Systems & Control Transactions

Original Research Article
Peer Reviewed Conference Proceeding
ESCAPE 36 - European Symposium on Computer Aided Process Engineering
Sheffield, UK. 21-24 June 2026

PSE PRESS

A Machine Learning Implementation for Fermentation Quality Prediction in Wine Manufacturing

Matthew A.J. Hill and Dimitrios I. Gerogiorgis*

Institute for Materials and Processes (IMP), School of Engineering, University of Edinburgh, Edinburgh
* Corresponding Author: D.Gerogiorgis@ed.ac.uk.

ABSTRACT

Wine consumers are increasingly health- and environmentally conscious. At the same time, white wine and rosé drinkers favour freshness and varietal aromas, which requires low-temperature regimes that extend fermentation time and increase energy demand. Additionally, global warming accelerates grape ripening which increases alcohol level in wine. To reduce cost and alcohol levels while maintaining quality, predictive tools that forecast how fermentation conditions impact fermentation time, and primary and secondary metabolite concentrations, can provide practical benefits to wineries by expediting oenological decisions-making and in turn reducing energy demand. Additionally, literature highlights static models in smart manufacturing suffer from performance degradation with data drift. In light of this, we successfully developed and evaluated pipelines for the automated design and training of three ML methods - support vector regression, random forest and artificial neural networks – to predict fermentation outcomes from ten initial features. The dataset employed captures the fermentation of nine commercially available *Saccharomyces cerevisiae* strains, each performed in triplicate, across two synthetic media and four temperature regimes, sampled at nine time points. The targets we focused on included time to complete fermentation, Ethanol, Acetate, given its undesired spoilage effects and ten secondary metabolite concentrations, given their synergistic impact on aroma. We were able to recommend effective pipelines for each task and identified our feature set contained adequate predictive signal for ethanol, acetate and fermentation duration predictions, but inadequate for the multi-output regression of secondary metabolite concentrations, which exhibited target heterogeneity. Ultimately, embedding these pipelines into optimisation frameworks, offers an actionable route to tailoring wine characteristics to evolving consumer preference, while reducing energy demand.

Keywords: alcoholic fermentation, machine learning, artificial neural network, support vector regression, random forest regression, fermentation time, efficacy, secondary metabolite concentration

INTRODUCTION

In 2024 the International Organization of Vine and Wine attributed the six-decade low wine consumption, 214.2 million hectolitres, to high average prices and emerging consumer trends [1]. These trends include prioritizing health, sustainability, and preferring low-alcohol beverages. The wine industry therefore requires solutions to produce environmentally friendly, low-alcoholic and high-quality wines, whilst reducing cost.

This is a particularly significant challenge in white-wine and rosé production as consumers prefer freshness and varietal aromas [2]. This requires low-temperature fermentation as it fosters the development and preservation of volatile compounds [3]. However, low-temperature isothermal programs (12-19 °C) can extend the duration of fermentation and increase the risk of sluggish and stuck fermentations. Given that alcoholic fermentation is exothermic, the process requires wineries to actively cool fermentation vats. This cooling accounts for between 45% and 90% of the total energy consumption in manufacturing. In the case of white wine and rosé, where fermentation is prolonged, the process becomes environmentally unfriendly and costly [4]. To address this challenge Minebois et al. investigated a non-isothermal temperature regime on 9 commercially available *Saccharomyces cerevisiae* strains across two synthetic grape must (SGM) compositions. The non-isothermal

temperature regime, henceforth referred to as the time-varying temperature regime (TVAR), began fermentation at 23°C, to leverage *S. cerevisiae's* higher optimal growth temperature, and upon entry into the stationary phase of fermentation switched to 15°C [2]. The two SGM compositions contained 210 g L^{-1} and 250 g L^{-1} total sugars (glucose and fructose in racemic quantities) and hence-forth will be referred to as MS210 and MS250, respectively. Additionally, the TVAR was compared to three isothermal regimes (12°C, 18°C, 25°C). The published dataset captures the complex interaction between temperature, strain, and fermentation media.

Initial fermentation conditions strongly influence the final composition and therefore quality of wine. These conditions affect the concentrations of primary metabolites such as ethanol, acetate, and glucose, as well as the concentrations of secondary metabolites that define the wines aroma profile. Glucose contributes to sweetness and elevated acetate levels can produce vinegary notes [2]. Therefore, predictive tools that forecast the impact of fermentation conditions on fermentation time, and, primary and secondary metabolite concentrations can accelerate oenological decision-making, whilst ensuring quality and reducing both costs and energy demand. Considering this, we looked to employ three machine learning (ML) methods – support vector regression (SVR), random forest (RF), artificial neural networks (ANN) – to forecast these outcomes, given their proven ability to map complex interactions [5]. However, in recognition that literature adopts static ML models, often with arbitrarily selected architectures and hyperparameters, where performance degrades with data shifts [6]. We therefore developed and evaluated pipelines an actionable workflow for the automated design and training of these methods.

Consumers have become increasingly conscious of adverse health effects and roadside safety, and, as climate change increases alcohol level in wines by accelerating grape ripening, research on reducing alcohol content in wine has intensified [7]. Subsequently, the primary metabolites we chose to predict were ethanol and acetate, considering that bacteria from grapes have been found to oxidize ethanol to acetic acid, a volatile acid that causes wine spoilage [8]. We chose to predict secondary metabolite concentrations for their synergistic effects on wine's aroma profile, a key component of wine quality.

Machine Learning Methods

In our investigation we explored supervised learning methods which aim to find parameters, θ, that minimize the average loss over a finite training set of input-output pairs $\{(x_i, y_i)\}_{i=1}^{N}$. Where, x_i encodes strain identity, synthetic must strength (MS210 or MS20), temperature regime and initial metabolites, and y_i denotes our predicted targets.

Support Vector Machines (SVM)

Support Vector Machines (SVM) are large-margin classifiers that separate labeled data by maximizing the margin between classes [9]. For linear SVM, the decision function is $f(x) = \omega^{\top} x + b$, where ω is the weight vector and b is the bias. The parameters (ω, b), are obtained by solving the soft-margin optimization problem. SVR extends this principle to continuous targets by introducing Vapnik's ε-insensitive loss, where an ε-tube is fitted around the regression function [9]. Deviations larger than ε incur a penalty and the optimization problem becomes,

$$min_{\omega,\xi,\xi^*} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \quad (1)$$

s.t. the constraints (Eq. 2, 3, 4),

$$y_i - f(x_i) \leq \varepsilon + \xi_i^* \quad (2)$$

$$f(x_i) - y_i \leq \varepsilon + \xi_i^* \quad (3)$$

$$\xi_i, \xi_i^* \geq 0 \quad (4)$$

where the slack variables, ξ_i and ξ_i^*, measure how far sample i lies outside the ε-tube, C adds a linear penalty to those violations and y_i is the true target value. To capture nonlinearity, SVMs employ the kernel trick, replacing the inner product $\omega^{\top} x$ with a kernel function $K(x, x')$ that behaves like an inner product in some (potentially higher) feature space. A common choice is the radial basis function (RBF) kernel, $K(x, x') = e^{\left(-\gamma\|x - x'\|^2\right)}$, where γ is a scaling parameter that controls the width of the Gaussian kernel function.

Random Forest (RF)

Random Forest (RF) is an ensemble learning method that combines many decision trees using bootstrap aggregation (bagging). Each tree is fit to a bootstrap sample of the training data, and at every split the algorithm considers only a random subset of the input features. This randomization reduces correlation and tends to improve generalization [10]. For regression, the forest prediction for an input vector x is the average of individual tree predictions,

$$\hat{y}(x) = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \quad (5)$$

where T is the number of trees in the forest and $f_t(x)$ is the prediction of t-th tree trained on its own bootstrap sample. Within each tree, splits are chosen to minimize the mean squared error (MSE) of the target values in the child nodes and a cost complexity pruning parameter, α, can be added to penalize larger trees.

Artificial and Recurrent Neural Networks (ANN)

An Artificial Neural Network (ANN) is a parametric

function approximator that learns the mapping f_θ: $\mathbb{R}^d \rightarrow \mathbb{R}^m$, where d and m are the number of input and output dimensions, respectively. Given an input vector $x \in \mathbb{R}^d$ and parameters $\theta = \{\omega, b\}$, each neuron computes a weighted sum $f(x) = \omega^\top x + b$, to which a non-linear activation function is applied, this mathematically replicates a biological neuron [10]. Each neuron then passes the processed signal forward as its output and multiple neurons constructs a layer. If we denote k as a discrete training step, the output of a single neuron can be written as,

$$y(k) = F\left(\sum_{i=0}^{m} w_i\,(k) \bullet x_i\,(k) + b\right) \tag{6}$$

where $F\,(\bullet)$ is the activation function, $x_i\,(k)$ is the i-th input component and $y(k)$ is the output [11]. A common activation function that applies a nonlinear transform is the rectified linear unit function (ReLU), which is applied component-wise,

$$ReLU(z) = \begin{cases} z, z > 0 \\ 0, z \leq 0 \end{cases} \tag{7}$$

where z is the pre-activation function within Eq.6 [8]. A Feed-forward ANN (FFANN) is one in which information flows strictly in one direction from the input layer to the output layer. A common FFANN is a multilayer perceptron (MLP) where all layers are fully connected (dense) [11]. An MLP with two or more hidden layers is referred to as a deep neural network and can learn hierarchical feature representations.

When training, a backpropagation algorithm was employed to minimize the mean squared error between predicted and observed targets, Eq. 8 [12]. Backpropagation applies the chain rule to efficiently compute the gradient of the loss with respect to each weight and bias, Eq. 9. Weights are then updated with the Adam optimizer, an adaptive variant of stochastic gradient descent that uses exponentially decaying, bias-corrected first and second order moment loss gradient estimates, eq. 11.

$$\mathcal{L}(\theta) = \frac{1}{m}\sum_{j=1}^{m}\left(y_j - \hat{y}_j\right)^2 \tag{8}$$

$$\frac{\partial \mathcal{L}}{\partial \omega_{kj}} = \frac{\partial \mathcal{L}}{\partial \sum_i \omega_{kj} a_j + b_k} \bullet \frac{\partial \sum_i \omega_{kj} a_j + b_k}{\partial \omega_{kj}} \tag{9}$$

$$\omega_{kj}{}^{(t+1)} = \omega_{kj}{}^{(t)} - \eta \frac{\hat{m}_{t,kj}}{\sqrt{\hat{v}_{t,kj}} + \epsilon} \tag{10}$$

where $\mathcal{L}(\theta)$ is the MSE loss-function, y_j is the true target value, $\hat{y}_j$ is the models prediction, a_j is the activation output of neuron $j, \frac{\partial \mathcal{L}}{\partial \omega_{kj}}$ is the loss gradient with respect to weight $\omega_{kj}(k), t$ is the current training step, $\hat{m}_{t,kj}$ and $\hat{v}_{t,kj}$ are the bias corrected first and second order moments and η is the learning rate.

METHODOLOGY

Minebois et al.'s dataset contains samples of biomass, primary metabolites, amino-nitrogen species, and volatile compounds at nine time points (T_0- T_8), where T_8 marks the completion of fermentation. Data was collected for nine commercially available *S. cerevisiae* strains across two synthetic must strengths (MS210, MS250), and four temperature regimes. As mentioned above, this included a time-varying temperature regime (TVAR) that switches from 23°C to 15°C at one-third completion of fermentation and 3 isothermal regimes ($12°C, 18°C, 23°C$). We developed supervised ML pipelines to predict fermentation outcomes at T_8 from the initial measurements at T_0.

A compact set of ten variables were identified using literature and spearman's correlations including: strain, temperature regime, MS, cell counts, ethanol, glycerol, organic acids, and NH_4Cl. Categorical features were one-hot encoded, and numerical features were median-imputed and standardized [13]. For the SVR and ANN pipelines, numerical features were also converted to orthogonal components via PCA to reduce collinearity and dimensionality (95% of variance retained). In contrast, PCA was *not* applied to the RF pipeline because tree-based ensembles are inherently robust to collinearity, and this could discard predictive low-variance signals. The models were evaluated on several train-test splits (80:20%, 70:30%, 60:40%, 50:50%) to mimic data limitations and distributional shifts using a grouped splitting method, ensuring that each *Strain-MS-Temperature-regime* combination was contained entirely within either the training or testing set to prevent information leakage.

For the SVR and RF, hyperparameters were tuned on the training set using an exhaustive grid search method with a 5-fold grouped cross-validation (CV) technique. The hyperparameters were selected using a mean MAE fitness across folds. The selected model was refit on the full training set and evaluated on the unseen test set. For the ANN, we replaced the grid search with a genetic algorithm (GA) as a more efficient search of both the architecture and hyperparameter space. The 5-fold grouped CV prevented leakage by keeping all samples from the same combination within a single fold and allowed us to gauge how well the model generalizes to unseen conditions.

Pipeline recommendations were primarily based on mean CV MAE (predictive accuracy) and CV MAE range (robustness) across splits, with the test set used to check for overfitting, and quantified through CV-Test MAE difference (Generalization behaviour). Additionally, deeper analysis was performed on the models out-of-fold (OOF) CV predictions at the 80:20 split, where models were least-data limited.

Results were performed in triplicate, so predictions were averaged within each Strain-MS-Temperature-regime combination. Minebois et al. implicitly define time T_8 as the point at which CO_2 evolution has effectively

ceased and the cumulative weight-loss curve reaches a plateau.

As mentioned, the primary metabolite regressed included concentration ($g\ L^{-1}$) of ethanol and acetic acid. The targets were predicted using separate models to optimize hyperparameters. The secondary metabolite regression included concentrations ($mg\ L^{-1}$) of ethyl acetate, isobutyl alcohol, isoamyl acetate, isoamyl alcohol, ethyl hexanoate, ethyl octanoate, phenylethyl acetate, phenyl ethanol, octanoic acid, and decanoic acid. Given the time constraints we choose predict all secondary metabolite concentrations in one pipeline. However, *scikit learns* SVR function does not natively support multi-output regression, therefore, we employed a wrapper that fits one model per target. Additionally, because some targets were right skewed (e.g. Fig. 1) the pipelines assessed if applying a logarithm transform to the targets improves predictive accuracy.

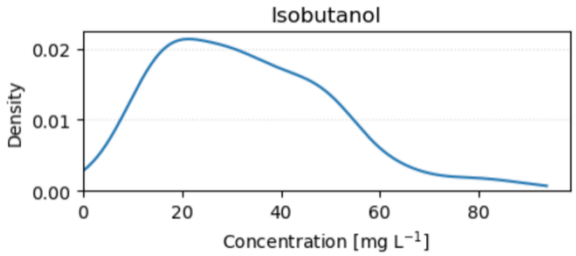

Figure 1: Kernel density estimate of Isobutanol concentrations.

RESULTS AND DISCUSSION

Across train-test splits, the Random Forest pipeline achieved the lowest mean cross-validated (CV) MAE (38.57h) and was the most robust to data partitioning (CV MAE range = 4.09h), outperforming the ANN (CV MAE = 40.59h, CV MAE range = 9.94h) and SVR (CV MAE = 42.5h, CV MAE range = 8.11h) pipelines. Generalization behaviour (CV-Test MAE gap) was also the most stable (-0.30h) indicating that the RF model performances transferred most consistently to the held-out test sets. Although larger training sets typically improve generalization, the RF pipeline achieved its lowest CV MAE at the 50:50 split (36.04h). This likely reflects split-composition effects in a small, heterogeneous, group-split dataset, where the presence/absence of harder regimes can dominate the CV error. Accordingly, we place stronger emphasis on 80:20 pipeline performance, as the split is the least data-limited and best reflects achievable performance given the feature set. The model at the 80:20 split achieved competitive CV metrics with strong test set performance (CV MAE = 39.65 h, test MAE = 32.68 h, both R^2 = 0.91). The model's configuration used 800 trees of unrestricted depth, that sampled 100% of feature at each split, with mild regularisation through minimum leaf size (4) and cost-complexity pruning ($\alpha = 1 \times 10^{-4}$), and chose to log transform targets. The 80:20 RF model, built in mean decrease in impurity (MDI) identified temperature (0.659) as the most important feature, consistent with the temperature-dependent yeast metabolism reported in literature [14].

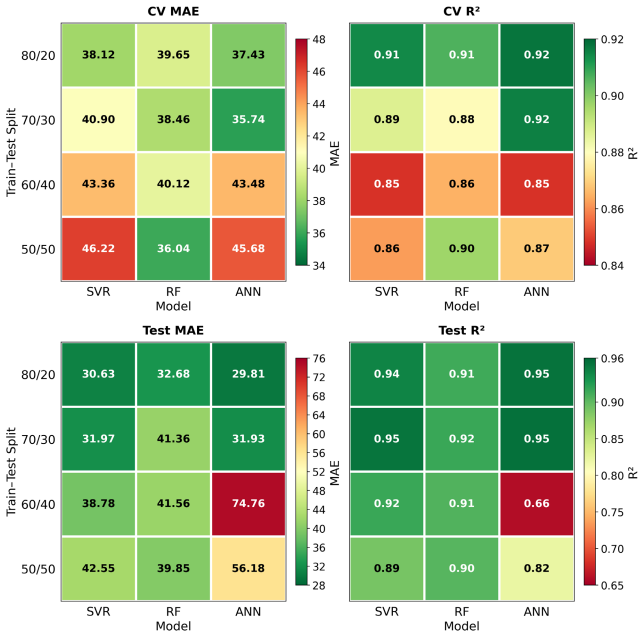

Figure 2: Cross-validated (CV) and test, MAE and R^2 scores for time to complete fermentation (hours) predictions across train-test splits and model pipelines.

Given the ANN's strong performance at higher splits (≥70:30), analysis of out-of-fold CV predictions relative errors (RE) and absolute relative errors (ARE) of both the RF and ANN is appropriate. Models at the 80:20 split were slightly positively biased, with RF achieving a closer to zero mean bias (RE = 0.63% vs +1.76%). RF was also more accurate than the ANN (mean ARE = 7.63% vs 8.60%), with an equivalent standard deviation (SD) (ANN: 0.104, RF: 0.101). Additionally, RF exhibited a heavier tail than the ANN (max. RE = +27.70% vs +22.04%, min. RE = -21.55% vs -18.99%), with a larger strain level bias. The RF model was found to be less biased to isothermal temperate regimes (12℃, 18℃, 25℃) and more biased under TVAR than the ANN (RE: +1.30% vs +1.10%), with the ANN being markedly more biased at 25℃ (RE = +3.98% vs -0.12%). RF exhibited a smaller MS level bias, with the ANN being notably more sensitive to MS210 (RE: +3.18% vs +0.78%), but both displayed poor predictive accuracy at MS210 compared to MS250 (MS210 vs MS250 mean ARE, **ANN: 9.38%** vs 7.95%, RF 9.50% vs **6.07%)**.

This supports recommending RF as the default pipeline, it was most robust to reduced training data and achieved the lowest mean error across splits. The 80:20 RF model was more accurate than the ANN overall, but

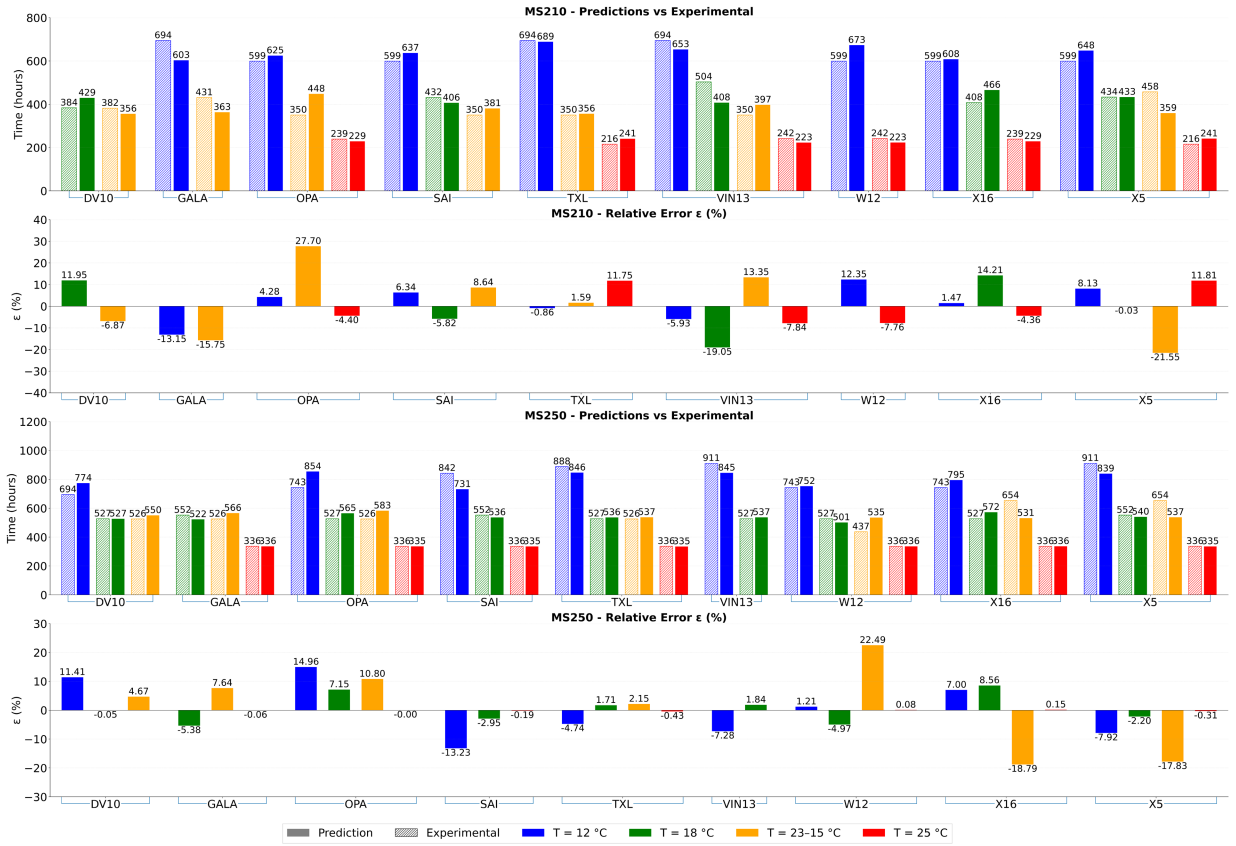

Figure 3. Random Forests 80:20 split out-of-fold cross-validation fermentation duration predictions relative

exhibited less stable tail behaviour. A key limitation is that Minebois et al. does not define an explicit quantitative stopping rule for complete fermentation, so time-to-completion may contain irreducible noise, regardless, the pipeline successfully predicted fermentation duration.

For the ethanol content (%) regression, across train-test splits ethanol was shown to be a learnable relationship on all three pipelines, but performance and stability varied. Over the splits, RF achieved the lowest mean CV MAE (0.563 %), outperforming both the ANN (0.566%) and SVR (0.605 %). However, SVR was the most robust across the train-test splits (smallest CV MAE range = 0.0319%) and exhibited the best generalisation behaviour (smallest CV-Test MAE gap: - 0.0799%).

Figure 3 shows RF achieved lower CV errors compared to SVR, across all splits, with RF being most competitive at the lower splits (≤60:40). The ANN marginally outperformed RF at the high splits (≥70:30) but was less robust overall. The best performing RF model occurred at the 50:50 split, again reflecting impact of the absence of difficult regimes in small, heterogeneous, group-split datasets. At 80:20 split the RF configuration employed 100 trees of restricted depth (max_depth = 10), considered 30% of feature set at each split, and applied regularisation through minimal cost-complexity pruning ($\alpha = 1 \times 10^{-2}$), and chose to log transform targets. The RF model at the 80:20 split's MDI identified must strength (0.695) as the most important feature, consistent with the fermentation's primary reaction. The 80:20 ANN exhibited competitive performance in Figure 3, with a four hidden-layer funnel configuration ($[48, 24, 12, 12]$), moderate drop-out rate (0.163), weak L2 regularisation ($\lambda = 3.36 \times 10^{-5}$) and a relatively high learning rate ($\eta = 3.64 \times 10^{-3}$), with the targets log transformed.

Analysis of out-of-fold CV predictions RE and ARE of both the RF and ANN at the 80:20 split shows that the ANN exhibited a slightly more negative mean bias (RE = -0.37% vs -0.06%), with the RF achieving a greater average predictive accuracy (mean ARE: RF = 3.07%, ANN = 3.56%). RF achieved a slightly tighter error distribution (SD = 0.0429 vs 0.0459), but both models exhibited stronger positive tails (max. RE: RF = +16.24%, ANN = +17.01%, min. RE: RF = -8.67%, ANN = -6.96%). At the temperature level, both models were unbiased under 12°C and TVAR, unpredicted at 18°C and overpredicted at 25°C, with the ANN exhibiting markedly larger under-predictions than RF at 18°C (RE = -1.86% vs -0.94%). However, RF was more accurate in absolute terms at every temperature regime. Strain level biases were small on both models (RE $\sim \pm 2\%$), but on harder to predict strains (OPA, X16) RF was consistently more accurate (lower mean ARE). Finally, both models were less

accurate at MS210 (mean ARE: ANN = 4.45%, RF = 4.00%) than at MS250 (mean ARE: ANN = 2.82%, RF = 2.28%), with RF being more accurate at both must strengths, with negligible MS level bias.

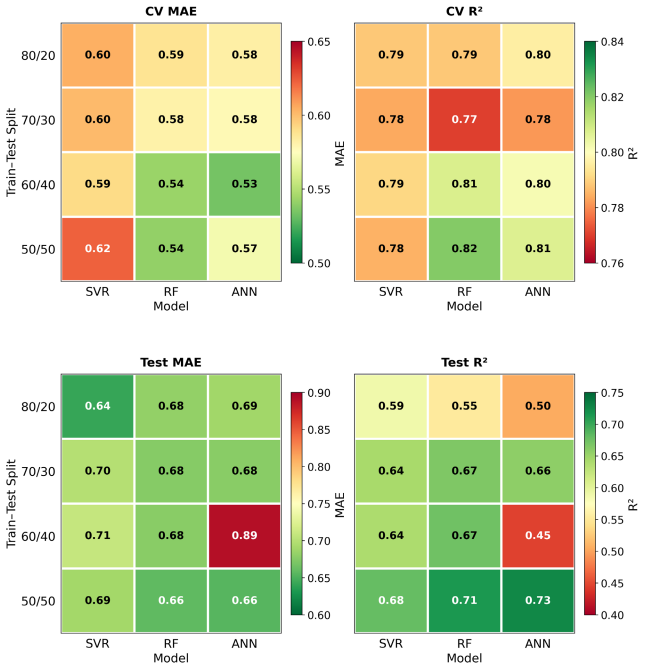

Figure 4: CV and test, MAE and R^2 scores for ethanol (%) predictions across train-test splits and model pipelines

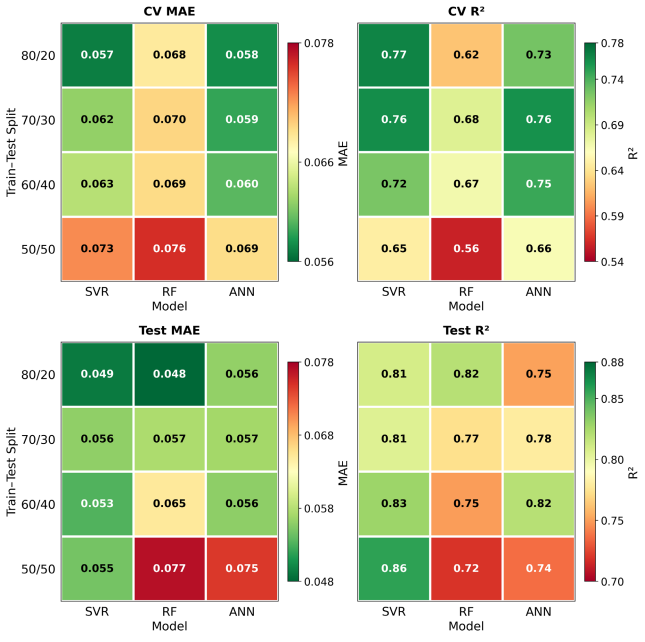

Figure 5: CV and test MAE and R^2 scores for acetate conc. (gL^{-1}) predictions across train-test splits and model pipelines.

Overall, the evidence supports recommending RF as the default pipeline because, although, SVR was the most robust to reduced training data, RF achieved the lowest mean error across splits and under the least data limited regime (80:20), RF achieved greater predictive accuracy, a tighter error distribution, and a smaller temperature and must strength dependency. However, the ANN produced fewer outlier (beyond $IQR \times 1.5$) predictions. A methodological refinement is to model ethanol yield (final – initial) instead of predicting final ethanol with initial ethanol included as a feature, which conflates baseline ethanol with yield.

For the acetate concentration task, across splits the ANN pipeline achieved the lowest mean CV MAE (0.0615 gL^{-1}) and the smallest CV-Test MAE gap (0.0007 gL^{-1}), therefore indicating stable generalization. The RF pipeline was the most robust to splits (smallest CV MAE range: 0.0075 gL^{-1}), however, achieves the worst predictive accuracy (mean CV MAE: 0.0708 gL^{-1}) and explained variance. The SVR pipeline achieved a close second in predictive accuracy across splits (mean CV MAE: 0.0638 gL^{-1}), whilst being the least robust (CV MAE range: 0.0155 gL^{-1}) and exhibiting the largest CV-Test MAE gap (0.0105 gL^{-1}), such that generalization was much stronger, indicating training **sampling noise** (test set was easier).

The SVR configuration at the 80:20 split used an **RBF kernel** with **C=100, ϵ=0.05$, γ=0.003**, and **no target transformation. This corresponds to a smooth nonlinear mapping with relatively strong penalization of training errors. The ANN at the 80:20 split used** a **shallow network** with one hidden layer ([5])**, a** low dropout rate (0.09), moderate L2 regularisation ($\lambda = 2.3 \times 10^{-3}$) and a relatively high learning rate ($\eta = 5.02 \times 10^{-3}$), with a log-transformed target.

Analysis on the 80:20 CV out-of-fold predictions indicate SVR was more positively biased than the ANN (RE = +1.25% vs +0.22%), but both achieved negative median RE (SVR −0.63%, ANN −1.52%), therefore on typical cases the models underpredicted, especially the ANN. The mean ARE highlights that the SVR was more accurate (mean ARE: 7.80% vs 9.15%) than the ANN, and given a typical condition is 52% more accurate (median ARE: 5.42% vs 8.22%). Additionally, the SVR's error distribution was tighter (SD = 0.0999 vs 0.1189). The ANN shows greater strain level bias than SVR, with SVR achieving a greater predictive accuracy for five of the nine strains (OPA, TXL, VIN13, X16, X5). The ANN exhibited particularly weak performance on X16. Both models were less accurate at **MS250** than **MS210**, with higher mean absolute relative error (ANN: 9.73% vs 8.46%, SVR: 8.39% vs 7.09%). Across MS, SVR was more accurate with a small positive bias and the ANN's bias changed sign, **overpredicting** at MS210 (mean RE = +2.56%) and **underpredicting** at MS250 (mean RE = −1.75%).

Finally, the evidence supports recommending the ANN pipeline when mean predictive accuracy, with moderate robustness across splits is preferred. However, the

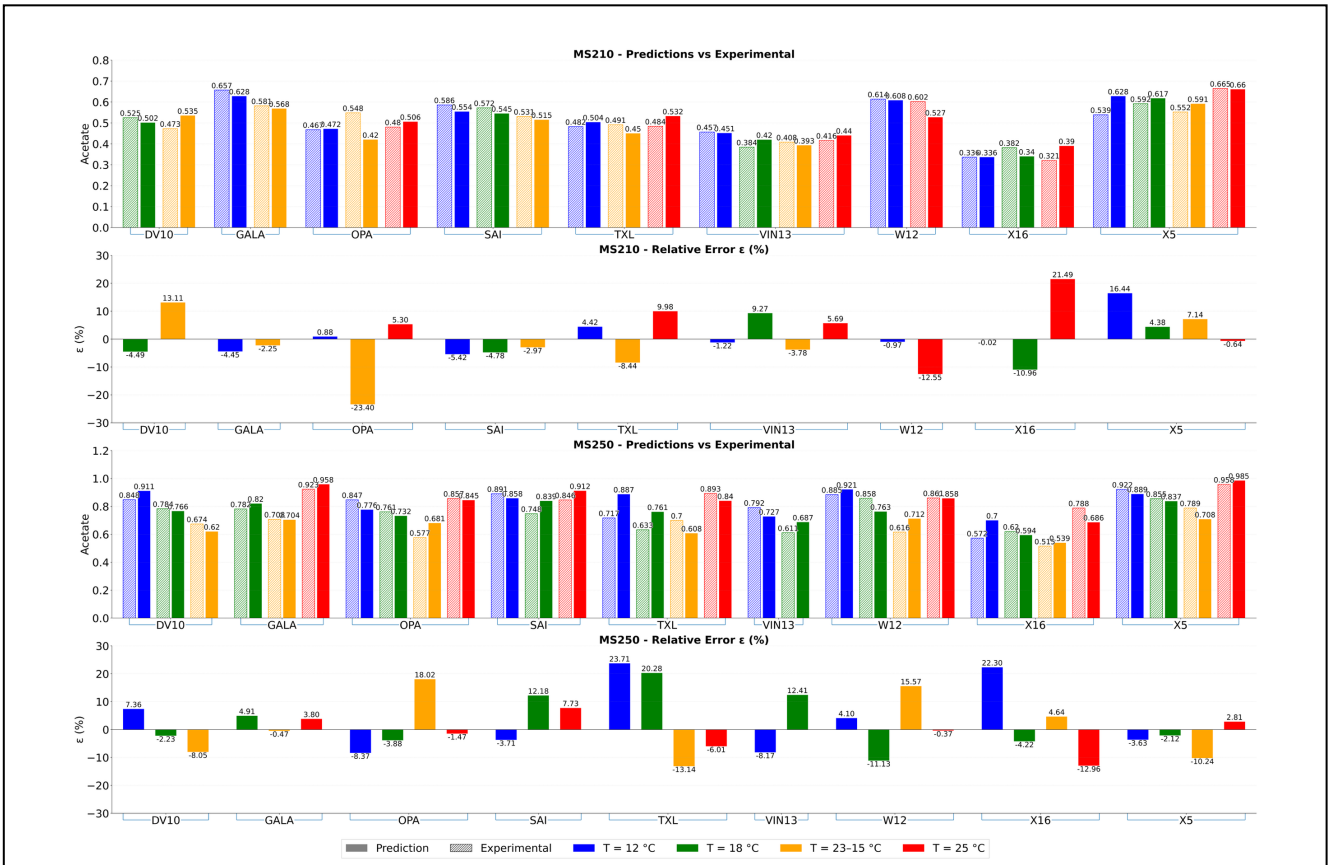

Figure 6. Support vector regression 80:20 split out-of-fold cross-validation acetate concentration (mgL^{-1}) predic–tions relative errors.

ANN's least data limited model (80:20) exhibited a wider error distribution and lower absolute predictive accuracy, with instability on the X16 strain, stronger MS depend-ence, and higher errors at **12°C and 25°C. Therefore, if predictive accuracy under non-data limited regimes is preferred, the SVR pipeline is recommended for its lower mean ARE, tighter error distribution and smaller MS and temperature dependence.**

For the secondary metabolite task, across the train test splits, SVR was both the most accurate and robust pipeline. On every split it achieved the lowest mean CV MAE across the 10 targets (4.97 mgL^{-1}) and highest mean CV R^2 (0.22 mgL^{-1}). However, the ANN pipeline achieved the smallest mean CV-test MAE gap (-0.05 mgL^{-1}), indi-cating the most stable generalization. Regardless SVR was the most competitive on the held-out test set.

Across the splits, the pipeline selected the same SVR configuration, an RBF kernel, $C = 100, \epsilon = 0.1$ and $\gamma = 0.001$, with a log-transform of the target variable, aside from at the 70:30 split, where the pipeline reduced ϵ (0.05). When coupled with the almost identical model configurations, the monotonic improvement of mean CV MAE and R^2 as training fraction increased is consistent with gradual loss of predictive signal under reduced sam-ple size, rather than model instability.

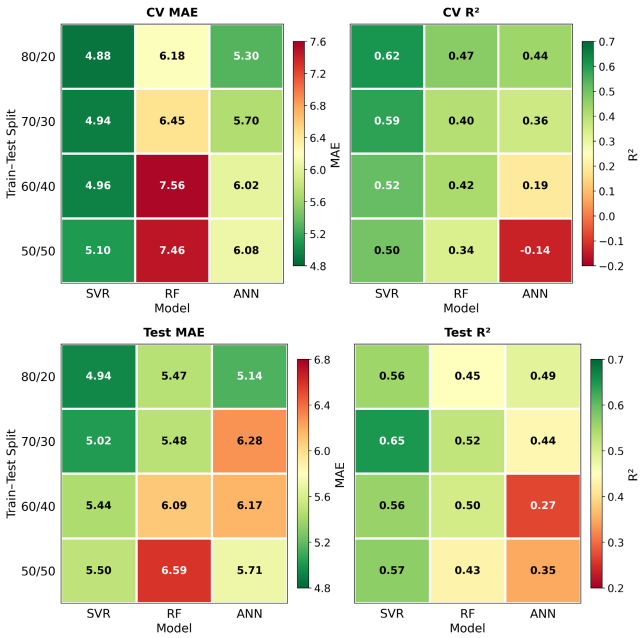

Figure 7: Mean CV and test, MAE and R^2 scores across secondary metabolite (mgL^{-1}) predictions from different train-test splits and model pipelines.

Although the mean results across the target show that the RF pipeline was most accurate and robust to train-test splits, the per target analysis at the 80:20 split reveals large target heterogeneity (Table 1). Specifically higher alcohols and their related esters (**PhenEtoh, Isobutanol, IsoamylAlcohol, PhenEthAcetate, IsoamylAcetate**) exhibit high learnability, whereas several medium-chain fatty acids and ethyl esters (OctanoicAcid, EthylHexanoate, DecanoicAcid, EthAcetate, EthOctanoate) remained difficult to learn.

Table 1: Per-target out-fold CV performance of SVR at the 80:20 train–test split, reported as normalized MAE (nMAE) and $R^{2.}$

Target	nMAE	R^2
PhenEtoh	0.144	0.886
Isobutanol	0.151	0.835
PhenEthAcetate	0.224	0.834
IsoamylAlcohol	0.115	0.814
IsoamylAcetate	0.286	0.759
OctanoicAcid	0.138	0.598
EthylHexanoate	0.237	0.444
DecanoicAcid	0.278	0.436
EthAcetate	0.133	0.343
EthOctanoate	0.271	0.206

The target learnability trend could be partially explained through feature set selection as medium-chain fatty acids and related ethyl esters depend on absent features like lipid/fatty-acid availability, thus limiting predictability.

Of the targets that exhibit high explained variance (R^2) PhenEthAcetate has an elevated nMAE which indicates predictions capture variance with poor practical accuracy. This can be caused through systematic bias and is common in targets with a small range, like PhenEthAcetate, where even modest absolute errors are large fraction of the data's range.

Analysing PhenEtoh, Isobutanol and IsoamylAlcohol out-offold CV predictions for their high learnability and low nMAE, we find that the three targets exhibit near-unbiased central behavior (median RE = -0.08-0.71%), but regime dependent heavy tails. For example, all targets performed relatively poorly on 12**°C** (larger positive bias, higher MARE, and outlier clustering). In terms of absolute accuracy and stability, the model performed best on IsoamylAlcohol (mean ARE = 11.05%, SD = 0.162, max. RE = +66.42%), while Isobutanol and PhenEtoh are less accurate on average (mean ARE = 16.07% and 16.32%) and with longer tails, especially Isobutanol which exhibited the largest upside tail risk (max RE = +157.58%). The mean ARE was higher at MS210 for IsoamylAlcohol and PhenEtoh regression, whereas Isobutanol regression showed a higher mean ARE at MS250. Additionally, the **SAI strain is repeatedly the most difficult to predict across the three targets,** indicating a strain specific failure.

The heterogeneous predictability across the ten metabolites indicates that a single, fixed model configuration is unlikely to be optimal for all targets. This reinforces concerns in the literature that applying static model (non-case specific) configurations limits industrial practicality.

Given the time constraints, we used a task-agnostic feature set and fitted one pipeline across all 10 targets. As a result, the chosen model configurations were likely a compromise that limited target instability on the least learnable targets, rather than refining configurations for highly learnable targets. We would therefore recommend future work expands the feature set to fatty acids/lipids, and include additional assimilable nitrogen sources (Lysine, Proline, Ornithine) and either, fit a pipeline per target or pipelines to groups of targets (higher alcohols, acetate esters, ethyl esters, medium-chain fatty acids).

CONCLUSION

This study successfully demonstrated that machine learning pipelines can map controllable fermentation conditions to fermentation kinetics and endpoint metabolite concentrations in *S. cerevisiae. Moving away from static ML models and motivated by industry constraints, we* developed pipelines that designed and trained three different ML models (SVR, RF, ANN), across four tasks (time to complete fermentation, and ethanol, acetate, and ten secondary metabolite concentrations). We evaluated the pipelines ability to cope with limited data and distributional shifts through evaluation on multiple train–test splits. We employed a *GroupKFold* splitting strategy to prevent information leakage and used five-fold cross-validation to improve generalization performance. Primary assessment criteria included mean MAE, CV MAE range, competitive R^2, and held-out tests were used to assess generalisation stability.

A compact ten variable feature set of operationally accessible predictors was identified: strain, temperature regime, must strength, biomass counts, early metabolites, and a key nitrogen source. This feature set was highly predictive for time, ethanol, and acetate tasks, supporting practical deployment, however, secondary metabolite predictions would likely benefit from additional predictors and per-target pipeline fits.

Performance was task-dependent, for **time-to-completion,** the RF pipeline is recommended as it

achieved the lowest mean CV error and highest robustness to split variation, although the ANN pipeline was competitive when sufficient training data was available (≥70:30). For **ethanol,** again the RF pipeline is the primary recommendation, as it achieved the lowest mean CV MAE with stable generalisation and the RF at the 80:20 split compared to the ANN achievd nearer-zero bias, with weaker temperature and must-strength dependence. For **acetate,** if robustness to split variations with competitive mean accuracy is the priority, the ANN is recommended. However, at the least data limited regime (80:20) the ANN exhibited lower accuracy and a wider error distribution than SVR, with stronger temperature and MS dependence. Therefore, if accuracy under high data availability is the priority, SVR is preferred due to lower mean ARE at the 80:20 split, a tighter error distribution and weaker MS/temperature dependence. For **secondary metabolites,** the per-target SVR models was both most accurate and most robust pipeline, highlighting the value of fitting separate models for targets with heterogeneous signal strength. Higher alcohols and associated acetate esters were highly learnable (R^2 ≥0.75), whereas several ethyl esters and medium-chain fatty acids remained weakly predicted, consistent with missing predictors such as lipids/fatty-acids.

Key methodological refinements for future work include, defining a quantitative fermentation endpoint, calculating and predicting ethanol yield (final–initial) to isolate production, adopting task-specific feature selection, expanding the secondary-metabolite feature space (Lysine, Proline, Ornithine, lipids/fatty acids), and increasing the GA search budget. Finally, we believe the recommended pipelines could be embedded in an optimisation framework to support industry and, to our knowledge, the present work was the first to train an machine learning models on a time-varying temperature regime in wine manufacturing.

ACKNOWLEDGEMENTS

D.I.G acknowledges support from UKRI/EPSRC (RAPID: ReAl-time Process Modelling & Diagnostics: Powering Digital Factories, EP/V028618/1) and a Royal Society International Exchanges grant (2023-25).

REFERENCES

1. International Organisation of Vine and Wine (OIV). State of the world vine and wine sector in 2024. *OIV*, April 2025. Accessed 22 November 2025.
2. Minebois R, Ramírez?Aroca L, Balsa?Canto E, Querol A. Toward less energy?consuming alcoholic fermentations in oenology: a laboratory?scale case study. Food Frontiers 6:2931-2941 (2025). https://doi.org/10.1002/fft2.70103
3. Du Q, Ye D, Zang X, Nan H, Liu Y. Effect of low temperature on the shaping of yeast-derived metabolite compositions during wine fermentation. Food Research International 162:112016 (2022). https://doi.org/10.1016/j.foodres.2022.112016
4. de Castro M, Baptista J, Matos C, Valente A, Briga-Sá A. Energy efficiency in winemaking industry: challenges and opportunities. Science of The Total Environment 930:172383 (2024). https://doi.org/10.1016/j.scitotenv.2024.172383
5. Gupta Y. Selection of important features and predicting wine quality using machine learning techniques. Procedia Computer Science 125:305-312 (2018). https://doi.org/10.1016/j.procs.2017.12.041
6. Sandeep Bharadwaj Mannapur . Understanding data drift and concept drift in machine learning systems. Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol 11:318-330 (2025). https://doi.org/10.32628/cseit25111239
7. Rodrigues AJ, Raimbourg T, Gonzalez R, Morales P. Environmental factors influencing the efficacy of different yeast strains for alcohol level reduction in wine by respiration. LWT 65:1038-1043 (2016). https://doi.org/10.1016/j.lwt.2015.09.046
8. Du Toit W. The enumeration and identification of acetic acid bacteria from south african red wine fermentations. International Journal of Food Microbiology 74:57-64 (2002). https://doi.org/10.1016/s0168-1605(01)00715-2
9. Montesinos López OA, Montesinos López A, Crossa J. Multivariate statistical machine learning methods for genomic prediction. Springer International Publishing (2022). https://doi.org/10.1007/978-3-030-89010-0
10. Biau G, Scornet E. A random forest guided tour. TEST 25:197-227 (2016). https://doi.org/10.1007/s11749-016-0481-7
11. Suzuki K. Artificial neural networks - methodological advances and biomedical applications. InTech (2011). https://doi.org/10.5772/644
12. Hameed AA, Karlik B, Salman MS. Back-propagation algorithm with variable adaptive momentum. Knowledge-Based Systems 114:79-87 (2016). https://doi.org/10.1016/j.knosys.2016.10.001
13. Mu, A.C., n.d. Introduction to Machine Learning with Python.
14. Molina AM, Swiegers JH, Varela C, Pretorius IS, Agosin E. Influence of wine fermentation temperature on the synthesis of yeast-derived volatile aroma compounds. Appl Microbiol Biotechnol 77:675-687 (2007). https://doi.org/10.1007/s00253-007-1194-3