Unsupervised Learning Reveals Structure in Breast Cancer Tumour Data

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

Unsupervised learning methods offer a systematic approach to uncover latent structure in high-dimensional biomedical data without reliance on diagnostic labels. We analyse the Breast Cancer Wisconsin dataset using principal component analysis (PCA) followed by k-means clustering to probe whether malignant and benign tumour classes emerge naturally from feature statistics alone. Correlation analysis reveals a strongly coupled cluster of morphological and nuclear features associated with malignancy, motivating dimensionality reduction. PCA shows that the first two components capture 74.2% of the total variance (80.2% with three components), with clear separation between tumour classes in the reduced space. Applying k-means clustering in this low-dimensional representation yields robust unsupervised class recovery, achieving and adjusted Rand index of 0.84 relative to known tumour classes. These results demonstrate that correlated cytological abnormalities organise the data into a low-dimensional structure that cleanly separates malignant and benign tumours, highlighting the utility of simple unsupervised approaches for structure recovery in 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 [1]. Advances in digital cytology have enabled the extraction of high-dimensional feature representations from tumour samples [2], motivating data-driven approaches to explore structure within tumour populations beyond individual features.

While supervised learning methods have achieved high diagnostic accuracy, understanding the intrinsic organisation of tumour data without reference to class labels remains important for exploratory analysis and interpretability. Unsupervised methods offer a complementary perspective by probing whether clinically relevant structure emerges directly from statistical regularities in the feature space. Principal component analysis (PCA) provides a natural tool for this purpose, identifying dominant directions of shared variance that often correspond to biologically meaningful combinations of correlated features [3]. Clustering methods applied in reduced representations can further expose intrinsic groupings, enabling comparison with known diagnostic categories without explicit supervision [4, 5].

Here, we apply unsupervised dimensionality reduction and clustering to the Breast Cancer Wisconsin dataset to investigate whether benign and malignant tumours occupy distinct regions of feature space. We show that correlated cytological features give rise to a low-dimensional structure in which the two diagnostic classes are partially separable, despite the absence of labels during training. These results demonstrate that dominant axes of variance in cytological feature space capture clinically meaningful morphological distinctions, highlighting the utility of simple unsupervised approaches for structure discovery in biomedical data.

METHODS

Dataset and preprocessing: We analysed the Breast Cancer Wisconsin cytology dataset, comprising 569 samples characterised by D = 30 morphological and nuclear features extracted from breast tissue samples. Features were standardised to zero mean and unit variance prior to analysis,

Correlation analysis: To characterise feature dependencies, we computed the Pearson correlation matrix

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

Revealing strongly correlated groups of cytological features associated with malignancy. This motivated dimensionality reduction to capture dominant modes of variation.

Principal component analysis: Dimensionality reduction was performed using principal component analysis (PCA). The 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 $\mathbf{v}_\ell$ denotes the eigenvalues and corresponding eigenvectors, ordered such that $\lambda_1 \geq \lambda_2 \geq \ldots$. Each data point $\mathbf{x}_i$ was projected onto the first K principal components via

$$\mathbf{z}_i = V_K^{\mathsf{T}} \mathbf{x}_i, \quad (4)$$

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

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, though no diagnostic label information was used during clustering.

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

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

which measures 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 feature space. In contrast, features associated with mitotic activity and nuclear texture exhibit weaker correlations with other variables ($r \lesssim 0.5$), suggesting partially independent sources of variability. These correlation patterns indicate that much of the variance in the data is shared across a subset of morphological features.

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%. Inspection of the PCA loadings shows that the first component is dominated by size- and shape-related features, including clump thickness and uniformity measures, while the second component receives stronger contributions from nuclear texture features such as bare nuclei and chromatin properties. This indicates that the dominant variance reflects a shared morphological abnormality axis, with secondary variation associated with nuclear structure.

Projection of tumour samples onto the first two principal components reveals partial separation between benign and malignant classes (Fig.). Although substantial

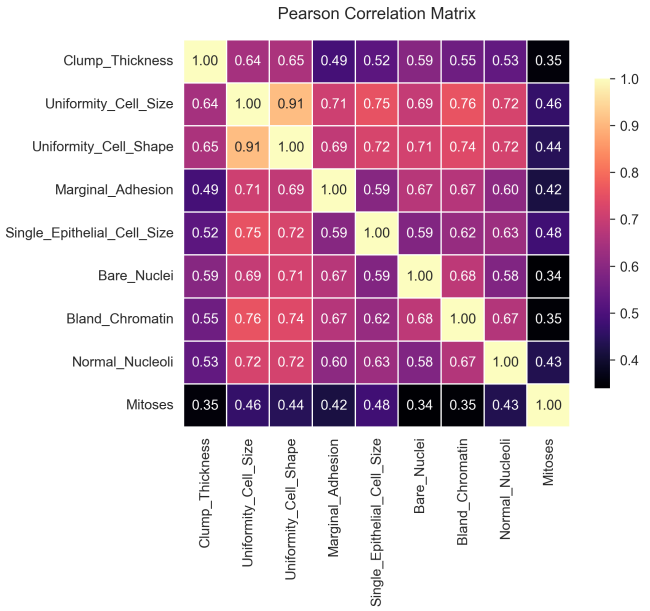

FIG. 1. Pearson correlation matrix of cytological features in the Breast Cancer Wisconsin dataset. Strong correlations among size- and shape-related features indicate a shared morphological abnormality axis.

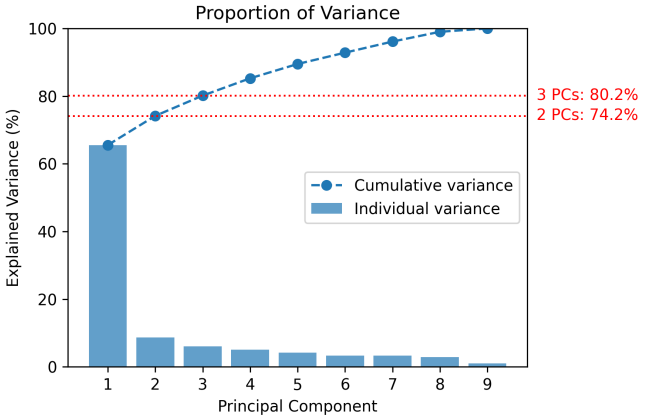

FIG. 2. Individual and cumulative variance explained as a function of the number of retained components. The first two components capture approximately 74% of the total variance, increasing to 80% when three components are included.

overlap remains, the two classes occupy distinct 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, 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. 3). The inferred cluster centroids are displaced along the axis associated with size- and shape-related cytological features, consis-

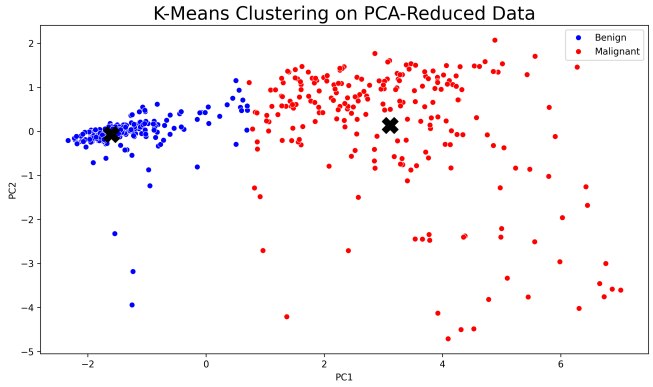

FIG. 3. K-means clustering of tumour samples in the space of the first two principal components. Points are coloured according to inferred cluster assignment, and crosses denote cluster centroids. Diagnostic labels are shown for visualisation only.

tent with the interpretation of PC1 as a morphological abnormality 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.

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 axes of variance driven primarily by correlated size- and shape-related features, capturing the majority of variability in the dataset. Clustering performed in this reduced representation recovers two principal groupings that show strong agreement with diagnostic labels, despite the absence of supervision during training. Although substantial overlap between classes remains, reflecting biological variability and measurement noise, the emergence of partial separation indicates that key pathological information is encoded directly in the statistical structure of the feature space. These results demonstrate that simple unsupervised pipelines can recover clinically meaningful organisation from high-dimensional biomedical data and provide a useful framework for exploratory analysis when labels are limited or uncertain. Future work incorporating richer feature representations or spatial and temporal information may further enhance structure discovery and interpretability in tumour datasets.

TABLE I. Adjusted Rand Index (ARI) between k-means clustering results and diagnostic labels as a function of the number of retained principal components.

Principal components	2	3	5	10
ARI	0.830	0.847	0.825	0.858

[1] 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).

[2] 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).

[3] I. T. Jolliffe, *Principal Component Analysis*, (Springer, New York, 2002).

[4] 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.

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, (Springer, New York, 2009).