Enhancing Radiomics Clustering: Nonlinear and Nonparametric Approaches in Biological Data Analysis

To the Editor:

The careful selection of radiomics analysis tools is crucial for ensuring accurate biological interpretations. Using tools with inappropriate assumptions can significantly distort results and lead to erroneous conclusions. Thus, researchers should treat radiomics data as inherently nonlinear and nonparametric and choose analytical methods that effectively accommodate these properties.

Li et al.[1] investigated a radiomics-based support vector machine to differentiate the molecular drivers of lung adenocarcinoma progression. For data clustering, the researchers utilized ConsensusClusterPlus, a tool that applies a consensus clustering approach through iterative resampling combined with user-defined algorithms—typically hierarchical clustering, k-means, or partitioning around medoids (PAM)—to robustly estimate the underlying data structure.

Nevertheless, concerns have been raised about using ConsensusClusterPlus in biological analyses. Its reliance on specific assumptions may not align with the inherently nonlinear and nonparametric nature of biological data. This misalignment can result in distorted clustering outcomes, potentially leading to misinterpretations and erroneous conclusions. Therefore, it is crucial for researchers to critically assess the methodological assumptions of such tools and consider alternative approaches that better accommodate the complex properties of biological data.

Common clustering methods—such as hierarchical clustering, k-means, and PAM—each have inherent limitations when applied to complex biological data. Hierarchical clustering uses standard distance metrics (e.g., Euclidean) to construct a nested dendrogram, which can misrepresent dissimilarities in data with intricate relationships. K-means assumes that clusters are spherical and uniform in size, summarizing them with centroids; nevertheless, irregular cluster shapes can undermine its effectiveness in minimizing squared distances. Although PAM is more robust when medoids and supporting arbitrary dissimilarity measures are used, a single representative point may not adequately capture a cluster's complexity.

Although supervised machine learning benefits from ground truth labels that validate prediction accuracy, unsupervised clustering lacks such benchmarks, making evaluation more challenging. To perform clustering effectively, it is crucial to assess clustering quality and accurately determine the optimal number of clusters. This article advocates the use of nonlinear and nonparametric methods, such as DBSCAN[2] and OPTICS,[3] which are well-suited for complex and irregular data structures that often challenge traditional clustering techniques for biological analysis. Unlike conventional methods, these approaches do not assume spherical shapes or equal cluster sizes, allowing greater flexibility in capturing data nuances.

Moreover, clustering quality and the optimal configuration can be rigorously assessed using metrics such as the Silhouette Score,[4] Davies-Bouldin Index,[5] and Gap Statistic.[6] The Silhouette Score gauges how well each data point fits within its own cluster compared with neighboring clusters, ensuring both cohesion and proper separation. The Davies-Bouldin Index quantifies the average similarity between clusters by comparing the dispersion within a cluster to the separation between clusters—a lower index value indicating better inter-cluster distinction. The Gap Statistic determines the optimal number of clusters by contrasting the observed intra-cluster variation with that expected from a random uniform distribution.

In contrast to hierarchical clustering, k-means, or PAM, these advanced methods bring several key advantages. For instance, unlike hierarchical clustering—which heavily depends on distance metrics and may fail with complex data structures—nonlinear and nonparametric methods do not assume specific shapes or sizes. Similarly, k-means and PAM often assume spherical clusters and equal sizes, limitations that these alternative approaches overcome by accommodating irregular cluster formations and noise. This ultimately leads to more reliable and data-reflective clustering outcomes.

CRediT Authorship Contribution Statement

Yoshiyasu Takefuji: Conceptualization, Writing, Formal analysis, Investigation.

Address correspondence to: Yoshiyasu Takefuji, PhD, Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan. E-mail: takefuji@keio.jp

Cite this article as: Takefuji Y. Enhancing radiomics clustering: nonlinear and nonparametric approaches in biological data analysis. *J Thorac Oncol* 2025;20:e65-e66

© *2025 International Association for the Study of Lung Cancer. Published by Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.*

ISSN: 1556-0864

https://doi.org/10.1016/j.jtho.2025.02.019

Disclosure

The author has no conflict of interest.

Yoshiyasu Takefuji, PhD
Faculty of Data Science
Musashino University
Tokyo
Japan

References

1. Li HJ, Qiu ZB, Wang MM, et al. Radiomics-based support vector machine distinguishes molecular events driving the progression of lung adenocarcinoma. *J Thorac Oncol*. 2025;20:52–64.
2. Cheng D, Zhang C, Li Y, et al. GB-DBSCAN: a fast granular-ball based DBSCAN clustering algorithm. *Inf Sci*. 2024;674:120731.
3. Grabowski E, Kuo J. Comparing K-means and OPTICS clustering algorithms for identifying vowel categories. *Proc Ling Soc Amer*. 2023;8:5488.
4. Shutaywi M, Kachouie NN. Silhouette analysis for performance evaluation in machine learning with applications to clustering. *Entropy (Basel)*. 2021;23:759.
5. Ros F, Riad R, Guillaume S. PDBI: a partitioning Davies-Bouldin index for clustering evaluation. *Neurocomputing*. 2023;528:178–199.
6. Khan IK, Daud HB, Zainuddin NB, et al. Determining the optimal number of clusters by Enhanced Gap Statistic in K-mean algorithm. *Egypt Inform J*. 2024;27:100504.

A Response to the Letter to the Editor: "Enhancing Radiomics Clustering: Nonlinear and Nonparametric Approaches in Biological Data Analysis"

To the Editor:

We sincerely appreciate the thoughtful comments from Prof. Yoshiyasu Takefuji and are pleased to respond. Prof. Takefuji raised important concerns regarding the tools used for unsupervised clustering of biological data and suggested several alternative methods that may be more suitable for such data.

In our article, we aimed to explore the biological similarity underlying radiological similarity.[1] The first step in addressing this problem was to define radiological similarity, and we carefully selected our methodology. Our approach was inspired by a previously published study by Perez-Johnston et al.,[2] who used consensus clustering with the R package ConsensusClusterPlus to classify lung nodules based on radiological features. Compared with their data set, our data set included early stage lung adenocarcinoma and was larger in scale, encompassing lung adenocarcinoma from pre-invasive to invasive stages. Therefore, we adopted a similar strategy for our analysis.

After dimensionality reduction using t-distributed stochastic neighbor embedding (t-SNE), we observed the potentially nonspherical structure of our data set and decided to use hierarchical clustering and partitioning around medoids (PAM) for further investigation. Although hierarchical clustering did not perform well in the consensus matrix, PAM with Manhattan distance demonstrated promising results (Fig. 1*A* and *B*). Subsequent t-SNE visualization confirmed that PAM worked effectively with our data set (Fig. 1*C*). As found in the t-SNE plot, the data structure was nonspherical, suggesting that Davies-Bouldin Index and Silhouette Score might not be suitable for evaluating clustering quality in this context. Instead, we used consensus clustering to enhance the robustness of the clusters and determine the optimal number of clusters. The Gap Statistic value was 1.48 (SD: 0.0029) when cluster number was 4, which indicated a strong clustering structure. After cluster identification, we conducted analyses at the clinicopathologic, genomic, and transcriptomic levels. The results revealed consistent biological characteristics across different levels and data sets, supporting the validity of our approach.

We greatly value the insights from Prof. Takefuji as an expert in data science. We fully agree that the selection of methods for biological interpretations should be tailored to the data structure. Although conventional tools can perform well when applied appropriately, we acknowledge the potential benefits of advanced methods and novel algorithms for analyzing biological data.

The comments of Prof. Takefuji have also inspired us to refine our future research. Given the progressive nature of lung cancer, lung nodules gradually exhibit behavioral changes on computed tomography scans. We

Address correspondence to: Wen-Zhao Zhong, PhD, Guangdong Lung Cancer Institute, Guangdong Provincial People's Hospital & Guangdong Academy of Medical Sciences, 106 Zhongshan Er road, Guangzhou, Guangdong 510080, People's Republic of China. E-mail: zhongwenzhao@gdph.org.cn

Cite this article as: Li HJ, Qiu ZB, Wang MM, et al. A response to the letter to the editor: "Enhancing Radiomics Clustering: Nonlinear and Nonparametric Approaches in Biological Data Analysis." *J Thorac Oncol* 2025;20:e66-e68

© 2025 International Association for the Study of Lung Cancer. Published by Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

ISSN: 1556-0864

https://doi.org/10.1016/j.jtho.2025.03.039