Low-Dimensional Structure in Breast Cancer Cytology Revealed by Unsupervised Learning

Gillian Macdonald, Lenka Okasova, and Bob Rice
School of Physics and Astronomy, University of Nottingham, UK

Unsupervised learning methods provide a principled framework for uncovering latent structure in biomedical data without reliance on diagnostic labels. We analyse the Breast Cancer Wisconsin (Original) dataset [1] using principal component analysis (PCA) followed by k-means clustering to assess whether benign and malignant tumours exhibit intrinsic separation based on cytological feature statistics alone. Correlation analysis reveals strong dependencies among size- and shape-related features, motivating dimensionality reduction. PCA shows that the first two components capture approximately 74% of the total variance (80% with three components), reflecting a low-dimensional structure dominated by shared morphological variation. Clustering in this reduced representation yields strong alignment between unsupervised clusters and diagnostic labels, achieving an adjusted Rand index of 0.84. Agreement with labels was assessed, post hoc. Although substantial overlap remains, these results demonstrate that correlated cytological features encode diagnostically relevant information, highlighting the utility of simple unsupervised approaches for exploratory structure discovery in biomedical data.

INTRODUCTION

Breast cancer diagnosis relies on quantitative analysis of cytological morphology, where systematic differences in cell size, shape, and nuclear properties are indicative of malignancy [2]. Advances in digital cytology enable multivariate tumour feature extraction, motivating data-driven exploration of cellular structure beyond individual measurements [3].

Despite high diagnostic accuracy of supervised methods, unsupervised analysis remains important for understanding intrinsic tumour structure and interpretability. Accordingly, such approaches provide complementary insight into whether clinically relevant structure emerges directly from feature-space regularities. Using principal component analysis (PCA) one can identify dominant directions of shared variance among correlated features [4]. Moreover, clustering of reduced representations reveals intrinsic groupings and enables unsupervised comparison with diagnostic categories [5].

Here, we apply unsupervised learning methods to the Breast Cancer Wisconsin (Original) dataset [1] to test whether benign and malignant tumours occupy distinct regions in cytological feature space. We show that correlated cytological features induce a low-dimensional structure in which the two diagnostic classes are partially separable, even without label information. Importantly, dominant variance axes (PC1 and PC2) capture clinically meaningful morphological distinctions, demonstrating the value of simple unsupervised methods for biomedical structure discovery.

METHODS

Dataset and preprocessing: We analysed the Breast Cancer Wisconsin (Original) dataset [1] from the UCI Machine Learning Repository, consisting of 699 cytology samples annotated as benign or malignant. Each sample is described by nine ordinal-valued features capturing graded assessments of cell size, shape, adhesion, nuclear properties, and mitotic activity derived from fine needle aspirates. Samples containing missing values were removed prior to analysis, and diagnostic labels were retained solely for post-hoc evaluation and visualisation, not for training. All features were standardised to zero mean and unit variance using the sample standard deviation (N-1 normalisation) prior to analysis.

Correlation analysis: To characterise dependencies among cytological features, we computed the Pearson correlation matrix

$$r_{jk} = \frac{\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{N}(x_{ik} - \bar{x}_k)^2}}. \quad (1)$$

where x_{ij} denotes the value of feature j for sample i and $\bar{x}_j$ is the corresponding feature mean. Strong correlations among multiple size- and shape-related features indicate substantial redundancy in the original feature space, motivating dimensionality reduction to capture dominant modes of shared morphological variation.

Principal component analysis: Dimensionality reduction was performed using principal component analysis (PCA), fit to the standardised feature matrix without access to diagnostic labels. PCA was chosen due to its interpretability and its ability to identify dominant correlated directions in linear feature spaces. Following standardisation to zero mean and unit variance, the sample covariance matrix

$$\Sigma = \frac{1}{N-1} X^{\mathsf{T}} X, \quad (2)$$

was diagonalised by solving the eigenvalue problem

$$\Sigma \mathbf{v}_\ell = \lambda_\ell \mathbf{v}_\ell, \quad (3)$$

where λ_ℓ and $\mathbf{v}_\ell$ denote the eigenvalues and corresponding eigenvectors, ordered such that $\lambda_1 \geq \lambda_2 \geq \ldots$. Each standardised data point $\mathbf{x}_i$ was projected onto the first K principal components via

$$\mathbf{z}_i = V_K^{\mathsf{T}} \mathbf{x}_i, \tag{4}$$

where $V_K = [\mathbf{v}_1, \ldots, \mathbf{v}_K]$. The number of retained components was selected based on the cumulative explained variance and inspection of the eigenvalue spectrum.

Unsupervised clustering: Clustering was performed in the reduced PCA space using the k-means algorithm. Given cluster centroids $\boldsymbol{\mu}_k$, assignments were obtained by minimising the within-cluster sum of squared distances,

$$\mathcal{L} = \sum_{i=1}^{N} \min_k \|\mathbf{z}_i - \boldsymbol{\mu}_k\|^2 . \tag{5}$$

The number of clusters was fixed to $k = 2$, corresponding to the expected benign and malignant tumour classes, although no diagnostic label information was used during clustering.

Evaluation: Clustering performance was assessed post hoc using the adjusted Rand index (ARI),

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathbb{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathbb{E}[\mathrm{RI}]}, \tag{6}$$

which quantifies agreement between inferred cluster assignments and reference diagnostic labels while correcting for chance. An ARI value close to unity indicates strong agreement.

RESULTS

Strong correlations are observed among multiple cytological features, particularly those related to cell size and shape (Fig. 1). Pairwise Pearson correlation coefficients among size- and shape-related features frequently exceed $r \approx 0.7$, indicating substantial redundancy in the original ordinal feature space. In contrast, features associated with mitotic activity and nuclear characteristics exhibit weaker correlations with other variables ($r \lesssim 0.5$), suggesting partially independent sources of variability. These correlation patterns indicate that a large fraction of the variance in the dataset is shared across a subset of morphologically related features. This strongly correlated feature block underlies the dominance of first principal component, which alone accounts for 65% of the total variance.

Principal component analysis reveals a rapidly decaying eigenvalue spectrum (Fig. 2). The first principal component accounts for approximately 65% of the total variance, with the second contributing an additional 9%, such that the first two components together explain 74% of the variance. Including a third component increases the cumulative explained variance to approximately 80%, with diminishing gains from higher-order components. Inspection of the PCA loadings (Fig. 3) shows that the first component is dominated by size- and shape-related cytological features, including clump thickness and uniformity measures, while the second component receives stronger contributions from nuclear-related features such as bare nuclei and chromatin properties. This indicates that the dominant variance reflects shared morphological variation, with secondary structure associated with nuclear characteristics.

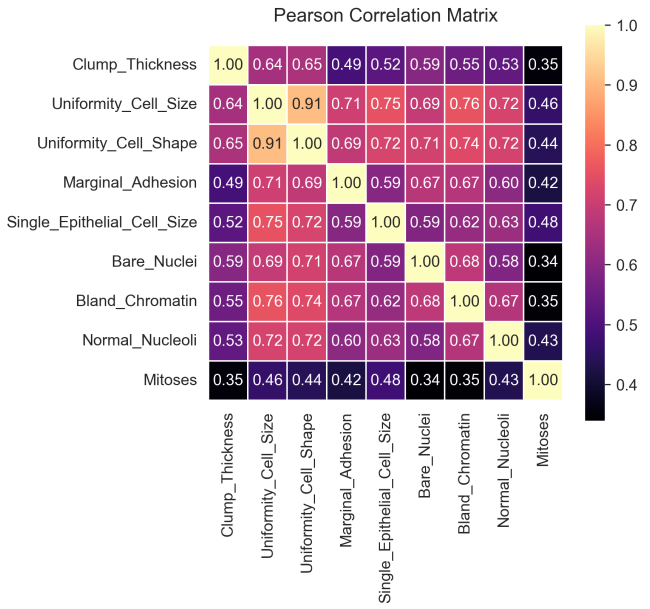

FIG. 1. Pearson correlation matrix of the nine cytological features in the Breast Cancer Wisconsin (Original) dataset. Strong correlations among size- and shape-related features indicate a substantial redundancy in the ordinal feature space and motivate dimensionality reduction.

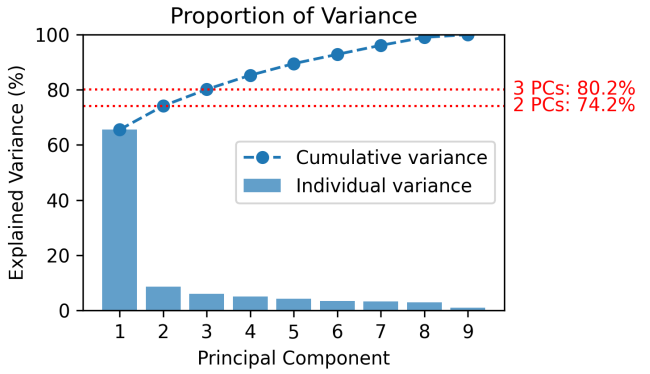

FIG. 2. Individual (bars) and cumulative (points) variance explained by principal components. The first two components capture approximately 74% of the total variance, increasing to 80% when a third component is included, consistent with a low-dimensional structure.

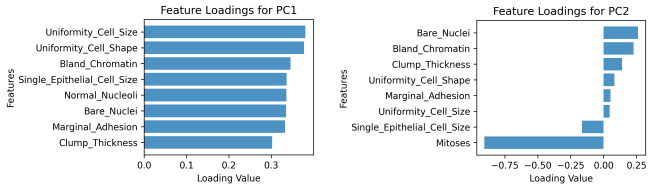

FIG. 3. Feature loadings for the first (top) and second (bottom) principal components. PC1 is dominated by size- and shape-related cytological features, while PC2 receives stronger contributions from nuclear-related features, indicating partially independent sources of variability.

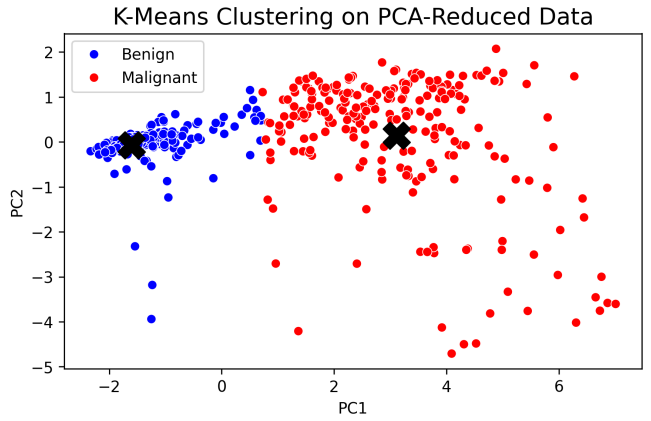

FIG. 4. K-means clustering of tumour samples in the space of the first two principal components. Points are coloured by inferred cluster assignment, and crosses denote cluster centroids. Diagnostic labels are shown for visualisation only.

Projection of tumour samples onto the first two principal components reveals partial separation between benign and malignant classes (Fig. 4). Although substantial overlap remains, the two diagnostic groups occupy systematically different regions of the reduced feature space, with separation occurring primarily along the first principal component. This demonstrates that unsupervised dimensionality reduction alone recovers diagnostically relevant structure from correlated cytological features, despite the absence of class labels during training.

Clustering performed in the space of the first two principal components yields two dominant groupings separated predominantly along PC1 (Fig. 4). The inferred cluster centroids are displaced along the axis associated with size- and shape-related cytological features, consistent with the interpretation of PC1 as a morphological variation axis. Despite the absence of supervision, the resulting cluster assignments show strong quantitative agreement with diagnostic labels, yielding an adjusted Rand index of ARI = 0.84.

Clustering performance remained high across PCA representations of varying dimensionality, with adjusted Rand indices ranging from 0.83 to 0.86 (Table I).

TABLE I. Adjusted Rand Index (ARI) between k-means cluster assignments and diagnostic labels as a function of the number of retained principal components. For each configuration, the reported ARI corresponds to the solution with minimum inertia over 30 random initialisations using k-means++ initialisation and a fixed random seed.

Principal components	2	3	5	9
ARI	0.847	0.847	0.847	0.836

Strong agreement is already achieved using only the first two principal components, with only modest and non-monotonic variation upon inclusion of additional dimensions. This suggests that the dominant morphological structure relevant to tumour classification is captured by a small number of correlated feature combinations.

CONCLUSION

We have shown that unsupervised analysis of breast cancer cytological data reveals a low-dimensional structure that aligns closely with pathological distinctions between benign and malignant tumours. Principal component analysis identifies dominant variance directions driven by correlated size- and shape-related features, capturing the majority of variability in the dataset. Clustering in this reduced representation recovers two principal groupings that show strong agreement with diagnostic labels, despite the absence of supervision. The saturation of clustering performance at low dimensionality is consistent with the rapid decay of the PCA eigenvalue spectrum, indicating that diagnostically relevant structure is governed primarily by shared morphological variation. Although substantial overlap between classes remains, reflecting biological variability and measurement noise, the observed partial separation demonstrates that key pathological information is encoded directly in the statistical organisation of cytological features. These results highlight the utility of simple unsupervised pipelines for exploratory structure discovery in biomedical datasets when labels are limited or uncertain.

[1] W. H. Wolberg, Breast Cancer Wisconsin (Original) dataset; UCI Machine Learning Repository (1990), https://doi.org/10.24432/C5HP4Z.
[2] W. H. Wolberg, O. L. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology, Proc. Natl. Acad. Sci. U.S.A. 87, 9193 (1990).
[3] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, Nuclear feature extraction for breast tumor diagnosis, Proc. IS&T/SPIE Symp. Electronic Imaging: Science and Technology **1905**, 861 (1993).

[4] I. T. Jolliffe, *Principal Component Analysis*, (Springer, New York, 2002).
[5] J. MacQueen, Some methods for classification and analysis of multivariate observations, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (University of California Press, Berkeley, 1967), pp. 281–297.