

Integrating autoencoder deep neural networks with principal component-based methods for subpopulation identification

Camila Ferreira Azevedo^{1✉} , Jaquicele Aparecida da Costa¹ ,
Ana Carolina Campana Nascimento¹ , Moyses Nascimento¹ 
and Cosme Damião Cruz² 

Crop Breeding and Applied Biotechnology
26(3): e544126313, 2026
Brazilian Society of Plant Breeding.
Printed in Brazil
<http://dx.doi.org/10.1590/1984-70332026v26n3a43>



Abstract: This study evaluated statistical dimensionality reduction techniques, Principal Component Analysis (PCA), Sparse PCA (SPCA), Independent PCA (IPCA), deep learning based on autoencoders (AE), and their hybrid combinations, for subpopulation identification in genomic data. High-dimensional single nucleotide polymorphisms (SNP) datasets pose computational challenges and may hinder clustering performance. We compared methods based on clustering agreement with known subpopulation labels using *Oryza sativa* data (413 genotypes; 44, 100 SNPs). PCA and SPCA exhibited strong performance (Adjusted Rand Index [ARI] = 0.715), with PCA achieving the lowest computational time (4.89 s). The standalone autoencoder performed poorly, likely due to the high-dimensional ($p \gg n$) setting. Among the hybrid approaches, IPCA-AE achieved the highest agreement (ARI = 0.758), demonstrating that combining dimensionality reduction with nonlinear representation learning improves population structure identification.

Keywords: Deep learning, Population structure, Clustering, Dimensionality reduction, Rice genomics

INTRODUCTION

Rice (*Oryza sativa* L.) accounts for more than half of the global population (Fukagawa and Ziska 2019). However, the increasing food demand has intensified the need to identify highly productive cultivars (Kubo and Purevdorj 2004, Ashfaq et al. 2012, Bharamappanavara et al. 2020). This species is broadly divided into two subspecies, Indica and Japonica, differing in their geographic distribution, agronomic traits, and genetic backgrounds (Lee et al. 2003, Datta and Datta 2006, Cordero-Lara 2020). Within these groups, molecular marker analyses have revealed additional subpopulations, including Indica and AUS, Temperate Japonica, Tropical Japonica, and Aromatic (Garris et al. 2005, Zhao et al. 2010, Huang et al. 2012). These subpopulations are associated with distinct domestication histories and patterns of genetic differentiation (Lu et al. 2009, Breseghello and Coelho 2013).

Assessing the genetic variability among rice genotypes is essential for breeding programs, as it supports diversity studies, conservation strategies, and the identification of superior genotypes (Akinwale et al. 2011, Vanniarajan et al.

✉ **Corresponding author:**

E-mail: camila.azevedo@ufv.br

Scientific Editor:

Luiz Antônio dos Santos Dias 

Received: 4 December 2025

Accepted: 19 July 2026

Published: 07 August 2026

¹ Universidade Federal de Viçosa (UFV), Departamento de Estatística, Avenida Ph Rolfs, s/n, 36570-000, Viçosa, MG, Brazil

² UFV, Departamento de Biologia Geral, Avenida Ph Rolfs, s/n, 36570-000, Viçosa, MG, Brazil

2012, Ranjith et al. 2018, Swarup et al. 2021). Molecular markers provide high-resolution information on genetic variation and are widely used to characterize population structures and genetic relationships (Zhao et al. 2011, Zhang et al. 2019, Gu et al. 2022). However, genomic datasets typically contain tens of thousands of single nucleotide polymorphisms (SNPs), resulting in high-dimensional data that can hinder clustering performance and increase computational costs.

Dimensionality-reduction techniques are commonly used to address these challenges (Tang et al. 2014, Lee et al. 2016, Lasalvia et al. 2022). Classical approaches, such as Principal Component Analysis (PCA), efficiently summarize genomic variation but rely on linear transformations, which may not fully capture complex relationships in genomic data, particularly under nonlinear linkage disequilibrium patterns (Smith 2020). Extensions, such as Sparse PCA (SPCA) and Independent PCA (IPCA), incorporate sparsity and statistical independence, respectively, improving interpretability or feature extraction while maintaining a linear framework.

In contrast, nonlinear methods, such as autoencoders (AE), can learn complex representations by mapping high-dimensional inputs into lower-dimensional latent spaces (Lewis 2016, Yuan et al. 2023). However, their performance may be limited in genomic applications characterized by a high-dimensional setting ($p \gg n$), where the number of variables far exceeds the number of observations. In such scenarios, deep neural networks are prone to overparameterization, which can lead to unstable training and reduced generalization ability (Kiarashinejad et al. 2020).

Hybrid strategies that combine statistical dimensionality reduction and deep learning may help mitigate these limitations. By reducing the dimensionality of the input space before applying nonlinear models, these approaches can improve computational efficiency and stabilize model training, while preserving relevant biological signals. Despite their potential, systematic evaluation of such hybrid methods for population structure identification in genomic datasets remains limited.

Therefore, this study evaluated the performance of PCA, SPCA, IPCA, autoencoders, and hybrid approaches (PCA-AE, SPCA-AE, and IPCA-AE) in identifying subpopulation structures in *Oryza sativa* genomic data. The methods were compared based on their ability to recover known subpopulation labels and their computational efficiency, providing insights into the advantages of integrating linear and nonlinear dimensionality-reduction techniques in high-dimensional genomic analyses.

MATERIAL AND METHODS

Genomic data

The study used Genomic data from the Oryza SNP Project and the Oryza Map Alignment Project (Ammiraju et al. 2006, Zhao et al. 2011), which are publicly available at <http://www.ricediversity.org/>. The dataset comprised 413 *Oryza sativa* genotypes originating from 82 countries. According to Zhao et al. (2011), six subpopulations are present: Indica (IND), AUS, temperate Japonica (TEJ), tropical Japonica (TRJ), aromatics, and an admixed group (ADMIX). In total, 44,100 single nucleotide polymorphisms (SNPs) were initially considered. Markers with a Minor Allele Frequency (MAF) of $< 5\%$ and a call rate of $< 70\%$ were removed, resulting in 36,901 markers. The final genotype matrix X ($413 \times 36,901$) encoded allelic dosages of 0, 1, or 2.

This configuration represents a typical high-dimensional setting ($p \gg n$), where the number of variables greatly exceeds the number of observations, which has important implications for model stability and learning capacity.

Definition of initial groups (D0)

An initial unsupervised clustering analysis was conducted to obtain an exploratory baseline partition, denoted as D0. Hierarchical clustering using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) was applied to the standardized marker matrix based on the Euclidean distances among genotypes.

The optimal number of clusters (K) was determined using the Average Silhouette Width (ASW) criterion (Rousseeuw 1987), which evaluates clustering quality by quantifying within-cluster cohesion and between-cluster separation (Batool and Hennig 2021). The final clustering configuration was obtained by cutting the dendrogram at a K value that maximized the ASW.

For observation i belonging to cluster A_k , let $a(i)$ denote the average Euclidean distance between i and all other observations in the same cluster. Let $b(i)$ denote the minimum average distance between i and observations in any other cluster $A_r \neq A_k$. The silhouette value is defined as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

and the Average Silhouette Width was computed as the mean of $s(i)$ across all observations.

In this study, DO was used exclusively as an exploratory baseline and the biological ground truth was not considered. The primary evaluation of clustering performance was based on the agreement between the predicted clusters and known rice subpopulation labels.

Statistical Methods for Pattern Recognition

Dimensionality reduction methods - PCA, SPCA, IPCA, and an AE - were applied to the marker incidence matrix to extract lower-dimensional representations. The PCA, SPCA, and IPCA generate linear combinations of the original variables, whereas the autoencoder learns nonlinear transformations. In all the cases, the number of extracted components or latent features was substantially lower than the number of markers. These representations were also used as inputs for the hybrid approaches (PCA-AE, SPCA-AE, and IPCA-AE).

PCA (Hotelling 1957, Kendall 1957) was performed using singular value decomposition of the marker matrix $X = UDV'$, where D is a diagonal matrix with elements $d_j = \sqrt{\lambda_j}$, with λ_j being the j -th non-zero eigenvalue of $X'X$ or XX' . The matrix V contains the eigenvectors of $X'X$, and U contains the eigenvectors of XX' associated with λ . The j th principal component is defined as $Z_j = u_j d_j$, where u_j is the j th loading vector. The proportion of total variance explained by the j -th component is given by $\frac{\lambda_j}{\sum_{i=1}^n \lambda_i}$. The first component captured most of the total variation, enabling dimensionality reduction while preserving orthogonality.

SPCA (Shen and Huang 2008) introduces sparsity into the loading vectors using Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani 1996) penalization, producing components with many zero loadings, and improving interpretability by selecting a subset of relevant variables.

Independent Component Analysis (ICA) (Jutten and Hérault 1991, Comon 1994) decomposes data into statistically independent components using the FastICA algorithm (Hyvärinen 1998). Unlike PCA, ICA does not rank components based on explained variance.

IPCA (Yao et al. 2012) combines PCA and ICA. First, PCA was applied to the marker matrix X to reduce dimensionality while retaining the directions of maximum variance. The covariance matrix of V is decomposed as: $V = P\Lambda^{-1/2}P'$, where P contains the eigenvectors and Λ is the diagonal matrix of eigenvalues. A subset of the first r principal components is selected to form the reduced matrix: $J = P_r \Lambda_r^{-1/2}$. To achieve statistical independence, ICA was applied to J using the FastICA algorithm, yielding a transformation matrix R such that the columns of VJ are statistically independent. Independent principal components were then calculated as: $Z = XS$, where $S = VJ/R$. By separating dimensionality reduction (PCA) from independence extraction (ICA), IPCA produces components that are both informative and statistically independent, while reducing the computational cost of directly applying ICA to high-dimensional data (Costa et al. 2020).

An AE (Lewis 2016) is a neural network composed of an encoder and decoder. The encoder maps the input vector x into a lower-dimensional latent representation through a nonlinear transformation. The encoder is defined as $h(x) = f(Wx + b_h)$, where x is the input vector, W is the weight matrix, b_h is the bias vector of the hidden layer, and $f(\cdot)$ where denotes the activation function. The decoder reconstructs the original input from a latent representation. The reconstructed vector $\hat{x}$ is defined as $\hat{x} = g(W_d h(x) + b_g)$, W_d is the decoder weight matrix, b_g is the decoder bias, and $g(\cdot)$ is the output activation function.

The AE is implemented as a fully connected encoder–decoder architecture. Two hidden layers were considered with candidate configurations of 512–256 and 256–128 neurons. The bottleneck dimensions are tuned to 32, 64, and 128 units. ReLU activation functions were used in the hidden layers, and linear activation was adopted in the bottleneck and output layers.

Model training was performed using the Adam optimizer with a learning rate 5×10^{-4} and the Huber loss function (Huber 1964), defined as

$$L_{\delta}(y, \hat{y}) = \begin{cases} \frac{1}{2} (y - \hat{y})^2, & \text{if } |y - \hat{y}| \leq \delta \\ \delta(|y - \hat{y}| - \frac{1}{2}\delta), & \text{otherwise} \end{cases}$$

The Huber loss was selected because of its robustness to outliers and improved numerical stability compared to the Mean Squared Error (MSE), which is particularly relevant for high-dimensional genomic data.

Additional hyperparameters tuned in the inner loop of the nested cross-validation included L2 regularization (10^{-7} and 10^{-6}) and dropout rates (0 and 0.05). Training was performed for up to 200 epochs with a batch size of 32. Early stopping (patience = 25 epochs) and learning-rate reduction on the plateau were used to prevent overfitting and improve convergence.

Given the high-dimensional setting ($p \gg n$), where the number of markers greatly exceeds the number of genotypes, the autoencoder may be prone to overparameterization. Therefore, dimensionality reduction methods have also been used as preprocessing steps in hybrid approaches to improve the stability and learning efficiency.

Hybrid approaches (PCA-AE, SPCA-AE, and IPCA-AE) were implemented by first reducing the dimensionality of the marker matrix and then using the resulting components as inputs to the autoencoder. This strategy reduces input dimensionality and facilitates model training.

After dimensionality reduction, clustering was performed using the UPGMA hierarchical method based on Euclidean distances. The optimal number of clusters in the training set was determined using the Average Silhouette Width (ASW) criterion, and the resulting number was then applied to the test set.

Measures for comparing pattern recognition methods

The performances of the evaluated approaches were assessed using a two-stage nested cross-validation procedure. The outer loop consisted of five folds and was used to obtain unbiased estimates of the clustering performance, whereas the inner loop comprised three folds and was used for hyperparameter tuning and model selection. This nested structure reduces the risk of optimistic bias, which may arise when model selection and performance evaluations are performed on the same data partition (Wilimitis and Walsh 2023).

For each outer fold, approximately 80% of the genotypes were used for training, and the remaining 20% were used for testing. Partitioning preserved the distribution of the known rice subpopulations, accounting for the unequal number of genotypes per group.

The clustering performance was evaluated by comparing the predicted clusters with known rice subpopulation labels. The agreement between the predicted clusters and reference labels was quantified using three external validation metrics: Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and purity.

ARI measures the similarity between two partitions while correcting for chance agreement, with values ranging from -1 to 1 , where higher values indicate stronger agreement. NMI quantifies the amount of shared information between the predicted and reference labels, reflecting how well clustering captures the underlying population structure. Purity measures the proportion of observations within each cluster that belong to the dominant class, indicating cluster homogeneity.

In addition, confusion matrices were constructed to visualize the distribution of genotypes from each known subpopulation across the predicted clusters. The compositions of the predicted clusters relative to the true subpopulations are also summarized.

Computational resources

All analyses were performed using the R software environment (R Core Team 2024), available at <http://cran.r-project.org>. Dimensionality reduction methods were implemented using the mixOmics package (Rohart et al. 2017), which provides functions for PCA, SPCA, and IPCA. Autoencoder models were implemented using the keras interface in TensorFlow (Allaire and Chollet 2023).

Cluster analysis and silhouette-based selection of the number of groups were performed using the cluster package (Maechler et al. 2023). Cross-validation partitions were generated using the caret package (Kuhn 2008). Clustering agreement metrics were computed using the mclust package (Scrucca et al. 2023) for ARI, and the aricode package (Chiquet et al. 2023) for NMI. Data manipulation and result aggregation were performed using the dplyr and readr packages.

The code used in this study is publicly available at <<https://github.com/licaeufv/autoencoder>>.

RESULTS AND DISCUSSION

Preliminary clustering analysis (D0)

A preliminary clustering analysis based on the original genomic marker matrix (D0) identified three major groups among the 413 rice genotypes (Figure 1). Cluster 1 contained the largest proportion of genotypes (62.71%, 259 genotypes), whereas clusters 2 and 3 comprised 23.24% (96 genotypes) and 14.04% (58 genotypes), respectively. These groups

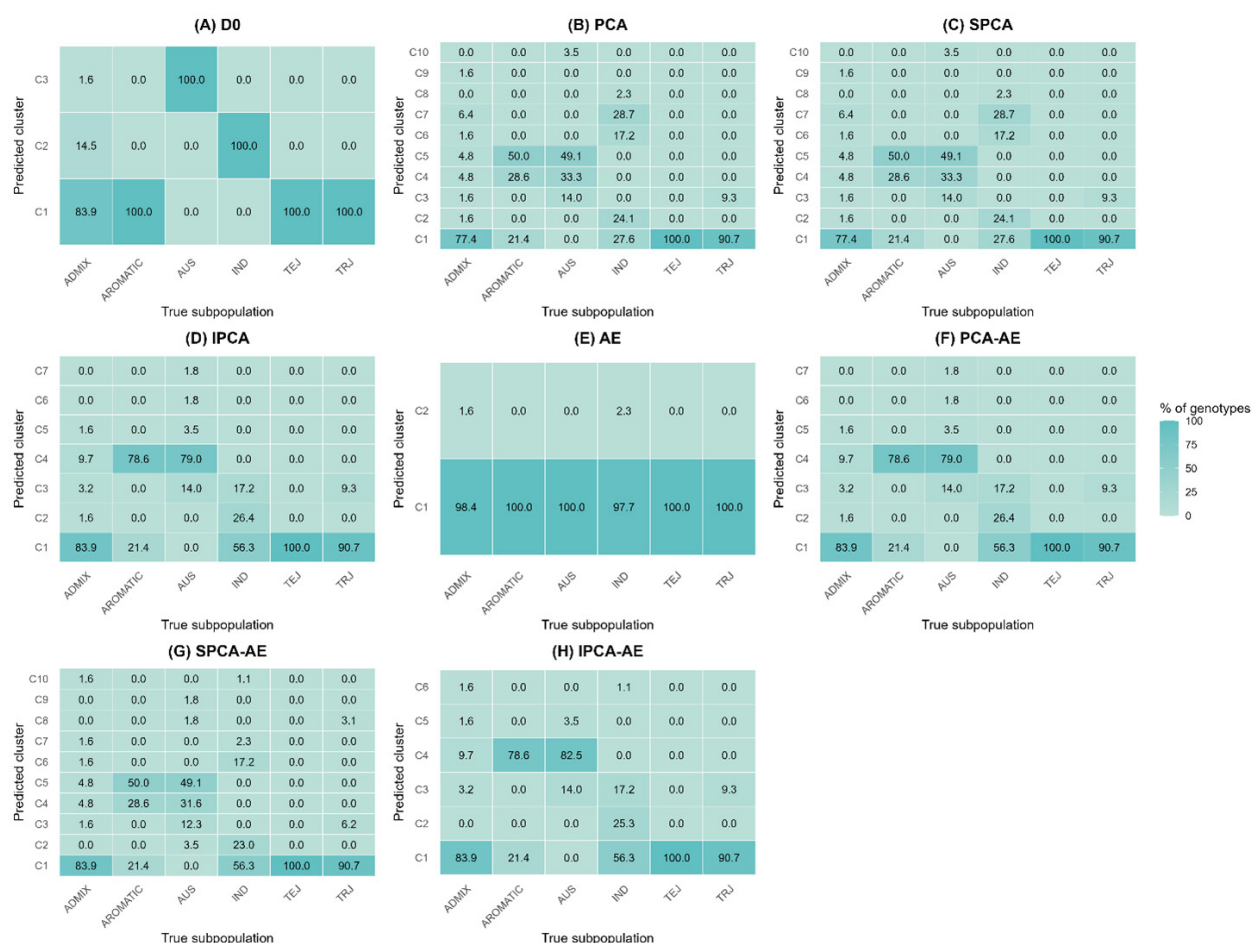


Figure 1. Confusion matrices showing the percentage of genotypes assigned to each predicted cluster relative to the known rice subpopulation labels for each evaluated method. Columns represent the true subpopulations (ADMIX, Aromatic, AUS, IND, Temperate Japonica – TEJ, and Tropical Japonica – TRJ), and rows represent the predicted clusters (e.g., C1 and C2). Values correspond to the percentage of genotypes from each true subpopulation assigned to each predicted cluster, averaged across the five outer folds of the nested cross-validation procedure. Darker colors indicate higher percentages. Panels correspond to the following methods: (A) baseline clustering (D0); (B) Principal Component Analysis (PCA); (C) Sparse Principal Component Analysis (SPCA); (D) Independent Principal Component Analysis (IPCA); (E) Autoencoder (AE); (F) PCA-AE; (G) SPCA-AE; and (H) IPCA-AE.

Table 1. Performance of dimensionality reduction and hybrid approaches for identifying rice subpopulation structure. Results are presented as the mean $\pm$ standard deviation across the five outer folds of the nested cross-validation procedure. Computational time (seconds) corresponds to the average total model training and evaluation time per fold

Method	ARI (True labels)	NMI (True labels)	Purity	Time (s)
D0	0.398 $\pm$ 0.004	0.498 $\pm$ 0.047	0.586 $\pm$ 0.007	11.59 $\pm$ 1.24
PCA	0.715 $\pm$ 0.069	0.783 $\pm$ 0.121	0.886 $\pm$ 0.134	4.89 $\pm$ 0.12
SPCA	0.715 $\pm$ 0.069	0.783 $\pm$ 0.074	0.886 $\pm$ 0.004	53.78 $\pm$ 18.33
IPCA	0.665 $\pm$ 0.133	0.743 $\pm$ 0.047	0.826 $\pm$ 0.007	15.76 $\pm$ 6.76
AE	0.394 $\pm$ 0.264	0.499 $\pm$ 0.047	0.593 $\pm$ 0.010	348.32 $\pm$ 100.59
PCA-AE	0.647 $\pm$ 0.173	0.738 $\pm$ 0.264	0.793 $\pm$ 0.098	17.50 $\pm$ 9.12
SPCA-AE	0.627 $\pm$ 0.113	0.736 $\pm$ 0.296	0.816 $\pm$ 0.223	52.59 $\pm$ 11.36
IPCA-AE	0.758 $\pm$ 0.031	0.808 $\pm$ 0.094	0.881 $\pm$ 0.106	20.45 $\pm$ 6.72

PCA, Principal Component Analysis, SPCA: Sparse Principal Component Analysis, IPCA: Independent Principal Component Analysis, AE: Autoencoder.

broadly reflected the main structure of rice population diversity reported in previous studies using the same dataset (Ammiraju et al. 2006, Zhao et al. 2011).

Although it captured part of the population structure, baseline clustering showed limited agreement with known subpopulation labels (ARI = 0.398) (Table 1). This result indicates that clustering based directly on high-dimensional marker data does not fully recover the underlying genetic structure, supporting the need for dimensionality reduction and flexible modeling approaches.

Comparison between dimensionality reduction and hybrid approaches

The performances of the evaluated methods are summarized in Table 1. Overall, dimensionality reduction substantially improved clustering performance compared with the baseline analysis (D0).

Among the linear methods, PCA and SPCA achieved identical performances (ARI = 0.715; purity = 0.886), indicating that the introduction of sparsity did not improve clustering performance on this dataset. This suggests that most markers contributed to the underlying population structure, thereby reducing the benefits of variable selection (James et al. 2013).

IPCA showed a slightly lower clustering agreement (ARI = 0.665) but still substantially outperformed the baseline. These results suggest that the main population structure is largely captured by dominant variance directions and that enforcing statistical independence does not necessarily improve clustering performance in this context.

The standalone autoencoder exhibited a lower clustering performance (ARI = 0.394). This result is consistent with the high-dimensional setting ($p \gg n$), in which the number of markers greatly exceeds the number of genotypes. In such scenarios, deep neural networks are prone to overparameterization, which can lead to unstable training and reduced generalization ability (Srivastava et al. 2014, Kiarashinejad et al. 2020).

Performance of hybrid methods

Hybrid approaches that combine dimensionality reduction and deep learning improved performance compared with the standalone autoencoder. Among them, IPCA-AE achieved the highest clustering agreement (ARI = 0.758, NMI = 0.808, purity = 0.881; Table 1).

This improvement is attributed to the combination of dimensionality reduction and nonlinear representation learning. By reducing the dimensionality of the marker matrix, IPCA improves the ratio of sample size to the number of variables, mitigating overparameterization and facilitating model training. In addition, the use of independent components helps retain informative and less redundant signals, which are further refined by an autoencoder (Babjac et al. 2022).

The PCA-AE and SPCA-AE also improved over the standalone autoencoder but did not surpass the best linear methods. These results indicated that dimensionality reduction stabilizes neural network learning, although nonlinear models do not always outperform simple linear approaches for high-dimensional genomic data.

Tradeoff between clustering performance and computational cost

Figure 2 illustrates the tradeoff between clustering performance and computational time. PCA achieved high clustering agreement with the lowest computational cost (4.89 s), highlighting its efficiency for large genomic datasets.

The IPCA-AE achieved the highest ARI while maintaining a moderate computational cost, representing a favorable balance between performance and efficiency. In contrast, the standalone autoencoder required a substantially higher computational time without improving clustering performance.

Interpretation of population structure

The confusion matrices presented in Figure 1 illustrate the correspondence between the predicted clusters obtained using each method and the known rice subpopulation labels. Overall, both the dimensionality reduction and hybrid approaches were able to recover the main population structure, although important differences in cluster composition were observed among the methods.

Linear dimensionality reduction methods, particularly PCA and SPCA, showed highly consistent clustering patterns and clearly separated most major rice subpopulations. As shown in Figures 1 and 3, both methods achieved complete or near-complete separation of the Aromatic, AUS, IND, TEJ, and TRJ groups, whereas misclassification was primarily concentrated in the ADMIX subpopulation. In both cases, ADMIX individuals were distributed across ADMIX-, TEJ-, and TRJ-dominated clusters, indicating that the admixed group did not form well-defined clusters in the genomic space.

The IPCA method also recovered the main population structure, achieving complete separation of aromatics, AUS, IND, and TRJ but showing a partial overlap between TEJ and TRJ. Approximately 19.79% of the TEJ genotypes were assigned to the TRJ-dominated cluster, a result consistent with that of previous studies reporting relatively low genetic divergence and substantial allele sharing between Temperate and Tropical Japonica groups (Zhao et al. 2011).

The standalone autoencoder exhibited the least consistent clustering pattern. As illustrated in Figures 1 and 3, this method resulted in substantial mixing among subpopulations, particularly involving ADMIX, TEJ, and TRJ, and reduced the separation of aromatics, AUS, and IND. For example, only 64.29% of the aromatic, 78.95% of the AUS, and 80.46% of the IND genotypes were assigned to their dominant clusters, indicating a limited ability to recover the known population structure.

The composition of the predicted clusters relative to the true subpopulations is summarized in Figure 3, which provides a complementary perspective by explicitly showing the proportion of genotypes assigned to each cluster. This representation highlights cluster homogeneity and allows for a direct comparison of how well each method captures the underlying population structure. In this context, baseline clustering (D0) recovered the broad structure of the dataset, but merged TEJ, TRJ, and aromatics into a single group and distributed ADMIX individuals across multiple clusters, indicating limited resolution.

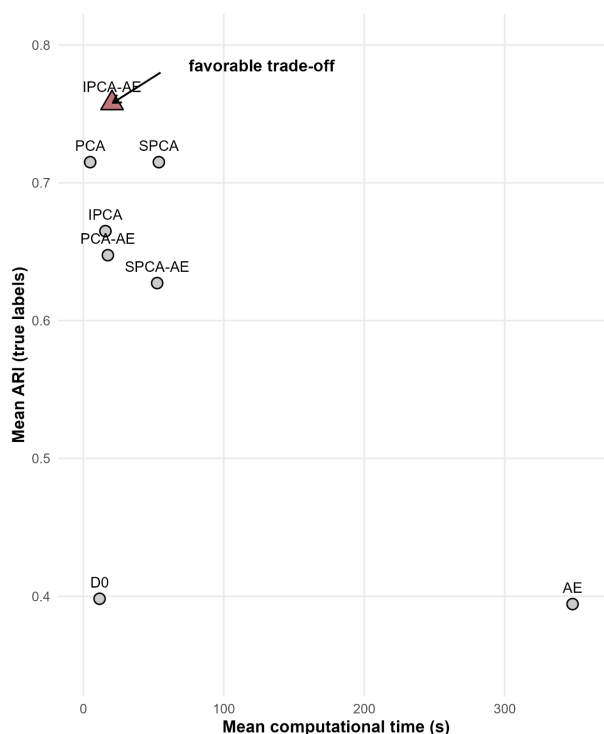


Figure 2. Tradeoff between clustering performance and computational cost for the evaluated methods. The x-axis represents the mean computational time (seconds), while the y-axis shows the mean Adjusted Rand Index (ARI), both computed across the five outer folds of the nested cross-validation procedure. Each point corresponds to one method. The IPCA-AE method showed the most favorable balance between clustering agreement and computational efficiency among the evaluated approaches. PCA, Principal Component Analysis; SPCA, Sparse Principal Component Analysis; IPCA, Independent Principal Component Analysis; AE, Autoencoder.

Among the hybrid approaches, IPCA-AE produced a clustering pattern that was the most consistent with the known population structure. This method achieved clear separation of aromatic, AUS, IND, TEJ, and TRJ genotypes, whereas ADMIX genotypes remained distributed across the ADMIX-, TEJ-, and TRJ-related clusters. This pattern is biologically plausible because admixed individuals are expected to occupy intermediate positions in the genetic space rather than forming a single compact cluster.

The PCA-AE and SPCA-AE approaches also improved upon the standalone autoencoder but did not outperform the best linear methods in all cases. PCA-AE preserved the complete separation of aromatics, AUS, and IND but showed partial overlap between TEJ and TRJ. The SPCA-AE exhibited a similar behavior, with a small degree of mixing involving aromatics and TRJ. These results indicate that combining dimensionality reduction with nonlinear feature learning can improve clustering performance compared with using an autoencoder alone, although the magnitude of this improvement depends on the input representation.

Overall, these results demonstrated that dimensionality reduction plays a key role in identifying population structures in high-dimensional genomic datasets. PCA and SPCA provided strong and stable recovery of the major subpopulations.

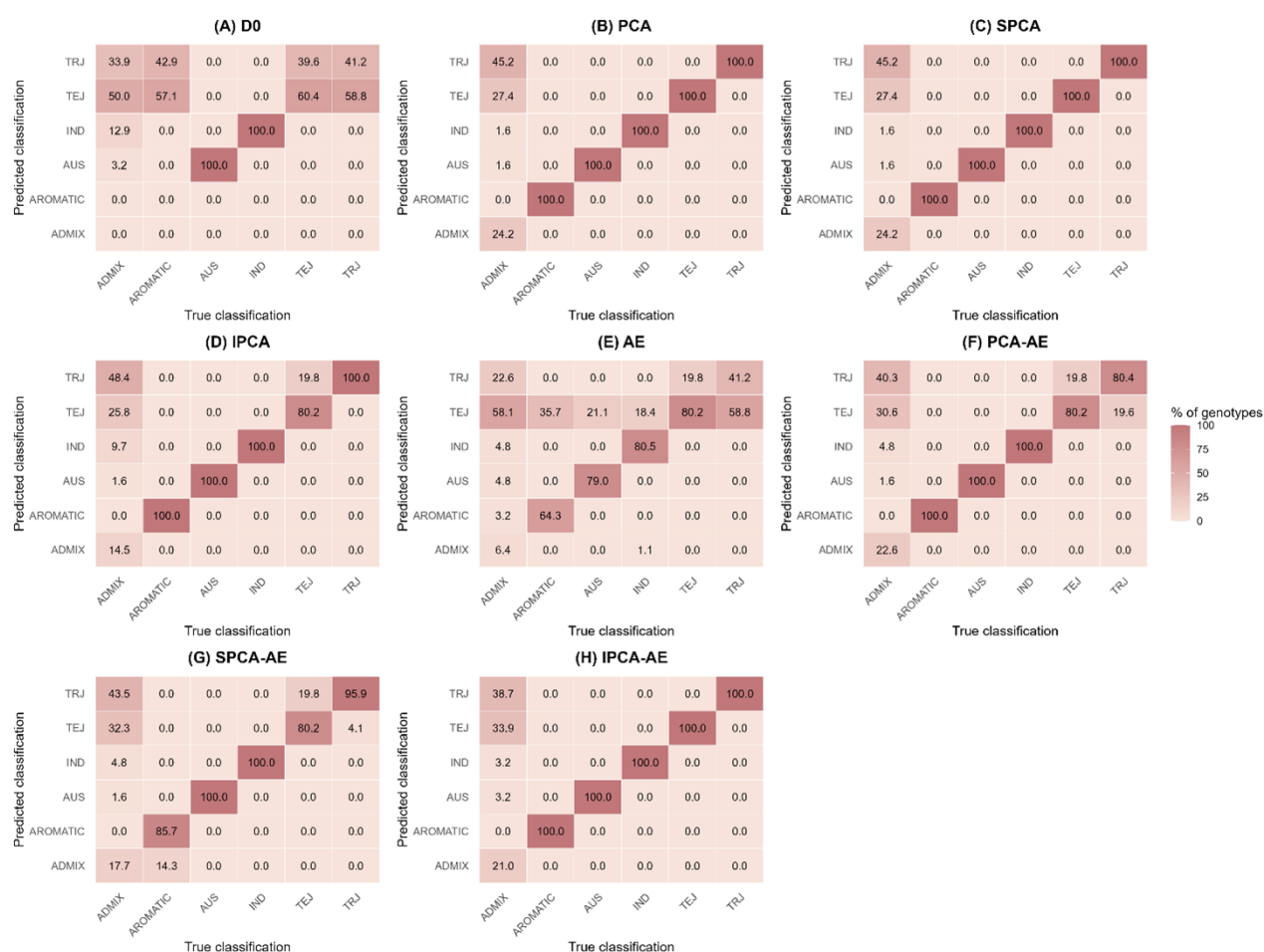


Figure 3. Composition of predicted clusters relative to the known rice subpopulations across the evaluated methods. Bars represent the proportion of genotypes from each true subpopulation assigned to each predicted cluster. Panels correspond to the following methods: (A) baseline clustering (D0); (B) Principal Component Analysis (PCA); (C) Sparse Principal Component Analysis (SPCA); (D) Independent Principal Component Analysis (IPCA); (E) Autoencoder (AE); (F) PCA-AE; (G) SPCA-AE; and (H) IPCA-AE. True subpopulations: ADMIX, Aromatic, AUS, IND, Temperate Japonica (TEJ), and Tropical Japonica (TRJ). This representation highlights the composition and homogeneity of predicted clusters, allowing direct comparison of how each method captures the underlying population structure.

In contrast, IPCA-AE achieved the highest overall agreement with known labels, particularly when considering the expected dispersion of the ADMIX group.

CONCLUSION

This study evaluated linear, nonlinear, and hybrid approaches for identifying population structures using high-dimensional rice genomic data. Dimensionality reduction substantially improved the clustering performance compared with direct clustering on the original SNP marker matrix.

Among the evaluated methods, the hybrid IPCA-AE approach achieved a favorable balance between clustering agreement and computational efficiency. By combining independent component extraction with nonlinear representation learning, the IPCA-AE captured relevant genetic patterns while mitigating the challenges associated with high-dimensional data.

In contrast, the standalone autoencoder showed limited performance and higher computational cost, highlighting the importance of combining dimensionality reduction with deep learning in settings characterized by high dimensionality ($p \gg n$).

Overall, these findings demonstrate the potential of hybrid strategies integrating classical statistical methods with deep learning for population structure analysis. However, the results were specific to the *Oryza sativa* dataset analyzed here, and further studies using different species and genomic datasets are needed to assess the general applicability of these approaches.

ACKNOWLEDGEMENTS

We thank the Foundation for Research Support of the State of Minas Gerais (FAPEMIG) for funding under projects APQ-02111-18 and APQ-01545-24 and the Brazilian National Council for Scientific and Technological Development (CNPq) for funding under projects 309856/2023–0, 310755/2023–9, and 408833/2023–8.

CREDIT STATEMENT

CFA; JAC: Conceptualization. CFA; JAC; MN; ACCN; CDC: Methodology. CFA; JAC: Data curation. CFA; JAC: Formal analysis; Investigation; Software; Visualization; Writing – original draft. CFA; JAC; MN; ACCN; CDC: Writing – review & editing.

DATA AVAILABILITY

The datasets generated and/or analyzed during the current research are available from the corresponding author upon reasonable request.

REFERENCES

- Akinwale MG, Gregorio G, Nwilele F, Akinyele BO, Ogunbayo SA and Odiyi AC (2011) Heritability and correlation coefficient analysis for yield and its components in rice (*Oryza sativa* L.). *African Journal of Plant Science* 5: 207–212.
- Allaire JJ and Chollet F (2023) keras: R Interface to ‘Keras’. R package version 2.13.0. **CRAN**. Available at <https://CRAN.R-project.org/package=keras>. Accessed on April 12, 2024.
- Ammiraju JSS, Luo M, Goicoechea JL, Wang W, Kudrna D, Mueller C, Talag J, Kim H, Sisneros NB, Blackmon B, Fang E, Tomkins JB, Brar D, MacKill D, McCouch S, Kurata N, Lambert G, Galbraith DW, Arumuganathan K, Rao K, Walling JG, Gill N, Yu Y, SanMiguel P, Soderlund C, Jackson S and Wing RA (2006) The *Oryza* bacterial artificial chromosome library resource: Construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus *Oryza*. *Genome Research* 16: 140–147.
- Ashfaq M, Khan AS, Ullah Khan SH and Ahmad R (2012) Association of various morphological traits with yield and genetic divergence in rice (*Oryza sativa*). *International Journal of Agriculture and Biology* 14: 55–62.
- Babjic A, Royalty T, Steen AD and Emrich SJ (2022) A comparison of dimensionality reduction methods for large biological data. **Proceedings of the 13th ACM international conference on bioinformatics, computational biology and health informatics (BCB '22)**. ACM, New York, p. 1–7.
- Batool F and Hennig C (2021) Clustering with the average silhouette width. *Computational Statistics and Data Analysis* 158: 107190.
- Bharamappanavara M, Siddaiah AM, Ponnuvel S, Ramappa L, Patil B, Appaiah M, Maganti SM, Sundaram RM, Shankarappa SK, Tuti MD, Banugu S, Parmar B, Rathod S, Barbadikar KM, Kota S, Subbarao LV, Mondal TK and Channappa G (2020) Mapping QTL hotspots associated with weed competitive traits in backcross population

- derived from *Oryza sativa* L. and *O. glaberrima* Steud. **Scientific Reports** **10**: 22103.
- Breseghello F and Coelho ASG (2013) Traditional and modern plant breeding methods with examples in rice (*Oryza sativa* L.). **Journal of Agricultural and Food Chemistry** **61**: 8277-8286.
- Chiquet J, Rigail G and Sundqvist M (2023) Aricode: Efficient computations of standard clustering comparison measures. R package version 1.0.3. Available at <<https://github.com/jchiquet/aricode>>. Accessed on April 13, 2024.
- Comon P (1994) Independent component analysis, a new concept? **Signal Processing** **36**: 287-314.
- Cordero-Lara KI (2020) Temperate japonica rice (*Oryza sativa* L.) breeding: History, present and future challenges. **Chilean Journal of Agricultural Research** **80**: 303-314.
- Costa JA, Azevedo CF, Nascimento M, Silva FF, Resende MDV and Nascimento ACC (2020) Genomic prediction with the additive-dominant model by dimensionality reduction methods. **Pesquisa Agropecuária Brasileira** **55**: e01713.
- Datta K and Datta SK (2006) Indica rice (*Oryza sativa*, BR29 and IR64). **Methods in Molecular Biology** **343**: 201-212.
- Fukagawa NK and Ziska LH (2019) Rice: Importance for global nutrition. **Journal of Nutritional Science and Vitaminology** **65**: S2-S3.
- Garris AJ, Tai TH, Coburn J, Kresovich S and McCouch S (2005) Genetic structure and diversity in *Oryza sativa* L. **Genetics** **169**: 1631-1638.
- Gu H, Liang S and Zhao J (2022) Novel sequencing and genomic technologies revolutionized rice genomic study and breeding. **Agronomy** **12**: 218.
- Hotelling H (1957) The relations of the newer multivariate statistical methods to factor analysis. **British Journal of Statistical Psychology** **10**: 69-79.
- Huang X, Kurata N, Wei X, Wang ZX, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W, Guo Y, Lu Y, Zhou C, Fan D, Weng Q, Zhu C, Huang T, Zhang L, Wang Y, Feng L, Furuumi H, Kubo T, Miyabayashi T, Yuan X, Xu Q, Dong G, Zhan Q, Li C, Fujiyama A, Toyoda A, Lu T, Feng Q, Qian Q, Li J and Han B (2012) A map of rice genome variation reveals the origin of cultivated rice. **Nature** **490**: 497-501.
- Huber PJ (1964) Robust estimation of a location parameter. **Annals of Mathematical Statistics** **35**: 73-101.
- Hyvärinen A (1998) New approximations of differential entropy for independent component analysis and projection pursuit. **Advances in Neural Information Processing Systems** **10**: 273-279.
- James G, Witten D, Hastie T and Tibshirani R (2013). **An introduction to statistical learning: With applications in R**. Springer, New York, 607p.
- Jutten C and Hérault J (1991) Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. **Signal Processing** **24**: 1-10.
- Kendall MGA (1957) **Course in multivariate analysis**. Griffin, London, 185p.
- Kiarashinejad Y, Abdollahramezani S and Adibi A (2020) Deep learning approach based on dimensionality reduction for designing electromagnetic nanostructures. **npj Computational Materials** **6**: 12.
- Kubo M and Purevdorj M (2004) The future of rice production and consumption. **Journal of Food Distribution Research** **35**: 128-142.
- Kuhn M (2008) Building predictive models in R using the caret package. **Journal of Statistical Software** **28**: 1-26.
- Lasalvia M, Capozzi V and Perna G (2022) A comparison of PCA-LDA and PLS-DA techniques for classification of vibrational spectra. **Applied Sciences** **12**: 5345.
- Lee KS, Choi WY, Ko JC, Kim TS and Gregorio GB (2003) Salinity tolerance of japonica and indica rice (*Oryza sativa* L.) at the seedling stage. **Planta** **216**: 1043-1046.
- Lee LC, Liong CY, Osman K and Jemain AA (2016) Comparison of several variants of principal component analysis (PCA) on forensic analysis of paper based on IR spectrum. **AIP Conference Proceedings** **1750**: 060012.
- Lewis ND (2016) **Deep learning made easy with R.A gentle introduction for data science**. Available at <<https://openlibrary.org/search?q=Deep+learning+made+easy+with+R.A+gentle+introduction+for+data+science>>. Accessed on March 18, 2026.
- Lu BR, Cai X and Xin J (2009) Efficient indica and japonica rice identification based on the InDel molecular method: Its implication in rice breeding and evolutionary research. **Progress in Natural Science** **19**: 1241-1252.
- Maechler M, Rousseeuw P, Struyf A, Hubert M and Hornik K (2023) Cluster: Cluster analysis basics and extensions. R package version 2.1.6. Available at <<https://CRAN.R-project.org/package=cluster>>. Accessed on 13 June, 2026.
- R Core Team (2024) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. Available at <<https://www.R-project.org/>>. Accessed on June 13, 2026.
- Ranjith P, Sahu S, Dash SK, Bastia DN and Pradhan BD (2018) Genetic diversity studies in Rice (*Oryza sativa* L.). **Journal of Pharmacognosy and Phytochemistry** **7**: 2529-2531.
- Rohart F, Gautier B, Singh A and Cao KAL (2017) mixOmics: An R package for 'omics feature selection and multiple data integration. **PLOS Computational Biology** **13**: e1005752.
- Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics** **20**: 53-65.
- Scrucca L, Fraley C, Murphy TB and Raftery AE (2023) **Model-based clustering, classification, and density estimation using mclust in R**. Chapman & Hall/CRC, London, 242p.
- Shen H and Huang JZ (2008) Sparse principal component analysis via regularized low rank matrix approximation. **Journal of Multivariate**

Analysis 99: 1015-1034.

- Smith RD (2020) The nonlinear structure of linkage disequilibrium. **Theoretical Population Biology 134**: 160-170.
- Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R (2014) Dropout: A simple way to prevent neural networks from overfitting. **Journal of Machine Learning Research 15**: 1929-1958.
- Swarup S, Cargill EJ, Crosby K, Flagel L, Kniskern J and Glenn KC (2021) Genetic diversity is indispensable for plant breeding to improve crops. **Crop Science 61**: 839-852.
- Tang L, Peng S, Bi Y, Shan P and Hu X (2014) A new method combining LDA and PLS for dimension reduction. **PLOS One 9**: e96944.
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society Series B 58**: 267-288.
- Vanniarajan C, Vinod KK and Pereira A (2012) Molecular evaluation of genetic diversity and association studies in rice (*Oryza sativa* L.). **Journal of Genetics 91**: 9-19.
- Wilimitis D and Walsh CG (2023) Practical considerations and applied examples of cross-validation for model development and evaluation in health care: Tutorial. **JMIR AI 2**: e49023.
- Yao F, Coquery J and Cao KAL (2012) Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. **BMC Bioinformatics 13**: 24.
- Yuan M, Hoskens H, Goovaerts S, Herrick N, Shriver MD, Walsh S and Claes P (2023) Hybrid autoencoder with orthogonal latent space for robust population structure inference. **Scientific Reports 13**: 2612.
- Zhang P, Zhong K, Zhong Z and Tong H (2019) Genome-wide association study of important agronomic traits within a core collection of rice (*Oryza sativa* L.). **BMC Plant Biology 19**: 259.
- Zhao K, Tung CW, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J, McClung AM, Bustamante CD and McCouch SR (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. **Nature Communications 2**: 467.
- Zhao K, Wright M, Kimball J, Eizenga G, McClung A, Kovach M, Tyagi W, Ali ML, Tung CW, Reynolds A, Bustamante CD and McCouch SR (2010) Genomic diversity and introgression in *O. sativa* reveal the impact of domestication and breeding on the rice genome. **PLOS One 5**: e10780.



This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.