PHOTOSYNTHETICA
International Journal for Photosynthesis Research

DOI 10.32615/ps.2025.031 PHOTOSYNTHETICA 63 (3): 296, 2025

LETTER TO THE EDITOR

Avoiding leakage when classifying drought stress from OJIP fluorescence – comment on Xia *et al*. (2025)

E.P. PASTORE

University of Calabria, Department of Biology, Ecology and Earth Science, 87036 Arcavacata di Rende (CS), Italy

Xia and colleagues report a support vector machine (SVM) classifier built on OJIP chlorophyll *a* fluorescence, aiming to distinguish four rice drought stress durations after dimensionality reduction with independent component analysis (ICA) (Xia *et al*. 2025). They obtain macro-averaged accuracies approaching 0.89. The paper addresses an important applied problem, but several details in the design and evaluation make the reported performance difficult to interpret as out-of-sample predictive ability.

First, the data-splitting scheme likely violates the principle of independence. The Methods state that the "ChlF of the same rice leaf under different drought stress levels (D0, D1, D2, and D4)… was measured," and that "80% of the samples were randomly selected as the training dataset and the remaining 20% as the testing dataset". If measurements from the same plant (or the same leaf at successive stress durations) are randomly divided between training and test, the classifier can exploit plant-specific idiosyncrasies rather than stress-related signal. This constitutes information leakage and leads to optimistic estimates of accuracy (Roberts *et al.* 2017, Kapoor and Narayanan 2023). When the experimental unit is the plant, the split should be performed by plant identity (grouped/blocked cross-validation), or by time block if stress duration induces temporal dependence (Roberts *et al.* 2017). A brief reanalysis with group-aware splits would clarify how much performance depends on within-plant redundancy.

Second, several preprocessing steps appear to have been applied before the split. The manuscript notes that datasets "need to be normalized according to Z-score standardization before analysis", and that feature extraction methods (PCA, SVD, Isomap, LLE, ICA) were applied to the OJIP induction. If the standardization parameters and the dimensionality-reduction mappings were estimated using all samples and only then evaluated on a held-out subset, the test set has effectively informed the training pipeline. Best practice is to fit every preprocessing step on the training portion only and apply the fitted transform to the test portion within each resampling loop; where model tuning is involved, error should be estimated with nested cross-validation to avoid selection bias (Varma and Simon 2006, Cawley and Talbot 2010). Even with fixed SVM hyperparameters, leakage at the preprocessing stage is sufficient to inflate performance (Kapoor and Narayanan 2023).

Third, because the four classes are imbalanced (counts reported as 1,335; 1,093; 1,322; and 1,146 samples), macro-averaging is appropriate, but readers would benefit from class-wise support and a confusion matrix built under the leakage-free splits suggested above. Reporting the number of plants represented in each fold would make the generalization target explicit. Finally, since the stated application is drought-stress screening, the most relevant deployment question is how the model performs on plants not seen during training, possibly under different batches or measurement days. A small, truly independent plant-level hold-out or a repeated, group-aware cross-validation would answer that directly.

References

Cawley G.C., Talbot N.L.C.: On over-fitting in model selection and subsequent selection bias in performance evaluation. – J. Mach. Learn. Res. **11**: 2079-2107, 2010.

Kapoor S., Narayanan A.: Leakage and the reproducibility crisis in machine-learning-based science. – Patterns **4**: 100804, 2023.

Roberts D.R., Bahn V., Ciuti S. *et al.*: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. – Ecography **40**: 913-929, 2017.

Varma S., Simon R.: Bias in error estimation when using cross-validation for model selection. – BMC Bioinformatics **7**: 91, 2006.

Xia Q., Tang H., Tan J.L. *et al.*: Determination of rice (*Oryza sativa* L.) drought stress levels based on chlorophyll *a* fluorescence through independent component analysis. – Photosynthetica **63**: 73-80, 2025.

Received 6 October 2025
Accepted 7 October 2025
Published online 7 October 2025

e-mail: emmanuel.pastore17@gmail.com

Conflict of interest: The author declares no conflict of interest.