IJAMML 19:1 (2025) 37-52 December 2025
ISSN: 2394-2258
Available at http://scientificadvances.co.in
DOI: http://dx.doi.org/10.18642/ijamml_7100122332

RESEARCH ON BREAST CANCER CLASSIFICATION BASED ON PCA AND K-NEAREST NEIGHBOR

LUJIN LYU, XINYU LIU, LEILEI BAO, QIRUI XIAO, LI LI and JIANQIANG GAO

School of Medical Information Engineering, Jining Medical University, Rizhao 276826, Shandong, P. R. China

---

Abstract

Breast cancer has become the most common malignant tumor, making early and precise diagnosis crucial for improving patient survival rates. However, although high-dimensional clinical data can provide rich information, it will introduce the 'curse of dimensionality'. dimensional disaster". This will increase computational costs and reduce generalization ability of traditional models. To solve this problem, this paper proposes a collaborative framework based on principal component analysis (PCA) and K nearest neighbor classification, denoted as PCA-KNN. The approach employs PCA for dimensionality reduction, compressing the feature space while preserving key variation information. In the low-dimensional subspace, K nearest neighbor algorithm is used to realize the classification task. The proposed method adopts cross-validation strategy, and according to the

*Corresponding author.
E-mail address: jianqianggaohh@126.com (Jianqiang Gao).

2020 Mathematics Subject Classification: 68T10.
Submitted by Jianqiang Gao.
Received December 26, 2025

contribution rate of principal components, the K value in the K nearest neighbor algorithm is obtained. Experimental results show that this proposed method significantly improves computational efficiency while maintaining high classification accuracy, achieving an effective balance between information retention and computational performance. The proposed method has the characteristics of lightweight, strong interpretability and easy operation. This not only provides a practical auxiliary tool for breast cancer recurrence risk prediction, but also establishes a universal paradigm for high-dimensional medical data analysis tasks such as imaging and multi-group fusion. This contributes to advancing the implementation of precision medicine in clinical practice.

Keywords: breast cancer, PCA, KNN, dimensionality reduction, classification.

---

1. Introduction

Breast cancer ranks among the most prevalent cancers affecting women globally, accounting for approximately 24.5% of all female cancers [1]. Early diagnosis is crucial for improving patient survival rates. With the deepening application of machine learning techniques in medical imaging and diagnostic support systems, constructing efficient and reliable classification models under limited clinical conditions has become a key research focus. While providing rich information, high-dimensional clinical feature data also introduces issues such as information redundancy and model overfitting, directly impacting the generalisation performance of classifiers.

PCA is a classical unsupervised dimensionality reduction method, which transforms raw features into a set of linearly independent principal components through orthogonal transformation. This achieves feature space compression while preserving the majority of data variability[2]. KNN classifier is widely used in pattern recognition because of its intuitive and non-parametric characteristics. However, its classification performance decreases significantly with the increase of data dimension, which is called ‘curse of dimensionality’ [3].

To overcome the 'curse of dimensionality' and strike a better balance between computational efficiency and classification performance, this study constructs a multi-stage PCA-KNN fusion framework method. Based on different levels of cumulative variance contribution rate, the proposed method gradually compresses the original high-dimensional clinical feature data into several low-dimensional principal component molecular spaces. In each compressed subspace, the best nearest neighbor parameter k is adaptively determined by cross-validation strategy. With the help of 'contribution rate-K' two-variable grid search mechanism, the proposed method realizes the dynamic regulation between information retention and model complexity. It not only effectively relieves the computational and storage burden of KNN algorithm in high-dimensional scenes, but also achieves or surpasses the performance of full feature modeling in many performance indexes, which provides a flexible, efficient and interpretable new way for the auxiliary diagnosis of early recurrence risk of breast cancer.

2. KNN Method

Researchers can directly use the classical KNN algorithm to classify breast cancer data, but the classification results are often unsatisfactory. The core idea of KNN algorithm is to give a sample to be predicted, find the nearest K training samples in the feature space, and decide its output by voting or weighted average. KNN algorithm does not need explicit training process, and only relies on local neighborhood information, so it has no assumptions about data distribution. However, its computational complexity increases linearly with sample size. The specific steps include:

Step 1: Select a distance metric (for example, Euclidean distance, Manhattan distance.).

Step 2: Determine the number of neighbors k and the weighting strategy (distance-weighted).

Step 3: Retrieve K nearest neighbors for the query sample and make a decision.

KNN algorithm is widely used in recommendation system, anomaly detection, medical diagnosis and multi-spectral image classification [4].

3. Proposed Method

This study proposed a two-stage breast cancer recurrence risk prediction framework integrating PCA with KNN classifier. This aims to achieve effective dimensionality reduction and efficient classification of high-dimensional clinical features.

The core idea of PCA is to project high-dimensional data into a new orthogonal coordinate system which is expanded by the eigenvector corresponding to the maximum eigenvalue of covariance matrix, so as to maximize the variance of the projected data, thus retaining as much original information as possible, while reducing the dimension of the data [5].

The operation of the proposed algorithm includes the following aspects: (1) centralization of data, (2) calculation of sample covariance matrix, (3) calculation of eigenvalue of covariance matrix. Firstly, the eigenvectors corresponding to the top k largest eigenvalues are selected to form a transformation matrix, and the original data are projected onto the transformation matrix to obtain a low-dimensional representation. The representations of covariance matrix, eigenvalue decomposition, projection matrix and dimension reduction data are as follows:

Covariance matrix: $C = \frac{1}{n-1} X^T X$.

Eigenvalue decomposition: $C = V \Lambda V^T$.

Projection matrix: $P = V_{:,\,1,\,:K}$.

Dimension-reduced data: $Y = XP$.

PCA has been widely applied across diverse scenarios including image compression, facial recognition, financial risk control, and agronomic trait analysis [6]. This study employs the UCI Wisconsin Breast Cancer Prognosis Data set, comprising 198 samples (151 non-recurrent and 47 recurrent cases) with 33 clinical features, to construct a binary prognostic model for recurrence risk assessment. With comprehensive cytometric quantification metrics and explicit classification labels, this data set provides a robust foundation for binary classification. The experimental steps are as follows:

3.1. Data preprocessing

(1) Missing value imputation [7]

The original feature matrix is denoted as $X = [X_{ij}] \in \mathbb{R}^{n\times d}$, and the label vector is $y = [y_i] \in \{R, N\}^n$, where n denotes the number of samples, d represents the feature dimension, R indicates recurrence, and N denotes non-recurrence.

Define the set of valid sample indices: $\mathcal{J}_{valid} = \{i | \forall j \in \{1, \ldots, d\}, x_{ij} \neq NaN\}$.

Retain valid samples, update feature matrix and label vector: $X \leftarrow X(\mathcal{J}_{valid}, :)$, $y \leftarrow y(\mathcal{J}_{valid})$.

The number of processed samples is defined as $m = |\mathcal{J}_{valid}|$.

(2) Feature standardization [8]

For each dimensional feature $j = 1, \ldots, d$, compute its mean and standard deviation: $\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij}$, $\sigma_j = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_{ij} - \mu_j)^2}$.

Standardized eigenvalue: $\tilde{x}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$, $i = 1, \ldots, m$; $j = 1, \ldots, d$.

The standardized characteristic matrix is denoted as $\widetilde{X} = [\widetilde{x}_{ij}] \in \mathbb{R}^{m\times d}$.

(3) Digitization of category labels [2, 7, 8]

Map textual labels to binary numerical labels: $y_i = \begin{cases} 1, \text{ if the original label is } R, \\ 0, \text{ if the original label is } N, \end{cases} i = 1, \ldots, m.$

The processed label vector is denoted as $y_{bin} \in \{0, 1\}^m$.

3.2. Dimension reduction

In order to solve the problems of computational redundancy and generalization ability decline caused by the 'curse of dimensionality' in high-dimensional data, PCA is employed for feature dimension reduction. This method uses orthogonal linear transformations to map original features onto an orthogonal basis space ordered by decreasing variance and then key principal components are then selected based on cumulative variance contribution to establish a low-dimensional projection subspace, achieving an optimal balance between data compression and preservation of essential structure [5]. Generally speaking, feature dimension reduction of PCA includes three main steps:

(1) Feature decomposition

For the standardized data matrix $X \in \mathbb{R}^{m\times d}$, compute the covariance matrix $\sigma = \frac{1}{m-1} X^T X$ and perform eigenvalue decomposition: $\sigma = V\Lambda V^T$, where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$.

(2) Dimension selection

Calculate the cumulative contribution rate of the first κ principal

components: $n(\kappa) = \frac{\sum_{j=1}^{\kappa} \lambda_j}{\sum_{i=1}^{d} \lambda_i}$.

Determine the minimum dimension κ^* that satisfies $\eta(\kappa) \geq 0.95$.

(3) Data projection

Construct the projection matrix: $P = [v_1, v_1, \ldots, v_{\kappa^*}] \in \mathbb{R}^{d \times \kappa^*}$ and get reduced-dimension features $Z = XP \in \mathbb{R}^{m \times \kappa^*}$. This method significantly enhances the efficiency of subsequent processing while retaining 95% of the data variance.

3.3. Classification

To extract the inherent value of reduced-dimension features while preserving non-parametric advantages, this study employs the KNN algorithm to construct a classifier. The specific workflow is as follows [10]:

(1) Selection of K value

Based on the candidate set $\mathcal{K} = \{3, 5, 7\}$, the optimal number of neighbors K^* is selected using 10-fold cross-validation, with the objective of maximizing the average accuracy rate.

For each fold f, the validation set accuracy is defined as: $Acc_f(K) = \frac{1}{|V_f|} \sum_{i \in V_f} \mathrm{II}[y_i = \hat{y}_i(K)]$, where V_f denotes the sample index set of the validation set for the $f-$th fold, and $\mathrm{II}[.]$ represents the indicator function. The final selection is $K^* = org \max_{K \in \mathcal{K}} \frac{1}{10} \sum_{f=1}^{10} Acc_f(K)$.

(2) Distance metric

For test sample $z \in \mathbb{R}^{\kappa^*}$, the distance to training sample z_i is

calculated by using Euclidean distance:

$$d(z, z_i) = \| z - z_i \|_2 = \sqrt{\sum_{j=1}^{\kappa^*} (z_j - z_{ij})^2}.$$

(3) Classification decision

Let $\mathcal{N}_{\kappa^*}(z) = \{i_1, \ldots, i_{\kappa^*}\}$ denote the set of indices for the K^* training samples with the smallest distances. The predicted label is then determined by majority voting $\hat{y} = \text{mode}[y_{i_1}, y_{i_2}, \ldots, y_{i_{\kappa^*}}]$, where $\text{mode}(\cdot)$ returns the category with the highest occurrence count (no recurrence=1, recurrence=2). This process operates within the low-dimensional subspace of $k^* \leq 7$, effectively balancing algorithmic intuitiveness with computational and storage efficiency [11]. Comparative experiments were configured as follows:

(i) Baseline group: Direct modelling based on the original 33-dimensional features;

(ii) Experimental group: PCA is used to reduce the dimension, and then KNN is used to classify.

3.4. Evaluation

To quantitatively assess model performance across multiple dimensions, this study employs an evaluation framework comprising Accuracy, Precision, Recall, F1-score, and Time. Based on the elements of the confusion matrix—TP (True Positives), TN (True Negatives), FP (False Positives), and FN (False Negatives)—the metrics are defined as follows:

Accuracy: Represents the proportion of correctly classified samples relative to the total sample size, reflecting overall classification correctness [12].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}.$$

Precision: Represents the proportion of samples predicted as recurrent that are actually recurrent, measuring the model's accuracy in predicting positive class [13].

$$Precision = \frac{TP}{TP + FP}.$$

Recall: The proportion of samples that actually recurred and were correctly identified as recurrent, measuring the model's ability to recognise positive class samples [13].

$$Recall = \frac{TP}{TP + FN}.$$

F1-score: The harmonic mean of precision and recall, providing a comprehensive assessment of the model's classification performance, particularly suitable for imbalanced data scenarios [13].

$$F1 = 2 \times \frac{Precison \times Recall}{Precison + Recall}.$$

Time: Represents the average time required for the model to make predictions on the test set, reflecting computational efficiency. Ten-fold cross-validation was employed in experiments to ensure result stability and reliability [11].

4. Results and Analysis

4.1. Experimental results

To quantify the dimensionality reduction effect of PCA, comparisons were made between the original 33-dimensional feature space and models retaining different proportions of principal components. Experiments employed 10-fold cross-validation combined with breast cancer has become the most common malignant tumor, making early and precise diagnosis crucial for improving patient survival rates. While high-

dimensional clinical data provide rich information, they also introduce the 'curse of dimensionality', which increases computational costs and reduces the generalization of traditional models. To address those problem, we proposed a collaborative framework based on PCA and KNN classification, denoted as PCA-KNN. The approach employs PCA for dimensionality reduction, compressing the feature space while preserving key variation information. Subsequently, the KNN algorithm enables efficient classification in the low-dimensional subspace. Cross-validation is used to simultaneously optimize two core parameters: the principal component contribution rate and the number of neighbors, i.e., K. Experimental results show that the proposed method significantly improves computational efficiency while maintaining high classification accuracy, achieving an effective balance between information retention and computational performance. Under three different indicators (i.e., accuracy, F1-score, and time), the UCI Wisconsin breast cancer prognosis data set was tested by using the proposed method framework, and Table 1 and Table 2 are obtained.

Table 1. The optimal dimensions corresponding to different cumulative contribution rates are obtained by PCA method

Original dimensions	Contribution rate/%	Optimal dimension
33	95	12
	85	7
	75	5
	65	3
	55	2

Table 2. Comparison of the results under different k values in KNN algorithm

Whether to employ PCA	Contribution rate /%	K	Accuracy	Precision	Recall	F1-score	Time/s
No		1	0.6821	0.3502	0.3900	0.3565	0.002272
		2	0.7734	0.5333	0.2200	0.2968	0.001115
Yes	95	1	**0.7024**	**0.4602**	**0.4450**	**0.4131**	**0.001119**
		2	**0.7982**	**0.6417**	**0.2300**	**0.3186**	**0.000738**
	85	1	**0.6971**	**0.3775**	**0.4000**	**0.3822**	**0.001811**
		2	**0.7782**	0.4000	0.1500	0.2162	**0.000656**
	75	1	0.6811	**0.4146**	0.3450	0.3305	**0.000839**
		2	0.7526	0.2000	0.0650	0.0952	**0.000936**
	65	1	0.6508	0.2800	0.2800	0.2717	**0.001123**
		2	0.7418	0.4000	0.1300	0.1786	**0.000976**
	55	1	0.6816	0.3000	0.3250	0.3072	**0.000853**
		2	0.7216	0.1583	0.0850	0.1071	**0.000943**

From Table 2, we can see that a good result (i.e., accuracy, precision, recalls, F1-score, and time/s) can be obtained with the proposed strategy by using PCA under the contribution rate is 95%. Hence, using PCA algorithm and KNN for classification can really improve the classification performance of the model and get the best results. For example, under the optimal configuration of 95% variance retention and $K = 2$, the model's accuracy and precision increased to 0.7982 and 0.6417, respectively, while recall and F1-score comprehensively surpassed the original high-dimensional space performance. Due to the deep compression of feature dimensions, computational efficiency achieved a qualitative leap. This framework achieves a delicate equilibrium between

information fidelity and computational efficiency. Figure 1 illustrates the performance evolution of the model across varying variance contribution rates

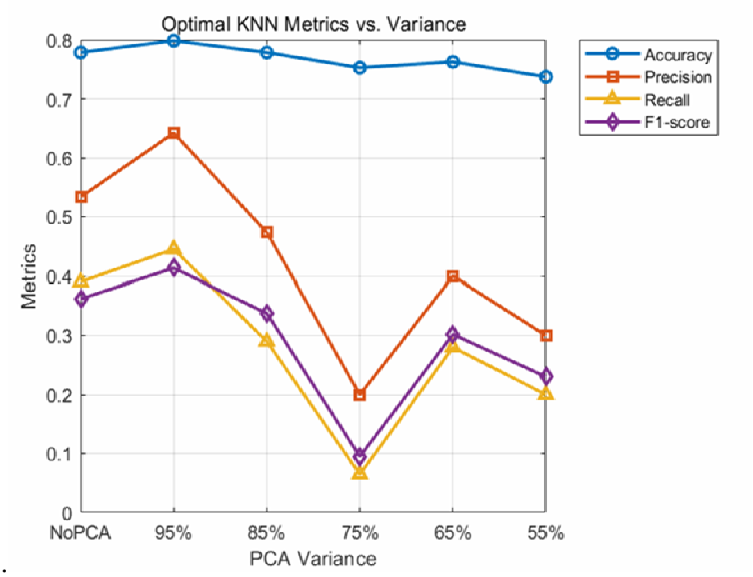

Figure 1. Comparison of optimal KNN model metrics under varying contribution rates.

Based on the data in Table 2, the optimal model performance for each evaluation metric is summarized according to whether PCA is applied and its varying contribution rates. The results are shown in Table 3.

Table 3. Optimal values of evaluation metrics under different PCA contribution rates

Evaluation indicators	Whether to use PCA	Contribution rate /%	K	Optimal values
Accuracy	No		2	0.7734
	Yes	95	2	0.7982
Precision	No		2	0.5333
	Yes	95	2	0.6417
Recall	No		1	0.3900
	Yes	95	1	0.4450
F1-score	No		1	0.3565
	Yes	95	1	0.4131
Time/s	No		2	0.001115
	Yes	95	2	0.000738

4.2. Discussion and analysis

Through the above experimental results, we can clearly see that the proposed framework method can significantly improve the classification effect of KNN classifier. In other words, before data classification, PCA technology is used to reduce the dimension, and then the obtained low-dimensional data are sent to KNN classifier, which will get satisfactory results. For example, in the original feature space (33 features), using KNN model, the highest accuracy is 0.7734. However, the original data from 33 dimensions are reduced to 12 dimensions (with 95% variance) by PCA technology, and the highest accuracy is 0.7982 by using KNN model (K=2), and the accuracy rate is also improved from 0.5333 to 0.6417. What is more noteworthy is that the recall rate and the peak value of F1-score also appear in the feature space after PCA processing. This promotion is not accidental, mainly because the principal component analysis effectively filters out noise and redundant information in the

compression process, weakens the interference of irrelevant dimensions on nearest neighbor search, and thus enhances the discriminant signal of samples in local space.

From the point of view of computational efficiency, the reduction of dimension will bring significant advantages. By analyzing the performance under five variance retention thresholds of 55% to 95%, it can be seen that although the feature dimension of the original data plummeted from 33 to 2 or 12, there was no performance degradation caused by excessive dimension reduction. However, the precision peak is formed in the 95% variance interval, which also shows that the actual redundancy of this data set is higher than the general estimation.

In other words, PCA not only realizes dimension compression, but also improves the classification ability of KNN model. On the whole, through the bivariate strategy of "contribution rate -K value", we successfully transformed the balance between accuracy and efficiency into a programmable and reproducible configuration strategy. If the ultimate accuracy is pursued, 95% variance +K=2 can be adopted. If the response speed is emphasized, 85% variance +K=2 can be selected. This strategy not only provides a reliable reference for the risk prediction of breast cancer recurrence, but also is simple and easy to operate, and also establishes a generalized method paradigm for the lightweight analysis of high-dimensional medical data such as imageology and multimethology.

5. Conclusions

Aiming at the "dimensional disaster" caused by high-dimensional clinical data characteristics in breast cancer recurrence risk prediction, we proposed a lightweight classification framework (PCA-KNN) based on UCI Wisconsin breast cancer prognosis data set. The framework adopts a two-stage strategy of ‘dimensionality reduction first and then classification’ to overcome the shortcomings of KNN in high-dimensional

space. Firstly, PCA is used to compress the original high-dimensional features, and the feature dimensions are reduced to 12, 7, 5, 3 and 2 dimensions under the five variance contribution rates of 95%, 85%, 75%, 65% and 55%, respectively. Then, based on each low-dimensional subspace, the "contribution rate -K value" bivariate strategy is adopted for KNN model to optimize the retention ratio of principal components and the number of neighbors K. Experimental results show that the proposed PCA-KNN framework can significantly improve the classification performance of KNN model.

In this study, PCA is used to filter the irrelevant features of high-dimensional medical data to enhance the classification performance of KNN model. In order to meet the needs of lightweight and low resource consumption in clinical deployment, we adopted a strategy without complex parameter tuning to ensure that it can be used in conventional hardware environment.

To sum up, PCA-KNN strategy is adopted for the real breast cancer recurrence risk prediction task, which not only achieves a controllable trade-off between accuracy and efficiency, but also provides a paradigm reference for the lightweight analysis of high-dimensional medical data such as imageology and genomics, and has important clinical transformation value and broad application development potential.

6. Acknowledgements

This work is partly supported by the Education Project of Industry-University Cooperation of the Ministry of Education under Grant No. 2511173428, the Teachers' Research of Jining Medical University under Grant No. JYFC2019KJ014, and the Doctoral Research Foundation of Jining Medical University under Grant No. 2018JYQD03, P. R. China. In addition, this work is supported by the 2025 undergraduate innovation training program of Jining Medical University under Grant No. cx2025121, P. R. China.

References

[1] J. Kim, A. Harper, V. McCormack, H. Sung, N. Houssami, E. Morgan et al., Global patterns and trends in breast cancer incidence and mortality across 185 countries, Nature Medicine 31(4) (2025), 1154-1162.

DOI: https://doi.org/10.1038/s41591-025-03502-3

[2] R. J. A. Little and D. B. Rubin, Statistical Analysis with Missing Data. 3rd ed. Hoboken, NJ: Wiley, 2020.

[3] S. S. Roy, S. Mallik, F. Ferretti, et al., KNN in high-dimensional spaces: Challenges and adaptive solutions for pattern recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 45(8) (2023), 9456-9472.

[4] S. S. Roy, S. Mallik, F. Ferretti et al., Adaptive-weighted KNN for on-device hyperspectral image classification in IoT remote sensing, IEEE Transactions on Geoscience and Remote Sensing 61 (2023), 1-15.

[5] I. T. Jolliffe and J. Cadima, Principal component analysis: A review and recent developments, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374(2065) (2016), 20150202.

DOI: https://doi.org/10.1098/rsta.2015.0202

[6] M. N. Gurcan, L. E. Boucheron, A. Can, et al., Nuclear feature extraction for breast tumor diagnosis using deep learning, Journal of Pathology Informatics 13(1) (2022), 45.

[7] D. Enders and P. K. Li, Multiple imputation: A flexible tool for handling missing data in medical research. Biostatistics 22(3) (2021), 510-525.

[8] L. Al Shalabi, Z. Shaaban and B. Kasasbeh, Data mining preprocessing: Z-score normalization for enhanced classification in imbalanced datasets, Journal of Computer Science 18(4) (2022), 289-301.

[9] Y. Zhang, L. Wang and X. Li, Nearest neighbor pattern classification in high-dimensional data: Modern adaptations and applications, IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(2) (2024), 789-802.

[10] T. G. Dietterich, A study of cross-validation and bootstrap for accuracy estimation and model selection in modern ML. Journal of Machine Learning Research, 23(1) (2022), 1-45.

[11] M. Sokolova and G. Lapalme, Approximate statistical tests for comparing supervised classification learning algorithms: Recent empirical evaluations, Neural Computation. 35(7) (2023), 1895-1923.

[12] C. D. Manning, P. Raghavan and H. Schütze, Introduction to Information Retrieval (2nd ed., updated 2021). Cambridge: Cambridge University Press, 2021.

■