ScienceDirect

Chemometrics and Intelligent Laboratory Systems

Volume 274, 15 July 2026, 105742

Engineering an interpretable chemometric pipeline for sugarcane wine fermentation using synthetic spectra and explainable ensembles

Ebenezer Aquisman Asare [a b], Dickson Abdul-Wahab [c], Elsie Effah Kaufmann [d], Rafeah Wahi [e], Zainab Ngaini [e], Archibold Buah-Kwofie [f], Abdul Rashid Dickson [g]

Show more

https://doi.org/10.1016/j.chemolab.2026.105742 ↗

Highlights

- Synthetic spectra framework achieves $R^2 > 0.99$ for fermentation monitoring models.
- SHAP reveals key spectral regions: O-H (3740 cm^{-1}) and C-H (2857 cm^{-1}) stretching.
- Tree-based models outperform deep learning for spectroscopic regression tasks.
- Bayesian bootstrap provides reliable uncertainty during fermentation transitions.
- Domain transfer remains critical challenge: all models fail on real-world data.

Abstract

Background

Real-time spectroscopic monitoring of sugarcane wine is becoming increasingly important for quality and safety control. However, it remains less studied than established fermentations such as beer or grape wine, especially from a chemometric perspective.

Objective

To develop a mechanistically grounded synthetic spectral generation framework that enables interpretable machine learning for fermentation monitoring without extensive experimental calibration datasets.

Methods

We integrated kinetic fermentation modelling based on extended Monod-inhibition equations with realistic spectral simulation incorporating multi-scale noise artefacts. Nine machine learning architectures (PLS, Random Forest, Gradient Boosting, DNN, CNN, LSTM, ResNet, Transformer, and Stacked Ensemble) were evaluated using SHAP-based explainability analysis and Bayesian bootstrap uncertainty quantification.

Results

Tree-based models achieved exceptional performance on purely synthetic validation data (R^2 up to 0.997, RMSE ≈ 1.1 g/L), but *all* architectures collapsed when evaluated under simulated real-world conditions that introduced unmodelled matrix variability and instrument artefacts (R^2 from −0.01 to −1.88). The simplest PLS model degraded the least but still failed to reach acceptable predictive accuracy, indicating a fundamental gap between the synthetic training distribution and realistic deployment scenarios.

Significance

These results show that even careful mechanistic, noise-aware synthetic spectra cannot guarantee successful domain transfer by themselves. Synthetic data remain valuable for model prototyping, architecture screening, and interpretability analysis, but must be complemented by targeted experimental calibration and synthetic-to-real adaptation strategies. Exploiting this limitation is critical for chemometric practice, as it reframes synthetic pipelines from “replacements” to “augmentation tools” for real fermentation monitoring.
