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 provides 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 solely on cytological feature statistics. Strong correlations among size- and shape-related features motivate dimensionality reduction, with the first two principal components capturing approximately 74% of the total variance (80% with three components), indicating a low-dimensional structure dominated by shared morphological variation. Clustering in this reduced space shows strong agreement with diagnostic labels (adjusted Rand index 0.84), assessed *post hoc*. Despite substantial overlap between classes, diagnostically relevant structure is encoded in correlated cytological features.

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 principal component one (PC1) and principal component two (PC2) capture clinically meaningful morphological distinctions, demonstrating the value of simple unsupervised methods for biomedical structure discovery.

We hypothesise that intrinsic correlations among cytological features induce a low-dimensional structure that partially separates benign and malignant tumours without supervision.

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, leaving N = 683 samples, and diagnostic labels were retained solely for *post hoc* evaluation and visualisation. All features were standardised to zero mean and unit variance using the sample standard deviation (N-1 normalisation) prior to analysis. Although features are ordinal, treating them as approximately interval-scaled is standard for this dataset.

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

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

where x_{ij} denotes the value of feature j for sample i and $\bar{x}_j$ is the corresponding feature mean. Pearson correlation coefficients were used to characterise linear dependencies among approximately interval-scaled features. Strong inter-feature correlations were used to assess feature redundancy and motivate subsequent application of PCA.

Principal component analysis: Dimensionality reduction was performed using 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

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

was diagonalised by solving the eigenvalue problem

$$\Sigma \mathbf{v}_\ell = \lambda_\ell \mathbf{v}_\ell, \tag{3}$$

where λ_ℓ and $\mathbf{v}_\ell$ denote the eigenvalues and corresponding eigenvectors, ordered such that $\lambda_1 \geq \lambda_2 \geq \dots$. 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 $V_K = [\mathbf{v}_1, \dots, \mathbf{v}_K]$. 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. The number of clusters was fixed to $k = 2$, corresponding to the two diagnostic categories in the dataset, while diagnostic labels were used only for *post hoc* evaluation. Because k-means is sensitive to initial centroid placement and can converge to

local minima, k-means++ initialisation was used with 30 random restarts (fixed seed), and the minimum-inertia solution was retained.

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

Given cluster centroids $\boldsymbol{\mu}_k$, sample assignments were obtained by minimising the sum of squared distances to the nearest centroid.

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 these features frequently exceed $r \approx 0.7$, indicating substantial redundancy in the 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 patterns indicate that a large fraction of the dataset variance is shared across a subset of morphologically related features, motivating dimensionality reduction and underlying the dominance of the first principal component.

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 contributions from higher-order components. Inspection of the PCA loadings (Fig. 3) shows that PC1 is dominated by size- and shape-related features, while PC2 receives stronger contributions from nuclear-related characteristics, indicating partially independent axes of morphological variation.

Projection of tumour samples onto the first two principal components reveals partial separation between benign and malignant classes (Fig. 4), with separation occurring primarily along PC1. Clustering in this reduced space yields two dominant groupings whose centroids are displaced along the axis associated with size- and shape-related features, consistent with the interpretation of PC1 as a global 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 in *post hoc* comparison.

Although clustering achieves strong agreement, substantial overlap remains, particularly for samples near the decision boundary. Misclassified tumours predominantly exhibit intermediate morphological characteristics, suggesting intrinsic biological ambiguity and measurement noise rather than algorithmic failure. Clustering performance remains high across

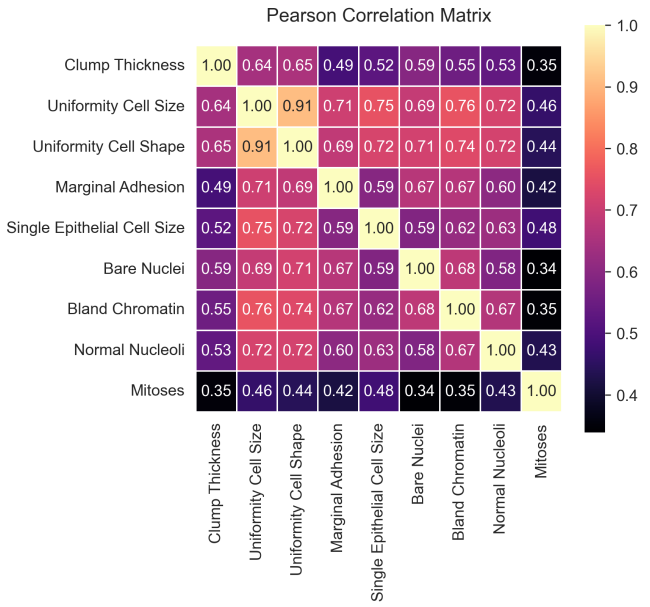

FIG. 1. Pearson correlation matrix of the nine cytological features in the Breast Cancer Wisconsin (Original) dataset, showing strong size- and shape-related dependencies that motivate dimensionality reduction.

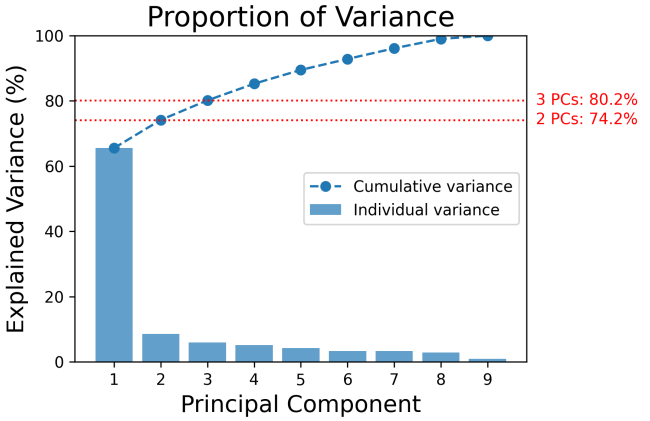

FIG. 2. Individual (bars) and cumulative (points) variance explained by principal components, showing that the first two PCs capture 74% of the total variance, rising to 80% with the inclusion of a third, consistent with a low-dimensional structure.

PCA representations of varying dimensionality, with adjusted Rand indices ranging from 0.83 to 0.86 (Table I). Strong agreement is already achieved using only the first two principal components, with only modest and non-monotonic variation upon inclusion of additional dimensions, indicating that diagnostically relevant structure is captured by a small number of correlated feature combinations.

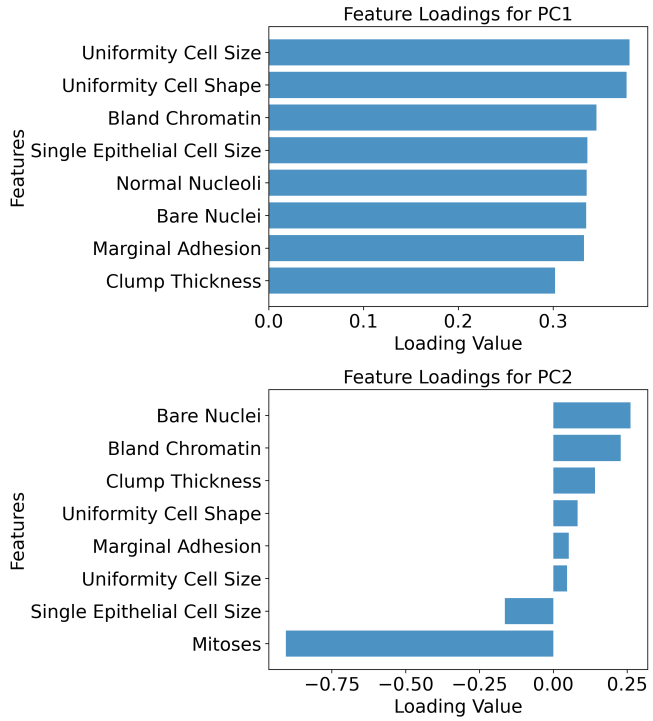

FIG. 3. Feature loadings for the first (top) and second (bottom) principal components. While PC1 is dominated by size- and shape-related cytological features, PC2 shows stronger contributions from nuclear-related features, reflecting distinct sources of biological variability. The sign of each loading value indicates the direction of feature contribution along the principal component axis, while its magnitude reflects relative importance.

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. ARI is reported post hoc for evaulation only and was not used to select the number of retained PCs

Principal components	2	3	5	9
ARI	0.847	0.847	0.847	0.836

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. Future work could explore alternative unsupervised models, such as hierarchical or mixture-based clustering, to further characterise structure in this feature space.

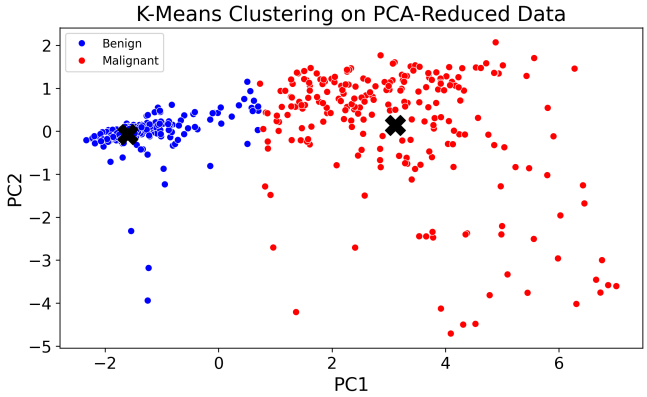

FIG. 4. K-means clustering of tumour samples projected onto the first two PCs. Points are coloured by computed cluster assignment, with crosses indicating cluster centroids. Diagnostic labels are shown for visualisation only and were not used during clustering.

[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.