



Limitations of principal component analysis in COVID-19 CT image classification

ARTICLE INFO

Keywords:

Nonlinear medical imaging data
PCA assumption violations
Feature selection bias
Nonparametric alternatives
Biological data complexity

ABSTRACT

This study critically evaluates the methodological approach employed by Nayak et al. in COVID-19 CT scan classification using their Dense-Res-Inception Ensemble Network with PCA dimensionality reduction. We demonstrate fundamental limitations in applying PCA to nonlinear biological imaging data, where key mathematical assumptions—including linearity, meaningful correlations, and homoscedasticity—are frequently violated. Using the MNIST dataset as a benchmark, we comparatively assessed three dimensionality reduction techniques: PCA, High Variance Gene Selection, and Feature Agglomeration (FA). Results confirm that FA significantly outperforms PCA (92.79% vs 83.76% accuracy) by preserving crucial spatial relationships within image data. This performance disparity highlights the critical importance of methodological alignment with data characteristics in medical imaging analysis. We propose that nonlinear dimensionality reduction approaches better accommodate the complex relationships inherent in biological systems, potentially enhancing both computational efficiency and diagnostic reliability in clinical applications requiring rapid assessment.

1. Introduction

Nayak et al. conducted an innovative study on the rapid and accurate classification of COVID-19 severity in CT scans using the Dense-Res-Inception Ensemble Network (DRIEN) model integrated with an advanced feature selection methodology [1]. Their analytical framework incorporated Principal Component Analysis (PCA) as a dimensionality reduction technique to manage the high-dimensional imaging data, while the Chaotic Enriched Kookaburra Optimization Algorithm was implemented for feature selection to identify the most significant imaging attributes that correlate with disease severity progression [1]. This hybrid approach aimed to enhance both computational efficiency and diagnostic accuracy in a clinical context where timely assessment is critical.

This paper, however, raises substantial methodological and empirical concerns regarding the application of PCA for feature reduction in biological and medical image analysis contexts. The fundamental issue lies in the inherent mismatch between PCA's underlying mathematical assumptions and the nonlinear, nonparametric nature of biological and medical imaging data. This incongruity potentially compromises the validity of the feature reduction process, subsequently affecting the reliability of the classifications and clinical interpretations derived from the reduced feature set.

PCA operates under several critical assumptions that are frequently violated in biological data analysis: it presupposes linear relationships between variables, requires meaningful correlations among the original features, assumes continuous and appropriately standardized data distributions, demands adequate sample sizes relative to feature dimensions, relies on homoscedasticity (uniform variance), and functions optimally with minimal outlier influence. When these assumptions are violated—as commonly occurs in complex biological systems and medical imaging data—the resulting principal components may not accurately represent the underlying data structure, potentially distorting

outcomes and leading to misleading conclusions [2–9]. Our systematic evaluation using benchmark datasets has demonstrated PCA's limitations compared to nonlinear unsupervised machine learning models such as Feature Agglomeration (FA) and High Variance Gene Selection (HVGS) in accurately preserving the information content of nonlinear biological data.

PCA exhibits two fundamental limitations that significantly impact its suitability for biological data analysis. First, it operates exclusively on the feature space without consideration of target variables, making it an unsupervised technique that cannot prioritize features based on their relevance to the outcome of interest (COVID-19 severity in this case). This fundamental characteristic means PCA may preserve variance that is mathematically significant but biologically irrelevant to the classification task, while potentially discarding features with lower variance that actually carry critical diagnostic information. The second limitation stems from PCA's inherently linear mathematical framework, which assumes that the principal axes of variation are straight lines in high-dimensional space. Biological systems, however, frequently exhibit complex nonlinear relationships and interactions, with data distributions that violate parametric assumptions of normality and homogeneity of variance. When applied to such data, PCA's linear transformations may create principal components that fail to capture the true nonlinear structure of the underlying biological relationships, resulting in suboptimal feature reduction that compromises downstream analysis and classification accuracy.

To substantiate PCA limitations in image-based feature selection, we utilized the MNIST dataset comprising 70,000 samples and 784 features from 28×28 pixel images. This methodological analysis employs well-characterized benchmark data since the original Nayak et al.'s dataset remains inaccessible. Our argument builds upon robust theoretical foundations and substantial empirical evidence demonstrating how linear methods (like PCA) applied to nonlinear data structures introduce significant distortions that undermine analytical validity—a

<https://doi.org/10.1016/j.bspc.2025.108353>

Received 13 May 2025; Received in revised form 12 June 2025; Accepted 21 July 2025

Available online 24 July 2025

1746-8094/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

fundamental methodological incongruence documented across over 300 peer-reviewed publications spanning multiple scientific disciplines.

We compared three dimensionality reduction methods: PCA, HVGS, and FA to select the top 30 features, followed by Random Forest cross-validation using these reduced features. Importantly, we applied no scaling, no transformation, or no normalization to isolate the methods' intrinsic performance. FA substantially outperformed both alternatives, achieving 92.79 % accuracy with the highest consistency (± 0.0020). This superior performance stems from FA's preservation of local spatial relationships critical in image data. HVGS and unscaled PCA demonstrated lower accuracies (84.41 % and 83.76 % respectively), with PCA showing slightly higher variability (± 0.0026).

These findings reinforce our central argument that methodological alignment with data characteristics significantly impacts performance—FA's hierarchical clustering approach effectively maintains structural information inherent in image data, while variance-only methods discard important spatial relationships. While we acknowledge PCA's value as an exploratory tool in certain contexts, particularly when assumptions of linearity are met, our results demonstrate its limitations when applied to inherently nonlinear data structures like medical imaging.

Our research identifies three categories of methodological misapplications: violations of statistical assumptions, ground truth challenges in model interpretation, and implementation errors in data preprocessing. Nayak et al.'s work illustrates the first category by applying linear assumptions to nonlinear medical imaging data. Rather than relying solely on PCA, we advocate for implementing complementary multi-faceted approaches that better accommodate the complex, nonlinear relationships in biological and medical imaging systems. All analysis code with cross-validation procedures, [pcavhgsfa.py](https://github.com/y-takefuji/mnist), is publicly available in GitHub repository to ensure full reproducibility [10].

Authors' contributions: Yoshiyasu Takefuji completed this research and wrote this article.

According to ScholarGPS, Yoshiyasu Takefuji holds notable global rankings in several fields. He ranks 54th out of 395,884 scholars in neural networks (AI), 23rd out of 47,799 in parallel computing, and 14th out of 7,222 in parallel algorithms. Furthermore, he ranks highest in AI tools and human-induced error analysis, underscoring his significant contributions to these domains.

CRedit authorship contribution statement

Yoshiyasu Takefuji: Writing – review & editing, Writing – original

draft, Visualization, Validation, Software, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest


The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

- [1] T.K. Nayak, A.C.S. Rao, Rapid and accurate classification of Covid-19 severity in CT scans using DRIEN model and advanced feature selection, *Biomed. Signal Process. Control* 109 (2025) 108052, <https://doi.org/10.1016/j.bspc.2025.108052>.
- [2] E.L. Dyer, K. Kording, Why the simplest explanation isn't always the best, *Proc. Natl. Acad. Sci. U. S. A.* 120 (52) (2023) e2319169120, <https://doi.org/10.1073/pnas.2319169120>.
- [3] P.M. Cristian, V.J. Aaron, E.D. Armando, et al., Diffusion on PCA-UMAP Manifold: the impact of data structure preservation to denoise high-dimensional single-cell RNA sequencing data, *Biology (basel)*. 13 (7) (2024) 512, <https://doi.org/10.3390/biology13070512>.
- [4] Y. Yao, A. Ochoa, Limitations of principal components in quantitative genetic association models for human studies, *Elife* 12 (2023) e79238, <https://doi.org/10.7554/eLife.79238>.
- [5] E. Elhaik, Principal component analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated, *Sci. Rep.* 12 (1) (2022) 14683, <https://doi.org/10.1038/s41598-022-14395-4>.
- [6] N. Mohseni, E. Elhaik, Biases of principal Component Analysis (PCA) in Physical Anthropology Studies require a Reevaluation of Evolutionary Insights, *Elife* 13 (2024) RP94685, <https://doi.org/10.7554/eLife.94685.2>.
- [7] M. Lenz, F.J. Müller, M. Zenke, A. Schuppert, Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data, *Sci. Rep.* 6 (2016) 25696, <https://doi.org/10.1038/srep25696>.
- [8] R. Dey, S. Lee, Asymptotic properties of principal component analysis and shrinkage-bias adjustment under the generalized spiked population model, *J. Multivar. Anal.* 173 (2019) 145–164, <https://doi.org/10.1016/j.jmva.2019.02.007>.
- [9] P. Mehta, C.H. Wang, A.G.R. Day, et al., A high-bias, low-variance introduction to machine learning for physicists, *Phys. Rep.* 810 (2019) 1–124, <https://doi.org/10.1016/j.physrep.2019.03.001>.
- [10] GitHub. [pcavhgsfa.py](https://github.com/y-takefuji/mnist). <https://github.com/y-takefuji/mnist>.

Yoshiyasu Takefuji 

Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku,
Tokyo 135-8181, Japan
E-mail address: takefuji@keio.jp.