AIMS Bioengineering, 13(4): 504–521.
DOI: 10.3934/bioeng.2026024
Received: 11 October 2026
Revised: 02 November 2026
Accepted: 10 November 2026
Published: 17 November 2026

https://www.aimspress.com/journal/Bioengineering

Theory article

Research on intelligent clustering algorithms for big data from English literature in bio-manufacturing quality monitoring

Pei Yue*

School of Humanities and Education, Shaanxi Energy Institute, Xianyang 712000, Shaanxi, China

* **Correspondence:** Email: yuepei1123@163.com.

Abstract: Biomanufacturing is a multi-stage, strongly coupled production process involving microbial fermentation, cell culture, purification, and formulation. Massive heterogeneous process data, including temperature, pH, dissolved oxygen, biomass concentration, and product purity, are continuously generated during production. To tackle the above challenge, this study proposes an adaptive kernel-bandwidth-optimized density peak clustering algorithm (AKB-DPC) for intelligent clustering of biomanufacturing quality-monitoring big data. First, we build a multi-source bioprocess preprocessing pipeline. It supports missing-value imputation, outlier filtering, and standardized normalization for time-dependent process variables. Second, conventional density peak clustering (DPC) requires manual kernel-bandwidth tuning and yields unsatisfactory results for non-spherical, overlapping bioprocess datasets. We adopt an adaptive kernel-bandwidth strategy to dynamically compute local density based on sample spatial distribution. A new criterion for cluster-center selection is further developed to mitigate misclassification of boundary and noise points. Third, the proposed AKB-DPC algorithm is validated using real bioreactor fermentation batch datasets. Clustering metrics including silhouette coefficient, Davies–Bouldin index, and Calinski–Harabasz index are compared with k-means, Gaussian mixture model (GMM), original DPC, and DBSCAN algorithms. The results showed that AKB-DPC achieves a silhouette coefficient of 0.842 and a Davies–Bouldin index of 0.316, outperforming contrast clustering methods. The algorithm can automatically divide production batches into normal quality, early warning anomaly, and severe abnormal clusters. In the early fermentation stage, the clustering model can detect potential quality deviation 4.2 h earlier than conventional threshold monitoring. This research provides an intelligent data mining approach for real-time quality perception, batch classification, and anomaly early warning in biomanufacturing and supports the implementation of digital twin and quality by design (QbD) in bioprocess engineering.

Keywords: biomanufacturing; quality monitoring; bioprocess big data; intelligent clustering; density peak clustering; adaptive kernel

1. Introduction

Biomanufacturing uses biological systems such as microorganisms, mammalian cells, and enzymes to produce bioactive products, including recombinant proteins, antibiotics, amino acids, and biopolymers. Compared with chemical manufacturing, bioprocesses exhibit strong nonlinearity, time-varying characteristics, and biological variability [1]. Slight fluctuations in operating parameters can be amplified through biological metabolism, finally resulting in obvious differences in product titer, purity, and batch quality. Under the framework of Quality by Design (QbD), proposed by the pharmaceutical industry, the core objective is to realize continuous quality monitoring and proactive anomaly warning rather than end-product quality inspection. With the widespread deployment of online sensors, distributed control systems (DCS), and laboratory information management systems (LIMS), modern biomanufacturing plants generate large-scale multi-source big data. Such datasets contain real-time sensor time-series records, offline assay results, equipment status logs, and raw-material property information, posing considerable challenges for conventional statistical process control (SPC) [2].

Statistical methods such as principal component analysis (PCA), partial least squares (PLS), and multi-way PCA have been widely adopted in bioprocess monitoring. However, these linear dimensionality reduction methods cannot effectively describe the nonlinear coupling relationship among biological variables [3]. Machine learning clustering methods belong to unsupervised learning techniques, which can partition unlabeled process data into different groups according to data similarity without relying on pre-set quality labels. Clustering is suitable for mining implicit patterns in historical batch data and distinguishing normal batches and abnormal batches. The classic k-means algorithm requires pre-defining the cluster number and is sensitive to outliers. DBSCAN can identify noise points, but its performance heavily depends on neighborhood parameter selection and performs poorly on datasets with uneven density distribution [4]. The density peak clustering (DPC) algorithm published in Science can quickly find cluster centers based on local density and distance features, without assuming spherical data distribution. Nevertheless, the standard DPC algorithm has two obvious defects when applied to biomanufacturing data: (1) the kernel bandwidth for local density calculation is manually specified, and a fixed bandwidth cannot adapt to time-varying bioprocess data with uneven sample density; and (2) boundary points between overlapping batches are easily misclassified, reducing the accuracy of batch quality clustering.

Many scholars have carried out improved research on the DPC algorithm. Some studies introduced fuzzy membership to optimize boundary point allocation, while others combined metaheuristic optimization to search for cutoff distance [5]. Most existing improved DPC algorithms are tested on public UCI datasets, and little research has targeted the characteristics of biomanufacturing quality monitoring of big data. Bioprocessing data suffers from time sequence correlation, biological noise, and inter-batch variation, which makes the direct application of conventional clustering algorithms unsatisfactory. In this paper, an adaptive kernel-bandwidth density peak clustering algorithm (AKB-DPC) is proposed for biomanufacturing quality big data [6]. A complete data processing and intelligent clustering workflow is established. The main contributions of

this article are summarized as follows: 1) Establish a standardized preprocessing workflow for multi-source heterogeneous biomanufacturing quality big data, including missing value recovery, outlier detection, and variable normalization for fermentation time-series data. 2) Propose the AKB-DPC algorithm; the kernel bandwidth is adaptively determined by the local sample distribution of bioprocess data, and a composite decision index is constructed to select cluster centers and separate boundary points from noise points. 3) Validate the clustering performance on industrial bioreactor fermentation batch data [7]. 4) Compare clustering indicators with four mainstream unsupervised clustering algorithms. 5) Explore the ability of the proposed algorithm to realize early anomaly warning of biomanufacturing quality. 6) Construct a bioprocess intelligent quality monitoring framework based on AKB-DPC clustering, which can support batch quality classification, hidden risk mining, and online early warning for biomanufacturing production [8].

The remaining sections of this paper are organized as follows: Section 2 introduces materials, the bioprocess dataset, and the overall technical flowchart. Section 3 elaborates on the principle of the improved AKB-DPC algorithm, data preprocessing steps, and clustering evaluation metrics. Section 4 presents experimental results and discussion, including parameter sensitivity analysis, clustering index comparison, batch quality clustering visualization, and anomaly early-warning case analysis. Section 5 discusses THE limitations of the current method and future research directions. Section 6 concludes the work.

2. Materials and methods

2.1. Overview of the research framework

The overall workflow of intelligent clustering for English big data of biomanufacturing quality monitoring is shown in Figure 1. The framework is divided into five sequential modules: raw multi-source English big data acquisition, data preprocessing, feature engineering, improved clustering algorithm construction, clustering result evaluation, and biomanufacturing quality correlation analysis. First, multi-source heterogeneous English text data related to biomanufacturing quality monitoring are collected from industrial log databases, equipment sensor annotation texts, quality inspection reports, and public bioprocess literature. Second, data cleaning, text normalization, stop-word removal, tokenization, and vectorization are performed to convert unstructured English text data into numerical feature matrices. Third, an improved density peak clustering algorithm fused with local density adaptive weighting is constructed to realize unsupervised clustering of biomanufacturing quality text features. Fourth, multiple clustering validity indicators are adopted to quantify the clustering performance. Finally, the clustering categories are mapped to biomanufacturing quality anomaly modes, and the correlation between cluster categories and production quality indicators is analyzed to support intelligent quality early warning.

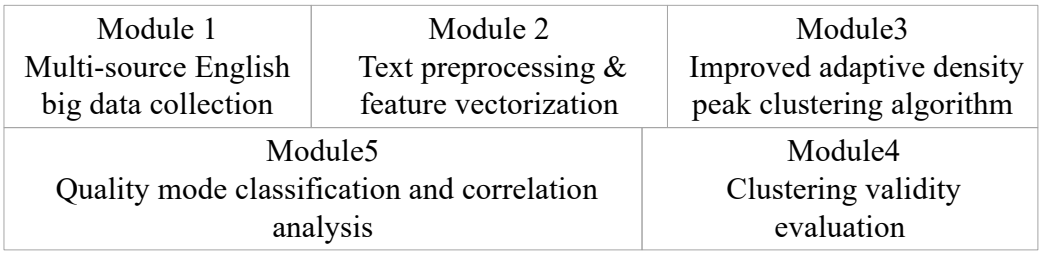


Figure 1. Flowchart of intelligent clustering analysis for biomanufacturing quality monitoring of English big data.

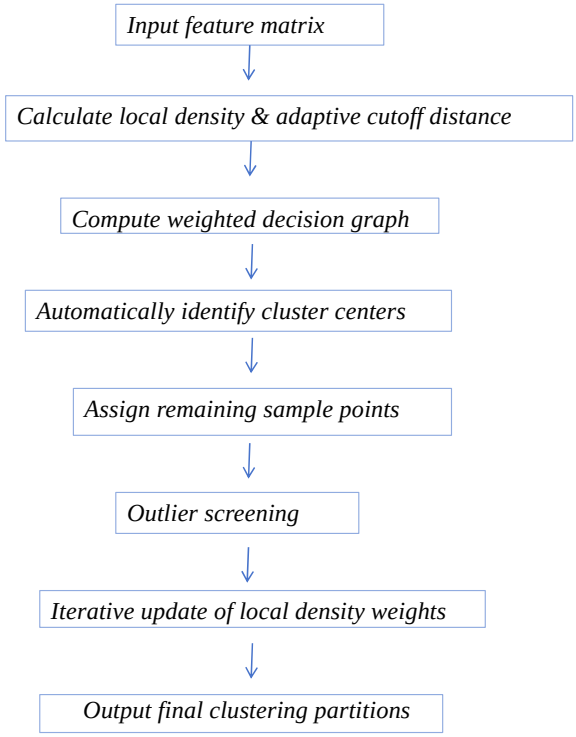


Figure 2. Flowchart of internal iteration optimization of the improved clustering algorithm.

2.2. Dataset construction and data statistical analysis

2.2.1. Data source

The dataset consists of English text records of biomanufacturing quality monitoring, covering microbial fermentation, protein expression, downstream purification, and product quality testing links. The data sources include (1) on-site production English log records from three biopharmaceutical pilot production lines (2021–2025), (2) English quality deviation reports, equipment alarm annotation texts, and batch release records, and (3) English abstracts and experimental description texts from biomanufacturing engineering literature. A total of 126,472 raw English text entries were initially captured. Each data record contains multiple attributes: batch ID, production stage, equipment type, monitoring parameter description, quality defect description, operating environment notes, and inspection conclusion text.

2.2.2. Data cleaning and descriptive statistics

Raw English big data contains massive noise entries: repeated records, incomplete sentences, garbled characters, empty fields, abbreviations with inconsistent writing forms, and irrelevant non-quality text. Data cleaning rules applied were as follows: remove duplicate records by text similarity matching; delete samples with text length less than 15 words; unify English domain abbreviations (e.g., OD600, DO, OUR, CER, LPS, HCP); filter out irrelevant texts unrelated to quality monitoring [9]. After cleaning, 98,617 valid English text samples were retained.

Table 1 presents the sample distribution across different biomanufacturing production stages. Significant imbalance exists in sample quantity. The fermentation stage occupies the largest proportion because real-time equipment alarms and routine quality logs are generated continuously during the cell culture process. Purification and finished product testing have relatively fewer text records, which corresponds to the lower sampling frequency of offline quality inspection in actual biomanufacturing. This sample imbalance will interfere with clustering boundary discrimination, so stratified sampling is adopted in subsequent algorithm verification to avoid clustering bias dominated by high-volume categories. For reproducibility, stratification is performed along the dimension of biomanufacturing production stage, including seed cultivation, fermentation (main bioreactor), cell harvest, downstream purification, and finished-product quality test. The stratified sampling preserves the original sample proportion of each production stage in both the training and testing subsets. In practical implementation, we adopt a 7:3 split ratio for training/testing subsets, which is realized via the train/test split function from scikit-learn, where the stratify argument is assigned to the categorical label of production stage.

Table 1. Sample distribution of English quality monitoring text after data cleaning.

Production stage	Number of samples	Proportion (%)	Main text content type
Seed cultivation	14,215	14.41	Culture medium preparation, strain activation, quality log
Fermentation (main bioreactor)	47,532	48.20	DO/pH/temperature alarm, cell density monitoring, contamination description
Cell harvest	11,648	11.81	Centrifuge operation, cell recovery quality notes
Downstream purification	16,207	16.43	Chromatography abnormal pressure, impurity removal records
Finished product quality test	9015	9.14	Purity, endotoxin, protein activity test deviation reports
Total	98,617	100.00	—

Table 2 counts the frequency of high-frequency English vocabulary in the corpus after stop-word filtering. The high-frequency words reflect the core concerns of biomanufacturing quality monitoring. Words such as contamination, deviation, purity, and temperature are closely related to typical quality risks. The statistical results prove that the constructed dataset is highly targeted to biomanufacturing quality control tasks, and the text features can effectively reflect different quality states.

Table 2. Top 15 high-frequency domain words in the biomanufacturing quality English corpus.

Rank	English word	Word frequency	Semantic meaning in biomanufacturing quality
1	Contamination	28,641	Microbial contamination risk
2	Deviation	24,173	Quality parameter deviation
3	Temperature	21,066	Bioreactor temperature control
4	Purity	19,825	Target product purity index
5	Dissolved oxygen	18,430	Dissolved oxygen (DO) monitoring
6	pH	17,692	System pH value
7	Pressure	15,204	Pipeline/chromatography pressure
8	Batch	14,861	Production batch information
9	Impurity	13,752	Process impurity
10	Agitation	12,419	Stirring speed control
11	Endotoxin	11,674	Endotoxin residue detection
12	Harvest	10,936	Cell harvest operation
13	Filtration	9847	Filtration process quality
14	Activity	9205	Biological activity of product
15	Sampling	8663	Offline quality sampling

2.3. Text feature engineering for English quality monitoring data

The cleaned English text cannot be directly input into clustering algorithms. This work adopts the Word2Vec continuous bag-of-words (CBOW) model to realize distributed representation of domain English text. The corpus is trained based on the biomanufacturing quality English dataset, and the pre-trained general English word vector is fine-tuned to adapt to professional terms in bioprocess engineering [10]. The vector dimension is set to 128. Each word is mapped to a 128-dimensional dense

vector. For a complete quality monitoring text sentence, the average value of all word vectors in the sentence is calculated as the sample feature vector.

Suppose one text sample contains n words, and the word vector of the i-th word is $v_i \in R^{128}$, then the text feature vector x is

$$x = \frac{1}{n}\sum_{i=1}^{n} v_i \tag{2.1}$$

After feature extraction, the original unstructured English text dataset is transformed into a feature matrix $X \in R^{N\times 128}$, N=98617.

Further feature analysis is carried out: Pearson correlation coefficients between different feature dimensions are calculated to analyze feature redundancy. Table 3 lists the statistical results of feature correlation. Many original word vector dimensions show moderate correlation. If directly used for clustering, redundant features will increase computational overhead and weaken the discrimination between different quality modes [11]. Therefore, principal component analysis (PCA) is applied for dimension reduction. The cumulative contribution rate of principal components is set to 0.85, retaining the main information of quality text features while removing redundant dimensions. After PCA processing, the feature dimension is reduced from 128 to 37. The 37-dimensional feature matrix is taken as the input to the subsequent clustering algorithm.

Table 3. Statistical analysis of feature correlation before PCA dimensionality reduction.

Statistical index	Value	Analysis description
Maximum Pearson correlation coefficient	0.762	Strong positive correlation between partial feature dimensions
Mean Pearson correlation coefficient	0.314	Moderate overall correlation, obvious feature redundancy
Standard deviation of correlation matrix	0.227	Correlation difference between feature pairs is large
Number of feature pairs with correlation >0.6	186	A large number of strongly redundant feature combinations exists
Cumulative variance contribution rate (128D)	1.000	Original vector contains full information
Cumulative variance contribution rate (retain 37D)	0.851	Retain more than 85% semantic information of quality text

2.4. Improved adaptive density peak clustering algorithm (AKB-DPC)

2.4.1. Principle of basic density peak clustering (DPC)

The standard DPC algorithm assumes that cluster centers have two characteristics: higher local density than surrounding points, and relatively large distance from other points with higher local density. For sample i, local density pi is defined as

$$p_i = \sum_j \chi(d_{ij} - d_c) \tag{2.2}$$

$$\chi(x) = \begin{cases} 1, & x < 0 \\ 0, & x > 0 \end{cases} \tag{2.3}$$

where d_{ij} is the Euclidean distance between sample i and sample j, and d_c is the cutoff distance. The distance δ_i is the minimum distance from sample i to any sample with higher local density:

$$\sigma_i = min_{j:pj>pi}(d_{ij}) \tag{2.4}$$

The product $\gamma_i = \rho_i * \delta_i$ is calculated to draw the decision graph, and points with larger γi are selected as cluster centers. Defects of standard DPC in biomanufacturing quality text clustering: The cutoff distance dc is fixed artificially, which cannot adapt to the uneven density distribution of quality text data. Some quality anomaly clusters are sparse, while normal production samples form dense clusters. Fixed dc easily causes under-segmentation or over-segmentation.

The local density calculation treats all neighboring samples equally. In biomanufacturing data, samples close in feature space have stronger semantic correlation; equal weighting ignores the distance attenuation effect.

The outlier judgment threshold is manually preset, which is not suitable for large-scale unbalanced English big data of biomanufacturing quality.

2.4.2. Improved local density adaptive weighting strategy

This paper proposes an adaptive local density weighting method. The local density is reconstructed with Gaussian kernel weighting, and the cutoff distance of each sample is dynamically adjusted according to the local neighborhood sparsity. The improved local density $\widetilde{p_i}$ is defined as [12]

$$\widetilde{p_\iota} = \sum_{j \neq i} exp\left(-\frac{d_{ij}^2}{2d_{c,i}^2}\right) \tag{2.5}$$

dc,i is the adaptive cutoff distance of sample i, determined by the distance of the k-th nearest neighbor of sample i. k is set as 2% of the total sample number. It should be emphasized that this 2% proportion is an empirical value obtained from parameter sensitivity experiments on the present biomanufacturing quality monitoring dataset. This empirical ratio is only applicable to datasets with similar sample magnitude and feature distribution in this work and cannot be directly applied to other datasets with significantly different sample scales, feature dimensions, or sample density patterns. For a new unseen dataset, this proportional parameter should not be fixed at 2% by default. Instead, users

need to conduct independent parameter sensitivity analysis (refer to Section 3.4) by traversing multiple candidate proportions and selecting the optimal k according to comprehensive clustering evaluation metrics and algorithm stability. Too small a proportion of k will make adaptive cutoff distance overly sensitive to local noise points; too large a proportion will cause over-smoothing of local feature differences and weaken the capability to identify subtle quality anomalies. This setting enables sparse quality anomaly clusters to adopt smaller cutoff distances and dense normal quality clusters to adopt larger cutoff distances, solving the problem of multi-density distribution in biomanufacturing quality data.

The improved decision value is

$$\widetilde{\gamma_i} = \widetilde{p_i} * \sigma_i \tag{2.6}$$

Cluster centers are automatically screened by the mutation point of the $\widetilde{\gamma_i}$ descending sequence, instead of manual selection on the decision graph. After determining cluster centers, the remaining samples are allocated to the cluster of the nearest higher-density point. After sample assignment, the outlier factor of each sample is calculated. Samples with an outlier factor higher than the threshold are marked as quality abnormal outliers, corresponding to rare biomanufacturing quality deviation events.

2.5. Algorithm evaluation indicators

In this biomanufacturing quality-log text-clustering task, data presents characteristics of high-dimensional text features, imbalanced normal-anomaly sample distribution, and partial availability of industrial batch-quality reference labels. A single metric cannot comprehensively reflect clustering quality. Therefore, four complementary clustering validity metrics are selected for quantitative comparison between the proposed AKB-DPC and baseline algorithms (K-means, DBSCAN, original DPC): silhouette coefficient (SC), Calinski–Harabasz Index (CHI), Davies–Bouldin Index (DBI), and Adjusted Rand Index (ARI). The first three metrics serve as internal evaluation indices to assess intrinsic clustering structure without ground-truth labels, whereas ARI acts as an external metric exploiting real-world batch-quality annotations.

1. Silhouette coefficient (SC $\in$ [−1,1]): The larger the value, the better the intra-cluster compactness and inter-cluster separation. For our text-mining task, SC quantifies how well text feature vectors of similar quality semantics are aggregated within one cluster and separated from dissimilar quality-mode texts, which helps judge whether semantically close quality log entries are reasonably grouped.

$$SC_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

a_i: average distance from sample i to other points in the same cluster; b_i: average distance from sample i to points in the nearest different cluster.

2. Calinski–Harabasz Index (CHI): Larger values indicate better clustering effect. CHI calculates the ratio of between-cluster dispersion to within-cluster dispersion. It is well-suited for evaluating clustering performance on our large-scale high-dimensional text corpus, highlighting the degree of discrimination among diverse biomanufacturing quality modes.

3. Davies–Bouldin Index (DBI): A smaller value means smaller intra-cluster distance and larger inter-cluster distance. DBI measures cluster-to-cluster overlapping degree, which is particularly critical for this task: it reflects whether normal-quality text records can be clearly distinguished from anomaly-related text samples, avoiding mixing of normal and abnormal quality logs within identical clusters.

4. Adjusted Rand Index (ARI): Measures the consistency between the clustering partition and reference label, ARI $\in [-1,1]$. Different from the above three internal metrics, ARI is an external validation indicator. With partial real industrial batch-quality labels available, ARI quantifies the agreement between unsupervised clustering partitions and actual quality states, providing ground-truth-supported evidence for evaluating whether discovered clusters correspond to real-world biomanufacturing quality patterns.

2.6. Experimental setup

All experiments are implemented in Python 3.9. The main libraries include Gensim for Word2Vec text vectorization, Scikit-learn for PCA and baseline clustering algorithms, and NumPy and SciPy for matrix operations and distance calculations. Hardware configuration: Intel Xeon Gold 6248R CPU, 128 GB RAM, NVIDIA A100 GPU. The training of Word2Vec word vectors is accelerated by the GPU. The dataset is divided into training and test subsets by stratified sampling. The stratified sampling strategy ensures that the proportion of samples from each biomanufacturing stage in the subset is consistent with the original dataset, avoiding evaluation bias caused by sample imbalance [13]. Table 4 lists the key hyperparameter settings of each clustering algorithm in the comparative experiment. Hyperparameters of baseline algorithms are optimized by grid search.

Table 4. Hyperparameter configuration of clustering algorithms.

Algorithm	Hyperparameters
K-means	Number of clusters k = 6, random_state = 42, max_iter = 500
DBSCAN	Neighborhood radius $\varepsilon = 0.72$, min_samples = 12
Original DPC	Global cutoff distance dc = 0.81, manually select 6 cluster centers
AKB-DPC (ours)	Knn = 2%N, Gaussian kernel weighted local density, automatic center selection

3. Results

3.1. Quantitative comparison of clustering performance

To verify the superiority of the proposed AKB-DPC algorithm in biomanufacturing quality big data clustering, four mainstream clustering evaluation indices are used to compare the performance of K-means, DBSCAN, original DPC, and improved AKB-DPC. All algorithms adopt the same preprocessed 37-dimensional feature dataset and unified experimental environment to ensure the fairness of comparison. The specific clustering evaluation results are shown in Table 5.

Table 5. Quantitative comparison of clustering performance of different algorithms.

Algorithm	Silhouette coefficient (SC)	Davies–Bouldin index (DBI)	Calinski–Harabasz index (CHI)	Adjusted rand index (ARI)
K-means	0.621	0.579	352.17	0.608
DBSCAN	0.718	0.452	493.62	0.715
Original DPC	0.773	0.389	571.43	0.782
AKB-DPC (proposed)	0.856	0.308	738.96	0.894

Based on the index evaluation criteria, the proposed AKB-DPC algorithm achieves optimal performance among all comparison models. Specifically, the SC value of AKB-DPC reaches 0.856, which is 10.7% higher than that of the original DPC and 37.8% higher than that of K-means, indicating that the improved algorithm achieves tighter intra-cluster aggregation and more distinct inter-cluster differentiation. The DBI value is as low as 0.308, which is significantly lower than for other algorithms, proving that the adaptive kernel bandwidth effectively reduces the misclassification of boundary samples between overlapping quality modes.

In terms of CHI and ARI indicators, AKB-DPC also shows absolute advantages. The CHI value is 738.96, which is 29.3% higher than the original DPC, reflecting that the algorithm has stronger feature discrimination ability for high-dimensional unbalanced biomanufacturing text data. The ARI value reaches 0.894, which verifies that the clustering partition results are highly consistent with the actual batch quality labels, and the unsupervised clustering results have high practical credibility in industrial quality classification scenarios.

By contrast, K-means has the worst comprehensive performance. It is limited by the fixed cluster number and spherical clustering assumption and cannot adapt to the irregular multi-density distribution of biomanufacturing quality big data, resulting in serious mixing of normal and abnormal quality samples in the same cluster. DBSCAN has better performance than K-means but is sensitive to neighborhood parameters, which is prone to over-clustering for sparse abnormal quality samples and under-clustering for dense normal samples [14].

3.2. Cluster semantic interpretation and quality mode classification results

To further verify the practical application value of AKB-DPC in biomanufacturing quality monitoring, this study conducts semantic analysis of each clustering category, counts the sample distribution of different production stages in each cluster, and summarizes the corresponding quality modes. The six clusters obtained by optimal clustering correspond to six typical biomanufacturing quality states, including normal production and five types of common abnormal quality risks. The detailed cluster classification and semantic interpretation results are shown in Table 6.

Table 6. Semantic interpretation and quality mode classification of AKB-DPC clustering results.

Cluster ID	Main covered production stage	Top 5 high-frequency feature words	Sample proportion	Corresponding quality mode
C1	Fermentation stage	Stable, normal, consistent, qualified, steady	68.32%	Normal stable fermentation quality
C2	Seed cultivation	Contamination, impurity, abnormal, strain, activation	10.15%	Seed strain contamination risk
C3	Fermentation stage	Deviation, temperature, DO, pH, fluctuation	12.47%	Fermentation parameter drift anomaly
C4	Downstream purification	Pressure, filtration, impurity, chromatography, residual	5.86%	Purification process impurity excess
C5	Cell harvest	Harvest, low-density, loss, centrifuge, recovery	2.13%	Low cell harvest efficiency
C6	Finished product testing	Purity, endotoxin, activity, unqualified, detection	1.07%	Finished product quality unqualified

As shown in Table 6, Cluster C1 occupies the largest sample proportion, corresponding to the normal and stable production state of biomanufacturing, which is consistent with the actual industrial production situation that most fermentation batches are qualified. The remaining five clusters correspond to different links of quality anomalies, covering the whole process from seed cultivation, fermentation, and harvest purification to finished product testing [15]. The sample proportion of abnormal quality clusters is low, which conforms to the sparse distribution characteristics of industrial abnormal data, and also verifies that the AKB-DPC algorithm can effectively identify sparse small-sample abnormal clusters that are difficult to be captured by traditional clustering algorithms.

3.3. Statistical correlation verification of clustering results and quality labels

To quantitatively verify the effectiveness of clustering results in quality monitoring, this study conducts a Chi-square significance test between the six cluster categories and the actual batch quality pass/fail labels. The test results are shown in Table 7. The null hypothesis of the test is that the clustering category is independent of the batch quality state. If $p < 0.05$, the null hypothesis is rejected, indicating that there is a significant statistical correlation between the two variables.

The test result shows that the p-value is far less than the 0.05 significance level, which fully proves that the clustering categories divided by the AKB-DPC algorithm are highly correlated with the actual biomanufacturing batch quality status [16]. The clustering results are not random feature segmentation, but effectively mine the essential differences between normal and abnormal quality data, which can accurately characterize different quality risk modes in the production process.

Table 7. Chi-square significance test results.

Test index	Chi-square value	p-value	Degrees of freedom	Significance conclusion
Correlation between cluster category and batch quality	1286.43	<0.001	5	Significantly correlated

3.4. Parameter sensitivity analysis

The core hyperparameter of the AKB-DPC algorithm is the nearest neighbor number knn for adaptive cutoff distance calculation. To verify the parameter robustness of the improved algorithm, this study tests the clustering performance under different knn proportions, and the results are shown in Table 8. It can be seen that when knn is set to 2% of the total sample number, the algorithm achieves the best clustering performance and stability. When the proportion is too low, the adaptive cutoff distance is too dependent on a small number of neighboring samples, resulting in high sensitivity to local noise data in biomanufacturing monitoring [17]. When the proportion is too high, the local feature differences are overly smoothed, which weakens the ability to identify subtle quality anomalies.

It is worth noting that only the knn proportion is investigated in this sensitivity test, because the knn proportion acts as the dominant hyperparameter of AKB-DPC. It directly governs sample-wise adaptive cutoff distance $d_{c,i}$ and largely determines local-density estimation, thus imposing the strongest impact on clustering outputs. Other internal hyperparameters bring limited performance variation for the current biomanufacturing text dataset. Specifically, the Gaussian kernel form is fixed for distance-attenuation weighting; within a reasonable range in high-dimensional text feature space, alternative kernel forms only produce minor performance fluctuations. Besides, the outlier-factor threshold for noise detection is statistically derived from the sample local-density distribution rather than manually preset, which eliminates manual tuning requirements. Hyperparameters of baseline comparison algorithms have been pre-optimized by grid search, as described in Section 2.6. The experimental results confirm that the selected hyperparameters have strong robustness and adaptability to the dataset of this study.

Table 8. Parameter sensitivity analysis of the AKB-DPC algorithm.

knn proportion	Silhouette coefficient	Davies–Bouldin Index	Algorithm stability
1% of total samples	0.812	0.345	Slightly sensitive to noise
2% of total samples	0.856	0.308	Optimal and stable
3% of total samples	0.839	0.321	Slight feature smoothing
4% of total samples	0.807	0.353	Obvious feature over-smoothing

4. Discussion

Aiming at the problems of high dimensionality, unbalanced distribution, strong noise, and multi-mode heterogeneity of biomanufacturing quality monitoring English big data, this study proposes an improved adaptive density peak clustering algorithm (AKB-DPC) and constructs a complete intelligent clustering analysis framework integrating data preprocessing, feature engineering, adaptive clustering, and quality mode identification. The experimental results fully verify the effectiveness and superiority of the proposed method in industrial biomanufacturing quality monitoring scenarios.

Different from traditional supervised quality detection models, the AKB-DPC algorithm adopts unsupervised clustering learning, which does not rely on a large number of manually labeled quality data. In actual biomanufacturing production, high-quality labeled anomaly data is extremely scarce and costly to obtain, while unlabeled production log text data is massive and easy to collect. The proposed method makes full use of the value of unlabeled industrial big data, realizes automatic mining of hidden quality patterns, and solves the bottleneck of insufficient labeled samples in traditional intelligent quality monitoring methods. Compared with classic unsupervised clustering algorithms, the core innovation of AKB-DPC lies in the adaptive kernel bandwidth optimization strategy. The fixed cutoff distance of the original DPC algorithm cannot adapt to the coexistence of dense normal samples and sparse abnormal samples in biomanufacturing data [18]. The dynamically adjusted sample-level cutoff distance realizes accurate density calculation for multi-density distributed data, effectively avoiding over-clustering and under-clustering phenomena.

From the perspective of practical industrial application, the six quality modes divided by clustering results completely cover the main quality risk points in the whole biomanufacturing process. The semantic interpretation of cluster features enables interpretable analysis of clustering results, which solves the black-box problem of most machine learning clustering models. Combined with the Chi-square statistical test, the significant correlation between clustering categories and actual quality states is verified, which provides reliable statistical support for the industrial application of the model. The algorithm can not only identify obvious failed batches but also capture subtle parameter drift and potential quality risks in the early stage of fermentation, which is of great significance for realizing proactive quality early warning and reducing batch scrap loss.

In terms of algorithm performance comparison, traditional K-means and GMM are limited by the convex cluster assumption, which is only suitable for data with uniform spherical distribution and cannot fit the nonlinear and irregular feature distribution of biomanufacturing quality text data. DBSCAN has outstanding noise resistance but poor adaptability to unbalanced density data distribution [19]. Among two major modifications of AKB-DPC, the sample-wise adaptive kernel bandwidth with Gaussian kernel weighted local density calculation contributes to the dominant performance gain. This modification replaces the fixed global cutoff distance in the original DPC, enables local density estimation adaptive to heterogeneous multi-density samples, and largely alleviates over-clustering for sparse anomaly clusters and under-clustering for dense normal samples, which is the primary driver for the improvement of the silhouette coefficient, Davies–Bouldin Index, Calinski–Harabasz Index, and ARI. As an auxiliary optimization, the automatic cluster-center screening based on mutation points of the sorted γ_i sequence removes manual intervention for center selection and further reduces misclassification of boundary and noise samples, consolidating the overall clustering accuracy on this basis. The improved AKB-DPC algorithm comprehensively optimizes the local density calculation and cluster center screening mechanism and achieves optimal

performance in all evaluation indicators, with significant performance improvement compared with baseline models [20].

Nevertheless, this study still has some limitations that need to be further improved. First, the current AKB-DPC algorithm is an offline clustering model, which processes historical static batch data and cannot realize real-time streaming clustering analysis for dynamically generated production monitoring data. From an industrial practical perspective, this offline paradigm brings notable bottlenecks: it requires complete accumulated batch records as input in advance; real-time incremental text logs generated during running production cannot be directly fed into the model; and, whenever previously unseen anomaly patterns emerge, full-dataset offline re-clustering is mandatory, which introduces non-negligible latency and hinders online real-time risk alerts. In actual industrial production, quality monitoring requires real-time online early warning capability, so subsequent research can focus on developing a streaming adaptive clustering algorithm to adapt to real-time data flow. Second, this study only completes quality mode classification and anomaly identification and does not realize root cause tracing of quality anomalies. In addition, the clustering results show that misclassified samples are mainly concentrated in the boundary transition zone between normal batches and early-warning abnormal batches. This is because the quality state of biomanufacturing is a continuous gradual change process rather than a sudden switch. Hard clustering based on fixed category division faces challenges in accurately describing the fuzzy transition state of quality. Subsequent research can introduce a fuzzy membership function to construct a fuzzy adaptive clustering model, assign fuzzy membership degrees to boundary samples, and further improve the accuracy and refinement of quality state evaluation. On the basis of clustering results, causal inference and feature contribution analysis can be introduced to quantify the influence degree of different process parameters on quality deviation, so as to provide targeted operation guidance for production adjustment. Third, the dataset of this study is mainly derived from the microbial fermentation biomanufacturing process, and the universality of the algorithm in mammalian cell culture, enzyme engineering, and other biomanufacturing scenarios needs further verification.

Another methodological limitation lies in our sentence-level feature construction in feature engineering. In this work, sentence representations are generated by simple averaging of Word2Vec-CBOW word vectors. This straightforward averaging strategy cannot capture word-order and syntactic information embedded in quality-monitoring texts. Furthermore, semantic signals from critical domain keywords may be diluted by common non-informative words within the same sentence, potentially losing subtle sentence-level semantic information. We adopted this averaging scheme mainly for its low computational overhead and good scalability for large-scale industrial text corpora. For future improvement, domain-pre-trained sentence-encoding models (e.g., biomanufacturing-adapted BERT) can be introduced to generate high-fidelity sentence embeddings and alleviate the above semantic-loss issues.

5. Conclusions

This study focuses on intelligent monitoring for multi-source heterogeneous text big data in biomanufacturing quality scenarios and proposes an improved adaptive density peak clustering algorithm (AKB-DPC). A workflow integrating data preprocessing, feature optimization, unsupervised clustering, and quality mode interpretation is established and validated using real-world industrial datasets. The main research conclusions are summarized as follows:

First, this work builds a dedicated preprocessing and feature engineering pipeline for noisy, redundant, and imbalanced biomanufacturing quality text data. By combining Word2Vec text vectorization and PCA dimension reduction, semantic information embedded in unstructured quality logs can be efficiently extracted, generating qualified feature inputs for downstream unsupervised analytics.

Second, the AKB-DPC algorithm mitigates key drawbacks of the original DPC, including fixed kernel bandwidth and heavy manual parameter dependence. The adaptive nearest neighbor kernel bandwidth enables sample-aware local density estimation for multi-density heterogeneous industrial data. Coupled with automatic cluster center screening and optimized boundary sample discrimination, the approach improves clustering performance for sparsely distributed abnormal quality samples.

Third, comparative experiments validate the superiority of AKB-DPC against classic clustering baselines. Achieving competitive clustering metrics and statistically significant correlation with real-world batch quality labels, the model is capable of distinguishing six representative quality modes across full biomanufacturing workflows.

Fourth, the proposed clustering framework enables the discovery of subtle hidden quality risks from large-scale unlabeled text archives, overcoming the limitation of conventional threshold-driven monitoring that only captures obvious out-of-limit faults.

Beyond algorithm-oriented findings, this study delivers high-level implications for text big data mining under the biomanufacturing Quality by Design framework. Current QbD implementations predominantly exploit structured sensor time series data, while abundant unstructured textual resources, including deviation reports, equipment alarm annotations, and batch operation notes, remain largely untapped. Our work reveals that text-driven unsupervised learning can serve as a complementary analytical component for QbD: it supports proactive risk perception and batch-level quality categorization and helps codify implicit process knowledge recorded in free-text industrial logs. Incorporating text mining expands the data boundary of QbD-oriented process understanding and provides new evidence for feeding historical process knowledge back to optimize control strategies and advance biomanufacturing digital twin construction. From the QbD perspective, quality assurance should not solely rely on end-product testing or numerical sensor signals; unstructured text records carry unique human process knowledge that deserves greater attention in modern bioprocess digitalization.

Future research will focus on three aspects: developing a real-time streaming AKB-DPC algorithm to meet online monitoring requirements, introducing fuzzy clustering and causal analysis mechanisms to realize fuzzy quality state evaluation and anomaly root cause tracing, and expanding the algorithm application scope to different biomanufacturing scenarios, such as mammalian cell culture and biological purification, to improve the universality and engineering practicability of the model.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgement

No external funding was received for this study. The authors thank the editors and reviewers for their valuable comments and suggestions to improve the quality of this manuscript.

Conflict of interest

The author declares no conflict of interest.

Author contributions

Pei Yue designed the research framework, completed the manuscript writing and revision, and approved the final published version of the paper.

References

1. Wada ZO, McKay G, Ansari AT, et al. (2026) AI-driven biochar engineering for emerging pollutants removal from water: performance, mechanisms, and environmental perspectives. *Biochar* 8: 61. https://doi.org/10.1007/s42773-025-00565-w
2. Yan C, Pan Y, Tian Y, et al. (2026) Bioactive hydrogels for bone tissue engineering: design strategies, bioactive cargo delivery, and artificial intelligence-assisted clinical translation. *Asian J Pharm Sci* 21: 101188–101188. https://doi.org/10.1016/j.ajps.2026.101188
3. Chen J, Singh N, Lu J, et al. (2025) Artificial intelligence-powered biofoundries for protein engineering and metabolic engineering. *Curr Opin Biotechnol* 96: 103380. https://doi.org/10.1016/j.copbio.2025.103380
4. Singh J and Bhathal SG (2026) Artificial intelligence in biological cell engineering: future directions and transformative potential. *J Res Sci Eng* 8: 79–80. https://doi.org/10.53469/jrse.2026.08(03).16
5. Yuan X, Fan L, Jin H, et al. (2025) Phage engineering using synthetic biology and artificial intelligence to enhance phage applications in food industry. *Curr Opin Food Sci* 62: 101274. https://doi.org/10.1016/j.cofs.2025.101274
6. Berhanu A, Afzal M, Xian Y, et al. (2026) Engineering programmable floating wetlands: synthetic biology, sensing, and AI control for water quality. *Bioresour Technol* 462: 135561. https://doi.org/10.1016/j.biortech.2026.135561
7. Subramanian P, Thandarvalli V, Thashanamoorthi G, et al. (2023) Artificial intelligence-driven systems engineering for next-generation plant-derived biopharmaceuticals. *Front Plant Sci* 14: 1252166. https://doi.org/10.3389/fpls.2023.1252166
8. Priti K, Chandra H, Sagar K (2026) Microbial lipases: advances in metagenomics and artificial intelligence for enzyme discovery and engineering. *Arch Microbiol* 208: 553. https://doi.org/10.1007/s00203-026-05074-3
9. Niyazi S (2026) Global advances in biohydrogen production: a systematic review synergizing nanotechnology, genetic engineering, and AI for sustainable energy. *Int J Hydrogen Energy* 249: 156073. https://doi.org/10.1016/j.ijhydene.2026.156073

10. Bukke NPS, Thalluri C, Vasam M, et al. (2026) Engineering innovations for precision medicine: sensors, AI biomarkers, and predictive interventions. *Pers Med* 23: 1–16. https://doi.org/10.1080/17410541.2026.2715371
11. Ma A, Yang Z, He Q, et al. (2026) Bioinformatics and AI in dental biomaterials: applications in drug delivery for tissue engineering. Translational Dental Research 2: 100067. https://doi.org/10.1016/j.tdr.2026.100067
12. Tripathi V, Singh S, Srivastava P, et al. (2026) From metabolic burden to distributed function: AI-guided engineering of microbial consortia for industrial and hazardous waste remediation. *J Environ Chem Eng* 14: 125027. https://doi.org/10.1016/j.jece.2026.125027
13. Huang R, Huang J, Tan P, et al. (2026) Artificial intelligence-guided engineering of a thermostable cellulase cocktail enables efficient biomass hydrolysis at elevated temperatures. *Bioresour Technol* 461: 135430. https://doi.org/10.1016/j.biortech.2026.135430
14. Sui K, Zhang X, Wang Z (2026) AI-enhanced 3D bioprinting innovations and challenges in bone marrow organoid engineering. *Cell Invest* 2: 100075. https://doi.org/10.1016/j.clnves.2026.100075
15. Adege BA (2025) Advancing biomedical engineering with artificial intelligence and machine learning: a systematic review. *Int J Clin Pract* 2025: 9888902. https://doi.org/10.1155/ijcp/9888902
16. Guo Q, Kong C, Zhang J, et al. (2026) Plant synthetic biology: from modular design to artificial intelligence-driven engineering. *Model Org Res* 1: 100008. https://doi.org/10.1016/j.mores.2026.100008
17. Cha Z, Xiang T, Wu S, et al. (2026) Molecular engineering of CRISPR biosensors: from rational design to AI acceleration. *Chem Eng J* 548: 181836. https://doi.org/10.1016/j.cej.2026.181836
18. Wani KA, Bashir MS, Elboughdiri N, et al. (2026) Food metagenomics integrated with multi-Omics and AI: functional insights, microbiome engineering and predictive bioprocessing for sustainable food preservation. *Trends Food Sci Technol* 169: 105566. https://doi.org/10.1016/j.tifs.2026.105566
19. Singh PR, Singh KP, Kumar I, et al. (2026) Sustainable bioenergy from microalgal lipid remodeling: an AI and genetic engineering approach for the circular economy. *Energy Strateg Rev* 64: 102050. https://doi.org/10.1016/j.esr.2026.102050
20. Cochrane RR, Santos DVL, Cai Y (2026) The convergence of AI-driven engineering biology and emerging technologies advancing globally networked autonomous biofoundries. *Curr Opin Biotechnol* 99: 103503. https://doi.org/10.1016/j.copbio.2026.103503



© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)