pathol.* 2019;249(3):286–94. <https://doi.org/10.1002/path.5331>.
11. Priyanka V, Narendra M, Dilleswar Rao T. A patient-aware benchmarking of CNN and transformer architectures for breast cancer histopathology classification. *Front Digit Health.* 2026;8:1752938. <https://doi.org/10.3389/fdgth.2026.1752938>.
 12. Prezja F, Pölönen I, Äyrämö S, Ruusuvaori P, Kuopio T. H&E Multi-Laboratory Staining Variance Exploration with Machine Learning. *Appl Sci.* 2022;12(15):7511. <https://doi.org/10.3390/app12157511>.
 13. Kumar R, Srivastava R, Srivastava S. Detection and classification of cancer from microscopic biopsy images using clinically significant and biologically interpretable features. *J. Med. Eng.* 2015, 457906 (2015). <https://doi.org/10.1155/2015/457906>
 14. Althubiti SA, Paul S, Mohanty R, Mohanty SN, Alenezi F, Polat K. Ensemble learning framework with GLCM texture extraction for early detection of lung cancer on CT images. *Comput. Math. Methods Med.* 2022, 2733965 (2022). <https://doi.org/10.1155/2022/2733965>
 15. Spanhol FA, Oliveira LS, Petitjean C, Heutte L. A dataset for breast cancer histopathological image classification. *IEEE Trans Biomed Eng.* 2016;63(7):1455–62. <https://doi.org/10.1109/TBME.2015.2496264>.
 16. Macenko M et al. A method for normalizing histology slides for quantitative analysis. In: *Proc. IEEE ISBI.* : 2009, pp. 1107–1110. IEEE (2009). <https://doi.org/10.1109/ISBI.2009.5193250>
 17. Zuiderveld K. Contrast limited adaptive histogram equalization. In: Heckbert PS, editor. *Graphics Gems IV.* Academic; 1994. pp. 474–85.
 18. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math.* 1987;20:53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
 19. Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell.* 1979;1(2):224–7. <https://doi.org/10.1109/TPAMI.1979.4766909>.
 20. Caliński T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat.* 1974;3(1):1–27. <https://doi.org/10.1080/03610927408827101>.
 21. Salloum S. K-Means clustering and classification of breast cancer images using histogram of oriented gradients features and convolutional neural network models: Diagnostic image analysis study. *JMIR Form Res.* 2025;9:e71974. <https://doi.org/10.2196/71974>.
 22. Öztürk Ş, Akdemir B. Application of feature extraction and classification methods for histopathological image using GLCM, LBP, LBGLCM, GLRLM and SFTA. *Procedia Comput Sci.* 2018;132:40–6. <https://doi.org/10.1016/j.procs.2018.05.057>.

Figures

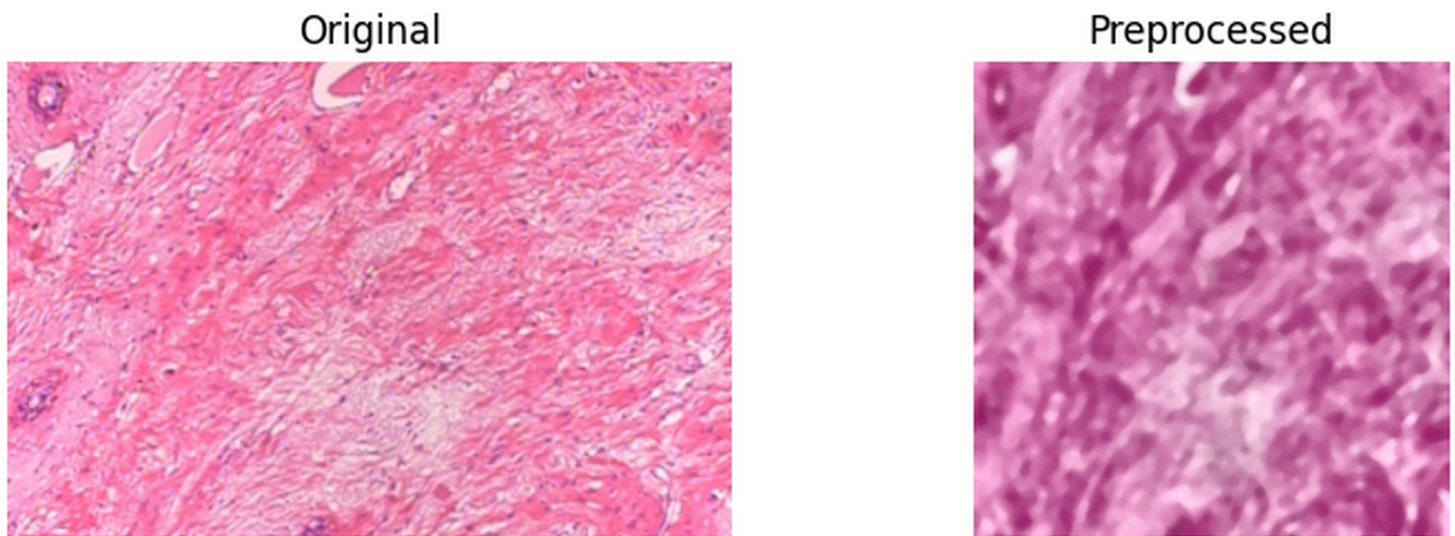


Figure 1

Preprocessing applied to a representative BreakHis image (SOB_M_LC-14-15570C-40-003). Left: original 700×460 image. Right: the same image after resizing to 224×224, Macenko stain normalisation to a fixed reference profile, Gaussian and median filtering, and CLAHE on the LAB luminance channel. The normalisation shifts the stain profile towards a deeper haematoxylin balance and the filtering visibly suppresses fine texture. The anisotropic resize is also apparent: horizontal structure in the right panel is compressed by a factor of 1.52 relative to the left.

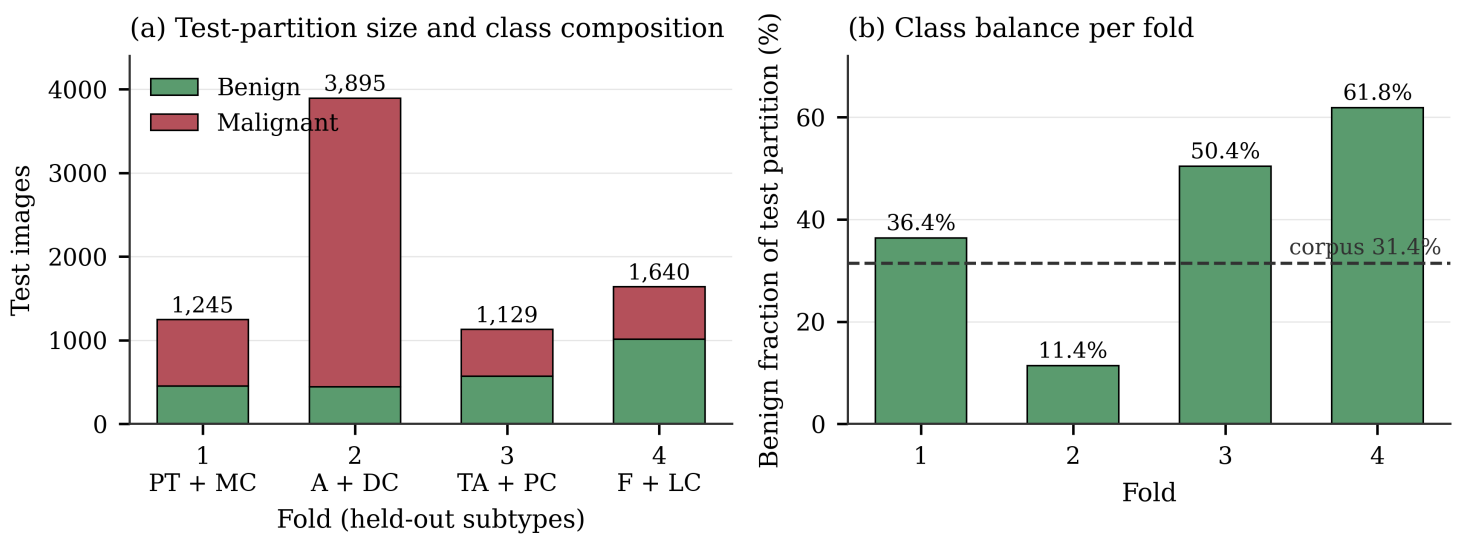


Figure 2

Composition of the four test partitions. (a) Absolute size and class composition; held-out subtype pairs are abbreviated PT (phyllodes tumour), MC (mucinous carcinoma), A (adenosis), DC (ductal carcinoma), TA (tubular adenoma), PC (papillary carcinoma), F (fibroadenoma), LC (lobular carcinoma). (b) Benign fraction of each test partition against the corpus-wide rate of 31.4% (dashed). Partitions vary 3.4-fold in size and

5.4-fold in benign fraction, so unweighted cross-fold means in Tables 4 to 6 are strongly influenced by fold composition.

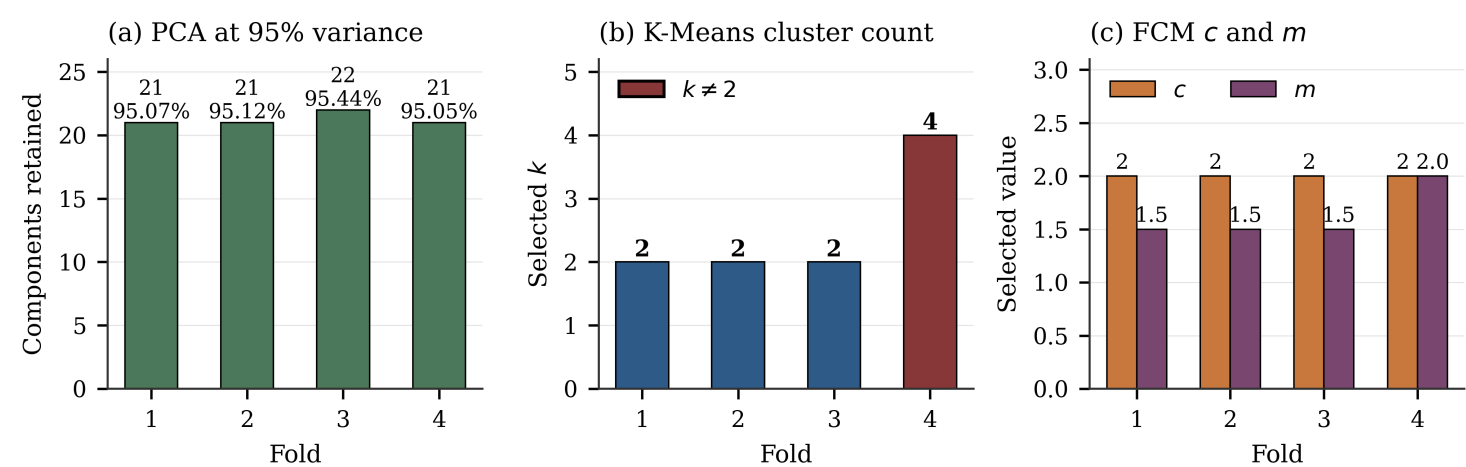


Figure 3

Dimensionality reduction and hyperparameter selection per fold, all determined on training partitions without reference to class labels. (a) Principal components retained at the 95% cumulative variance threshold, annotated with the variance actually retained. (b) Cluster count k selected for K-Means by silhouette maximisation; Fold 4 departs from the $k = 2$ selected elsewhere (red). (c) Cluster count c and fuzzification exponent m selected for Fuzzy C-Means; $c = 2$ throughout, with m at the lower grid bound in three folds.

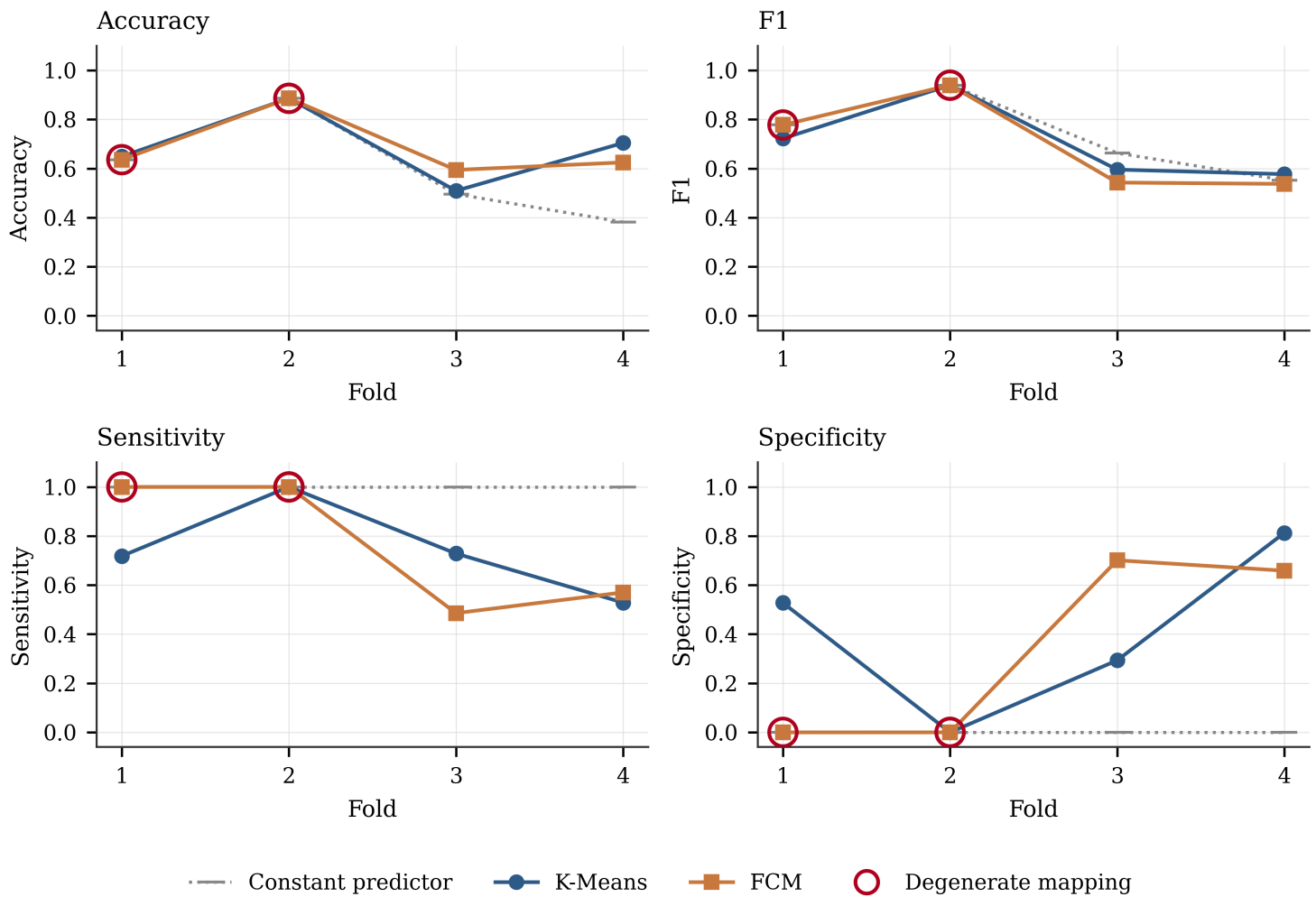


Figure 4

Post-hoc label-based metrics per fold for both algorithms, with the constant-predictor reference (dotted grey). Red rings mark the three evaluations in which the cluster-to-class mapping degenerated to a constant malignant assignment; their sensitivity of 1.000 and specificity of 0.000 are properties of the mapping rather than of the clustering. Accuracy and F1 track the constant predictor closely in Folds 1 to 3, with K-Means separating from it only in Fold 4.

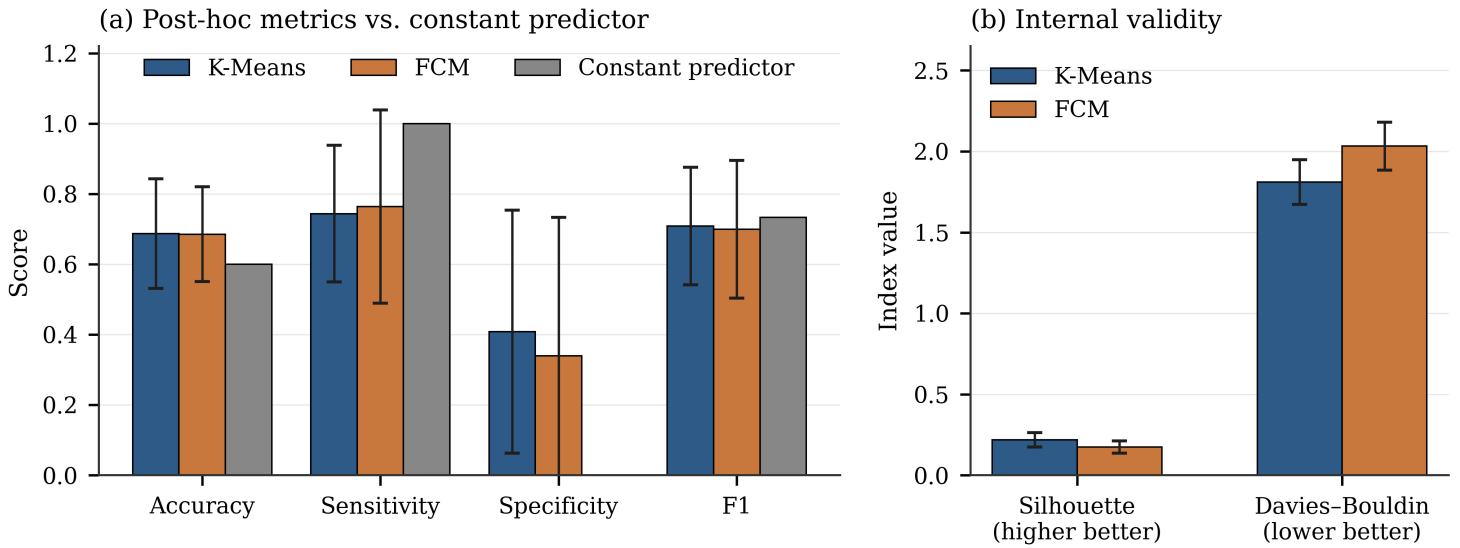


Figure 5

Cross-fold summary, unweighted means with error bars giving the sample standard deviation over four folds. (a) Post-hoc metrics for both algorithms against the constant-predictor reference. Accuracy is the only measure on which either algorithm exceeds the baseline; on F1 both fall below it. These figures derive from a mapping fitted on test labels and are optimistic upper bounds. (b) Internal validity indices, which are label-free and unaffected by the mapping. Numerical values are given in Table 6.

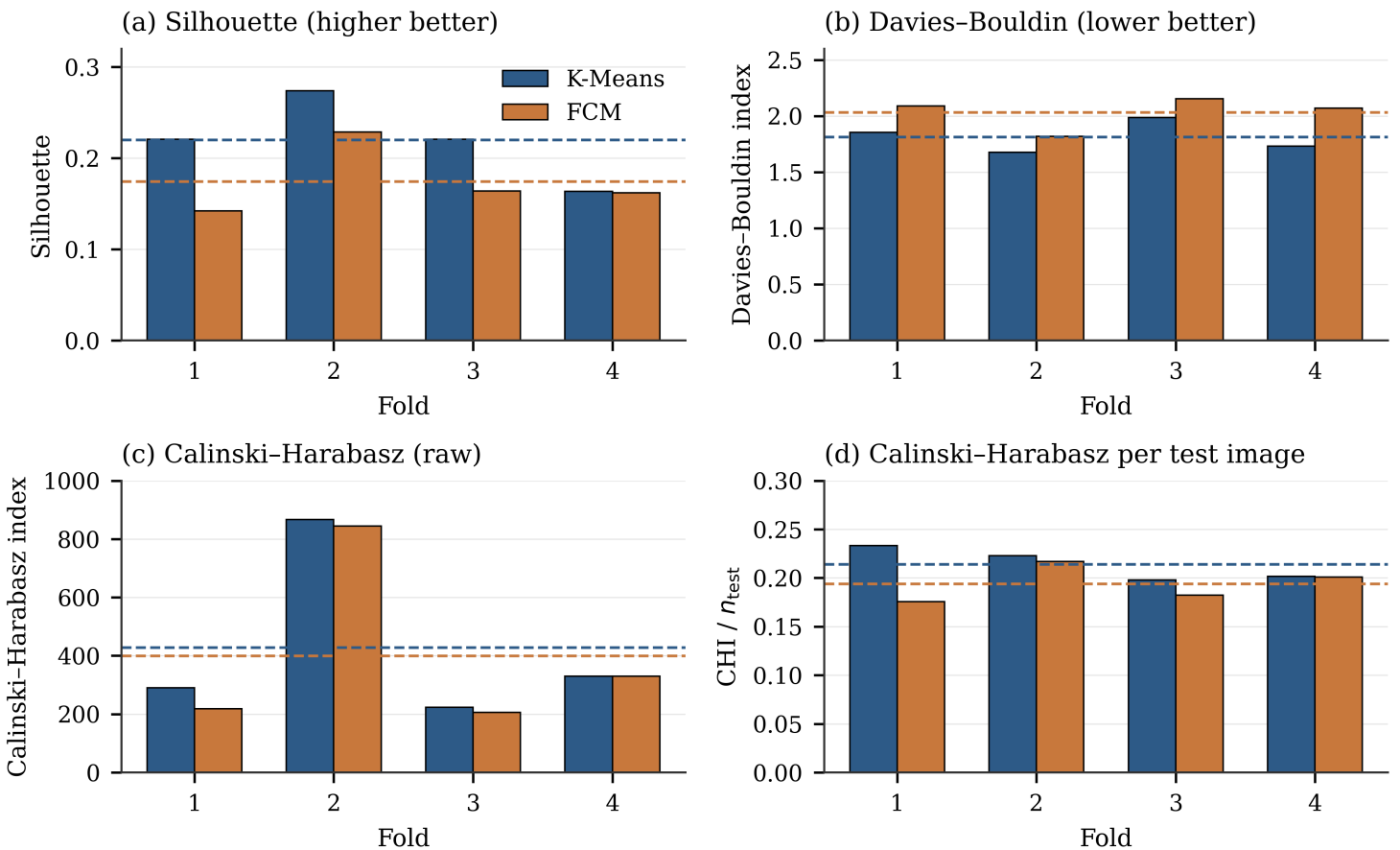


Figure 6

Internal validity indices per fold; dashed lines mark cross-fold means. (a) Silhouette coefficient, higher is better. (b) Davies–Bouldin index, lower is better. K-Means is superior on both in every fold. (c) Raw Calinski–Harabasz index, which varies almost fourfold across folds. (d) The same index divided by test-partition size, which is nearly constant. Panels (c) and (d) together show that the cross-fold variation in the raw index reflects partition size rather than cluster structure.