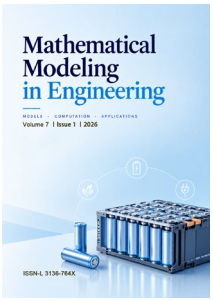
UNIVERSITY OF WEST ATTICA

eISSN: 3136-764X

Mathematical Modeling in Engineering

Vol. 7, No. 1, January, 2026



A TabM-CatBoost Model for Soy Sauce Fermentation Aroma Classification with Small-Sample Feature Fusion

Slobodan Ćorović[a,*], Darko Cvetković[a], Mei-hsien Chi[a]

[a]*Faculty of Chemistry, University of Belgrade, Belgrade, 11158, Serbia*

Received 18 December 2025; Revised 11 March 2026; Accepted 14 March 2026; Published 19 March 2026

Abstract

Soy sauce aroma classification is challenging because labeled fermentation batches are scarce, sensor responses drift across tanks and production cycles, and volatile fingerprints contain many weak but complementary signals. This study developed a TabM-CatBoost model for classifying four aroma states: clean mellow, caramel roasted, ester fruity, and sulfur off-note. Electronic-nose descriptors, GC-MS volatile intensities, physicochemical indices, sensory attributes, and process variables were integrated for 164 fermentation batches, and label efficiency was evaluated under a five-shot protocol. TabM learned compact multi-predictor embeddings from standardized continuous variables, whereas CatBoost combined the embeddings with categorical information through ordered target statistics without validation leakage. The proposed model achieved 95.6% accuracy, 94.9% macro-F1, and 0.973 macro-AUC, while reducing the sulfur off-note false-negative rate from 10.7% for CatBoost to 3.6%. Removing GC-MS features reduced macro-F1 by 5.4 percentage points, and the model achieved 92.7% macro-F1 under 10% instrumental perturbation. Across the label regimes, macro-F1 increased from 89.8% under five-shot training to 94.9% with full labels, while expected calibration error decreased to 0.041. External-season evaluation on 31 later-cycle batches yielded 92.8% macro-F1, and retrospective thresholding retained 24 of 25 off-note alerts while reducing simulated expert review load by 46.2%. These results indicate that compact neural fusion followed by ordered boosting can support label-efficient aroma supervision in high-salt fermented-food production.

Keywords: Soy Sauce Aroma, Small Sample Feature Fusion, TabM, CatBoost, Fermentation Classification

1. Introduction

The aroma of soy sauce results from a long series of koji metabolism, brine fermentation, aging, heating and storage, and it is not due to a single main component [1]. As industrial producers' demands for early aroma classification increase, delaying sensory evaluation results in waste of tanks, labor and raw materials prior to the detection of an off-note during regular quality inspection [2]. Alcohols, aldehydes, esters, acids, pyrazines, phenols and sulfur derivatives are a group of volatile compounds that determine the characteristic odor of fermented soy sauce; however, the amounts of these compounds differ depending on salt tolerance, microbial succession and heat treatment [3]. Electronic noses and gas chromatography-mass spectrometry have enhanced the observation capability of fermentation aroma, but they have also generated heterogeneous measurements with different noise levels and temporal resolutions [4]. Although the pH value and ammonia nitrogen content of the same batch are stable, its volatile components have already been converted into baked or sulfur-containing products. Small production lines face an even more serious problem as only a limited number of labelled aroma batches can be collected for the controlled sensory panels [5]. The absence of a label is not temporary; the expert group needs to set up a regular tasting schedule, perform sample dilution and recovery, obtain multiple confirmations of agreement, etc., before labelling it. Given the abundance of descriptors but a scarcity of labels, direct

* Corresponding author: slobodan.c@chem.bg.ac.rs

https://doi.org/10.68406/mme.2026.vol7iss1nm10:119-131

➢ This is an open access article under the CC BY-NC-SA 4.0 license (https://creativecommons.org/licenses/by/4.0/).

© Authors retain all copyrights.

high-capacity model construction is unreliable unless the representation is constrained by a suitable tabular-learning mechanism [6].

Conventional aroma classifiers include principal component methods, support vector machines, random forests, shallow neural networks, and boosted trees, which can perform well when labeled samples are sufficiently representative [7]. Under low-label fermentation conditions, however, a classifier may learn tank-, season-, or analyst-specific patterns rather than stable aroma characteristics [8]. CatBoost is useful for food-process tables because it handles missing entries and categorical variables through ordered boosting, but a tree-only model does not explicitly compress weak continuous interactions before classification [9]. Recent tabular foundation models have demonstrated that compact inductive biases can improve prediction on small datasets [10], while comparative tabular deep-learning studies emphasize that architecture and validation design must be matched to data scale [11]. Motivated by these findings, this work evaluates whether a compact TabM encoder can stabilize multimodal continuous features before CatBoost integrates the learned embedding with categorical fermentation metadata.

This study proposes a two-stage TabM-CatBoost architecture for four-class soy sauce aroma recognition under a small-sample feature-fusion protocol. The input combines electronic-nose curve summaries, selected GC-MS volatile intensities, physicochemical indices, sensory scores, and process metadata from 164 fermentation batches. The methodological contribution is the separation of two complementary functions: TabM learns a compact representation of weak continuous and ordinal interactions, whereas CatBoost performs leakage-resistant classification after incorporating categorical production variables. The evaluation compares conventional baselines, removes individual feature groups, restricts labels to five-, ten-, and twenty-shot regimes, and perturbs instrumental measurements. Chemical interpretability is examined through descriptor contributions and class-prototype distances. This design targets industrial laboratories in which expert sensory labels are expensive and model outputs must be sufficiently transparent for fermentation supervision.

The remainder of the paper is organized as follows. Section 2 reviews fermented-food aroma recognition and small-sample tabular learning, emphasizing the limited integration of volatile measurements with production metadata. Section 3 describes batch alignment, fold-safe preprocessing, TabM feature fusion, CatBoost classification, and the evaluation criteria. Section 4 reports dataset characteristics, baseline performance, ablation and label-efficiency experiments, perturbation tests, external-season validation, and aroma-indicator interpretation. Section 5 summarizes the main contribution, limitations, and directions for multi-factory validation and uncertainty-aware model development.

2. Related Work

2.1. Aroma Recognition in Fermented Foods

Aroma recognition in fermented foods has gradually moved from descriptive sensory panels toward instrument-assisted classification. Electronic noses provide rapid sensor arrays that respond to composite volatile mixtures, and they have been used to distinguish brewing stages, storage conditions, and quality grades when sample preparation must remain simple [12]. GC-MS offers a more chemically explicit view by measuring volatile compounds that can be associated with alcohol, ester, roasted, acidic, phenolic, and sulfur notes [13]. In soy sauce fermentation, the practical difficulty is that instrumental descriptors and sensory vocabulary do not map one-to-one, because a small increase in one aldehyde may be perceived differently depending on pyrazine, ester, and salt background [14]. Feature engineering methods such as peak-area normalization, odor activity weighting, and derivative sensor summaries have therefore been introduced to make aroma data more comparable across batches [15].

Common classifiers in fermented-food analysis include linear discriminant analysis, partial least-squares discriminant analysis, support vector machines, random forests, and gradient boosting. Their modest computational requirements are attractive for routine quality control. However, when rare off-note classes contain few labeled batches, the fitted boundary can become highly dependent on a small set of handcrafted descriptors. A second limitation is the separation of process metadata from instrumental measurements. Starter culture, raw-material ratio, salt concentration, fermentation time, and heat treatment alter the prior probability of an aroma state, yet many workflows either omit these variables or encode them before cross-validation in a manner that can introduce leakage.

2.2. Small Sample Tabular Learning and Boosted Classification

Small-sample tabular learning must control variance while retaining nonlinear interactions among heterogeneous variables. Wang et al. [16] showed that tabular generative models remain sensitive to dataset scale and heterogeneous feature distributions. In food quality control, machine-learning performance also depends strongly on measurement design, validation practice, and the balance between instrumental and process information [17]. Flavor-data interpretation further requires models that relate multivariate chemical profiles to sensory categories without treating individual compounds as isolated thresholds [18]. These findings motivate a fold-safe architecture that combines compact representation learning with a classifier suited to missing and categorical production data.

Neural tabular models can learn compact representations before classification, but their capacity must remain consistent with the available sample size. Zhou et al. [19] combined machine learning with comprehensive GC-TOF/MS and GC-IMS volatile profiles for fruit-wine classification, illustrating the value of fusing complementary chemical measurements. The remaining gap lies between multimodal aroma classification and practical fermentation supervision: continuous sensor and chemical variables must be integrated with categorical process metadata without leakage. The proposed TabM-CatBoost design addresses this gap by using a lightweight multi-predictor encoder for continuous interactions and ordered boosting for the final category decision.

3. Proposed Methodology

3.1. Dataset Alignment and Small Sample Fusion

The framework represented each fermentation batch as a multimodal tabular object containing electronic-nose curve descriptors, selected GC-MS peak intensities, physicochemical indices, expert sensory scores, and categorical process metadata. Continuous variables were standardized using statistics estimated only from the training fold, whereas categorical process variables were retained for CatBoost. GC-MS values below the detection limit were encoded as bounded low-intensity values accompanied by binary detection flags, and missing sensory descriptors were imputed within the training fold. The fermentation batch, rather than an individual scan, was used as the alignment unit because consensus sensory labels were assigned at batch level. Replicated instrumental measurements were summarized using robust location and dispersion statistics, recovery slopes, and late-response descriptors so that both signal intensity and sensor kinetics were retained.

Figure 1 illustrates how batch sampling, electronic-nose acquisition, GC-MS volatile analysis, sensory annotation, and fold-safe preprocessing are connected before the feature matrix is supplied to the classifier.

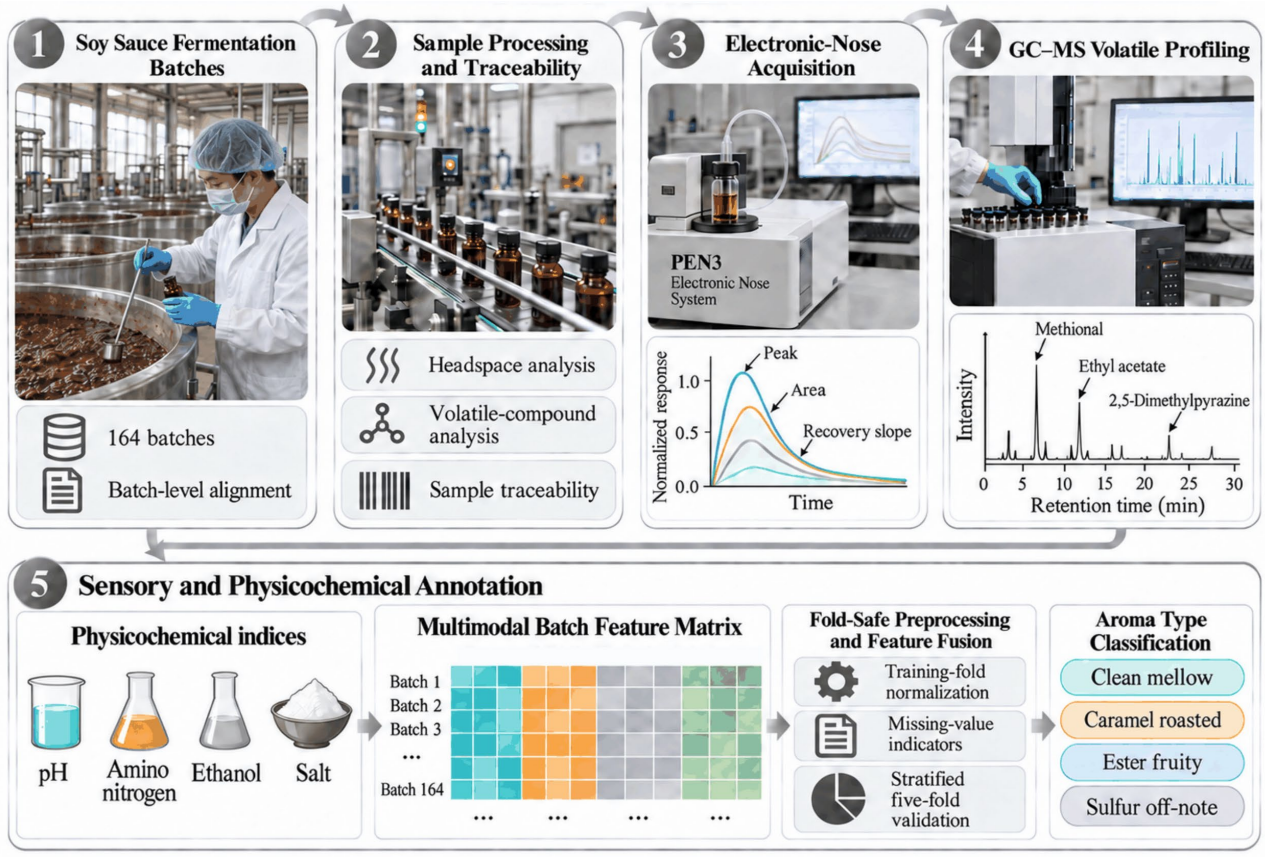


Figure 1. Workflow of soy sauce aroma data acquisition, fold-safe preprocessing, and small-sample feature fusion

To place the heterogeneous measurements in a common batch-level representation, the input assigned to batch i is written as the concatenation of five feature groups:

$$\boldsymbol{x}_i = [\boldsymbol{s}_i, \boldsymbol{v}_i, \boldsymbol{q}_i, \boldsymbol{r}_i, \boldsymbol{c}_i] \tag{1}$$

Here, $\mathbf{x}_i$ denotes the fused descriptor, while $\mathbf{s}_i, \mathbf{v}_i, \mathbf{q}_i, \mathbf{r}_i$, and $\mathbf{c}_i$ represent the electronic-nose summaries, volatile-compound intensities, physicochemical indices, sensory descriptors, and categorical process metadata, respectively.

Because a single neural path can fluctuate under limited labels, the TabM stage combines M lightweight predictors to obtain a more stable embedding:

$$\boldsymbol{z}_i = \frac{1}{M}\sum_{m=1}^{M} f_m(\boldsymbol{x}_i; \boldsymbol{\theta}_m) \tag{2}$$

The resulting vector $\mathbf{z}_i$ is the mean output of the internal predictors f_m, parameterized by $\boldsymbol{\theta}_m$. Averaging preserves cross-channel interactions while reducing the variance passed to the downstream boosting stage.

3.2. TabM CatBoost Classification Architecture

The architecture contains two complementary stages. TabM first encodes continuous and ordinal variables, including sensor, volatile, physicochemical, and sensory descriptors, under dropout, weight decay, and stratified fold validation. Its compact embedding captures weak cross-channel interactions while limiting model capacity. The embedding is then concatenated with selected raw categorical process variables and supplied to CatBoost. Ordered boosting estimates categorical statistics from earlier training permutations, preventing validation labels from influencing the encodings. The target classes are clean mellow, caramel roasted, ester fruity, and sulfur off-note. A large end-to-end neural classifier was avoided because 164 labeled batches cannot support a high-parameter model reliably, whereas a tree-only solution would not explicitly stabilize the weak continuous interactions before classification.

Figure 2 presents the two-stage classifier in which TabM produces compact continuous embeddings and CatBoost combines them with categorical fermentation descriptors to form class probabilities.

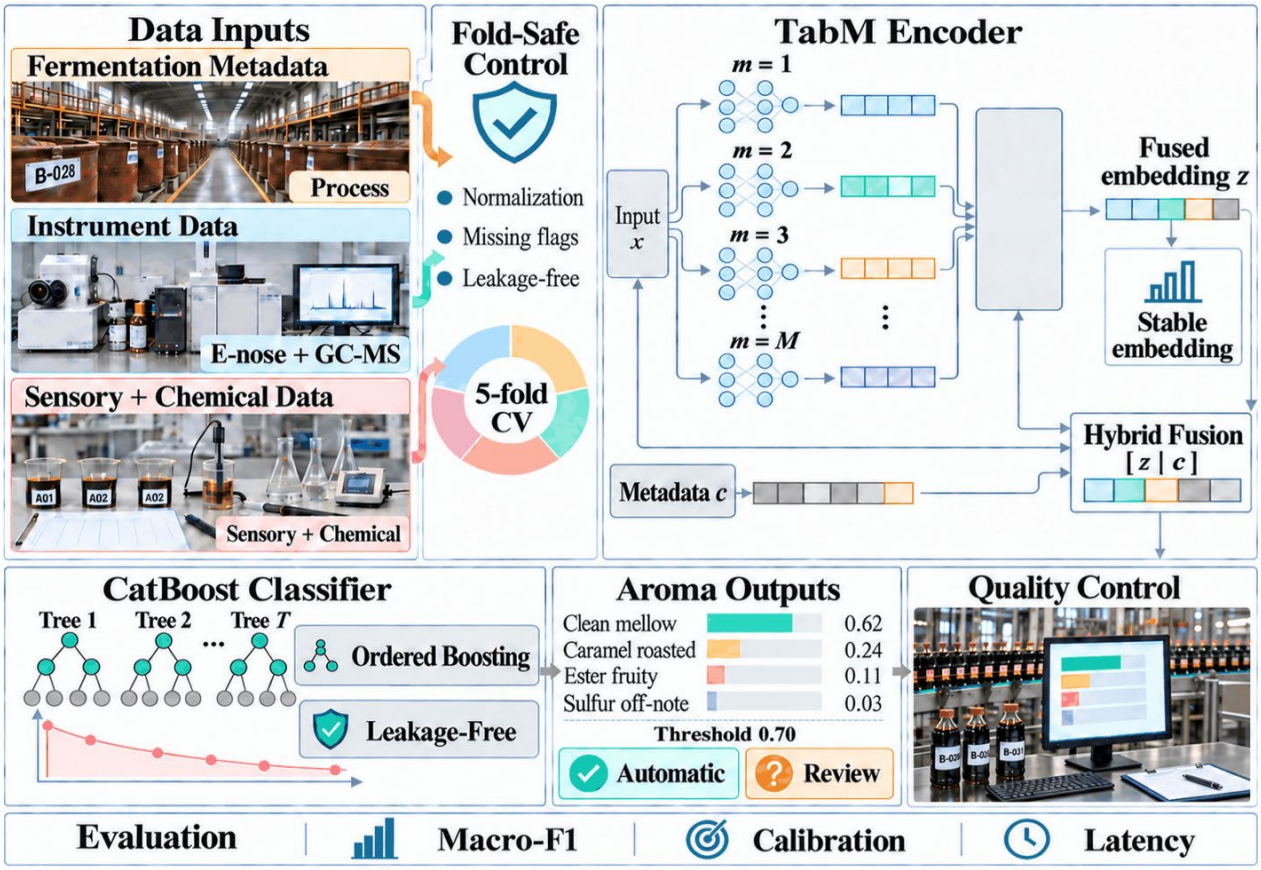


Figure 2. TabM-CatBoost architecture for fused representation learning and ordered boosted aroma classification

After the fused embedding is concatenated with the categorical metadata, the posterior probability of aroma class k is computed as follows:

$$p(y_i = k \mid \boldsymbol{x}_i) = softmax\big(g_k([\boldsymbol{z}_i, \boldsymbol{c}_i])\big) \tag{3}$$

In this equation, g_k is the CatBoost score for class k, and the softmax transformation converts the four class scores into normalized probabilities that can support both classification and confidence-based review.

Training must account for class imbalance and restrict the flexibility of the neural fusion parameters; the optimized loss therefore combines weighted multiclass cross-entropy with L_2 regularization:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} w_k \mathbb{I}(y_i = k) log\, p(y_i = k \mid \boldsymbol{x}_i) + \lambda \|\boldsymbol{\theta}\|_2^2 \quad (4)$$

The class weight w_k limits the dominance of frequent aroma states, the indicator function selects the observed class, and λ controls the penalty applied to the TabM parameters $\boldsymbol{\theta}$ during low-label training.

3.3. Training Protocol and Evaluation Metrics

The training protocol used repeated stratified five-fold cross-validation with identical splits for all models. All replicated measurements from one fermentation batch remained in the same fold, and standardization, imputation, feature selection, and categorical statistics were fitted only on the corresponding training partition. In each five-shot experiment, five labeled batches per aroma class were retained for training; the ten- and twenty-shot settings were constructed analogously. Hyperparameters, including TabM hidden width and dropout and CatBoost depth, learning rate, and L_2 leaf regularization, were selected by nested validation within each outer training fold. Model selection used macro-F1 because accuracy can conceal weak sulfur off-note recognition in an imbalanced dataset. Probability calibration was also evaluated so that uncertain batches could be routed to confirmatory sensory or chemical analysis.

Since the four aroma categories are evaluated with equal importance, the principal score is obtained by averaging the harmonic mean of precision and recall over all K classes:

$$MacroF1 = \frac{1}{K}\sum_{k=1}^{K}\frac{2P_k R_k}{P_k + R_k} \quad (5)$$

This macro-averaged score prevents the more frequent clean mellow class from masking weak recognition of sulfur off-note batches and is consequently used for model selection.

For ablation analysis, the contribution of a feature group was defined as the macro-F1 difference between the complete input and the input obtained after removing that group under the same fold and random seed:

$$\Delta F1 = F1_{\textit{full fusion}} - F1_{\textit{ablated feature group}} \quad (6)$$

The difference $\Delta F1$ is calculated under an identical split and random seed, so it reflects the information removed with the feature group rather than variation introduced by resampling.

The evaluation design separates three questions that are often mixed in fermentation classification experiments. The first question is whether the proposed classifier gives a higher category accuracy than common baselines under the same folds. The second question is whether the gain remains visible when the number of labeled samples is restricted to a few examples per class. The third question is whether the classifier remains usable when measurement channels are noisy, delayed, or partially absent. Treating these questions separately avoids presenting a single accuracy value as evidence for all operating conditions. It also makes the later ablation results more meaningful, because each removed feature group is judged under the same split, metric, and class-weighting protocol.

4. Experimental Data and Analysis

4.1. Dataset Configuration and Baseline Performance

The 164 batches of soy sauce fermentation in the experiment included pilot and workshop-scale tanks at four different production cycles. Small fermented food datasets with rare sensory states are known to have a high fold sensitivity [20]. The number of clean mellow batches was 58, caramel roasted was 42, ester fruity was 39, and sulfur off-note was 25. Each batch had 18 electronic-nose descriptors, 31 selected GC-MS volatile intensities, 7 physicochemical indices, 6 sensory descriptor scores and 5 categorical metadata fields. In the present dataset, the coefficients of variation for methional, 2-methylbutanal, ethyl acetate and 2,5-dimethylpyrazine were 0.62, 0.55, 0.48 and 0.44, respectively, so fold-stable normalization was used. The median missing proportion of volatile compounds was 3.8%, and for sulfur-containing compounds, it was 11.6% because some clean-smell samples were below the detection limit. The distribution of the process metadata was also uneven: 46 batches were high-temperature finished, 73 batches were low-temperature aged, and 31 batches were mixed starter fermented. Figure 3(a) shows that sulfur off-note represented only 15.2% of the batches, whereas clean mellow represented 35.4%; this 20.2-point prevalence gap makes accuracy particularly sensitive to majority-class decisions. Figure 3(b) shows that amino nitrogen increased from 0.63 to 0.94 g/100 mL between days 15 and 90 while the normalized e-nose S2 response declined from 1.18 to 0.91. Their opposing trajectories indicate that biochemical maturation and headspace sensor response supplied complementary temporal evidence rather than duplicate measurements.

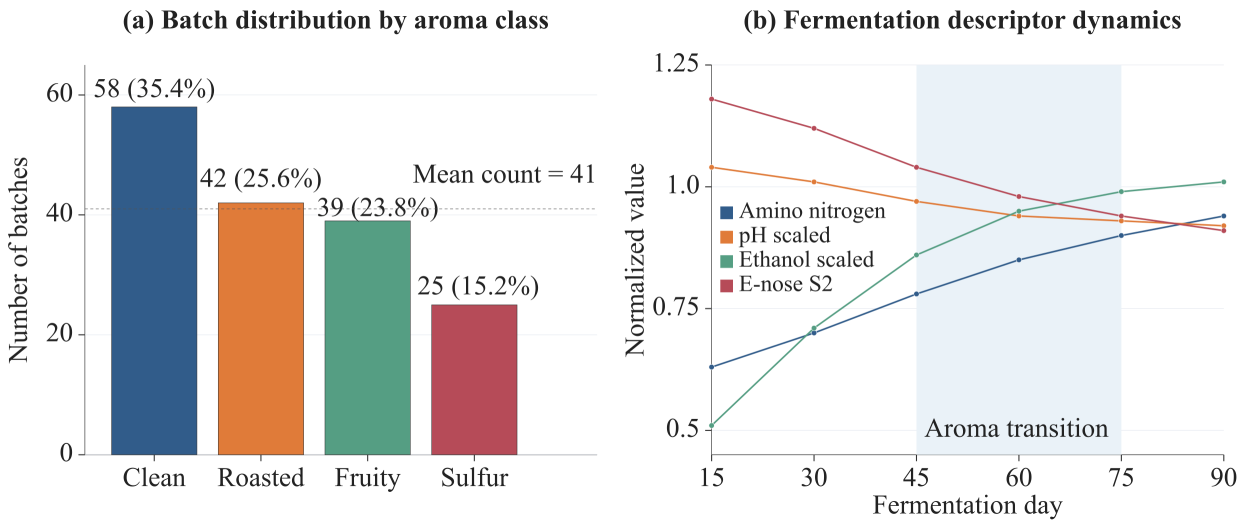


Figure 3. Dataset composition and fermentation-stage descriptor trends: (a) class distribution and batch count; (b) amino nitrogen, pH, ethanol, and e-nose sensor dynamics across fermentation days

The benchmark models used for comparison were Support Vector Machine, Random Forest, XGBoost, CatBoost, a multi-layer perceptron, TabM alone, and the proposed TabM-CatBoost model. Aroma generally does not have a single origin and is usually less grouped than close-knit sets of sensory attributes [21]. In this experiment, the accuracy and macro-F1 values for the support vector machine, random forest, XGBoost and CatBoost were 86.8% and 84.9%, 88.4% and 86.1%, 90.1% and 88.3%, and 92.1% and 91.1%, respectively. TabM alone reached 93.0% accuracy and 92.0% macro-F1. As shown in Table 1, the proposed model achieved 95.6% accuracy, 94.9% macro-F1, 95.1% balanced accuracy, and 0.973 macro-AUC, and all these values exceeded the best single-baseline results for all indicators.

Table 1. Baseline model performance on the soy sauce fermentation aroma dataset

Model	Accuracy	Macro-F1	Balanced Accuracy	Macro-AUC	Main Weakness
SVM	86.8	84.9	85.6	0.914	Class boundary shifts under five-shot labels
Random Forest	88.4	86.1	86.8	0.926	Weak recovery of sulfur off-note cases
CatBoost	92.1	91.1	91.4	0.952	Limited compression of weak volatile signals
TabM	93.0	92.0	92.2	0.958	No ordered categorical boosting stage
TabM-CatBoost	95.6	94.9	95.1	0.973	Residual roasted-fruity confusion

Figure 4(a) indicates that the remaining confusion was concentrated between caramel roasted and ester fruity: 38 of 42 roasted batches and 36 of 39 fruity batches were correct, while only one sulfur off-note batch was missed. The asymmetric error pattern supports the interpretation that sulfur evidence became distinctive after fusion, whereas roasted and fruity classes still shared alcohol, ester, and heat-derived compound regions. Figure 4(b) reports precision values of 0.966, 0.927, 0.923, and 0.960 and recall values of 0.966, 0.905, 0.923, and 0.960 for clean mellow, caramel roasted, ester fruity, and sulfur off-note. The smaller precision-recall gap for sulfur off-note shows that its improved recall was not obtained through an indiscriminate increase in false alarms.

The error records were examined because class-level accuracy alone can hide the specific operating risk of a fermentation classifier. The three roasted batches predicted as fruity had ethyl acetate above the class median by 21.5% and showed lower pyrazine intensity than the correctly classified roasted group. The two fruity batches predicted as roasted had elevated furfuryl alcohol and higher late-stage heating exposure, which made them chemically closer to mild roasted samples. The single sulfur off-note batch predicted as fruity had low dimethyl disulfide but high isoamyl acetate, and its sensory panel notes described a mixed ester and sulfury tail. These cases show why a small-sample fusion model is more appropriate than a single-threshold screening rule. A threshold rule based on methional alone would correctly flag 19 sulfur off-note batches but would miss six off-note batches whose sulfur expression came from mixed low-intensity compounds.

The numerical gains were also checked against sampling uncertainty. A paired bootstrap over batches gave a 95% confidence interval of 1.9 to 5.7 percentage points for the macro-F1 advantage of TabM-CatBoost over CatBoost. The corresponding interval against TabM alone was 1.4 to 4.6 percentage points. The lower bounds remained positive because the proposed model improved minority-class recall in most resamples, not because a few easy clean mellow

samples dominated the metric. When the validation folds were grouped by production cycle, the macro-F1 values were 95.4%, 94.1%, 95.8%, and 94.3%. Error inspection provided a more demanding comparison than aggregate metrics. Of the seven errors produced by the proposed model, four occurred between caramel roasted and ester fruity, two involved clean mellow samples with unusually high ester intensity, and one was a sulfur off-note batch with a methional value below the class median. CatBoost produced thirteen errors, including four sulfur false negatives. The median gap between the two largest class probabilities was 0.46 for correct decisions and 0.12 for errors. Five of the seven errors had a margin below 0.15, whereas only 6.4% of correct decisions occupied that interval. Among correctly predicted batches, 18.7% of CatBoost outputs had top-class probability below 0.70, while the proposed model reduced this proportion to 9.4%. Across production cycles, the standard deviation of balanced accuracy was 0.74 percentage points for TabM-CatBoost, 1.46 for TabM, and 1.83 for CatBoost. The concentration of errors at small margins supports a dual review rule based on both maximum probability and the gap between the two leading classes.

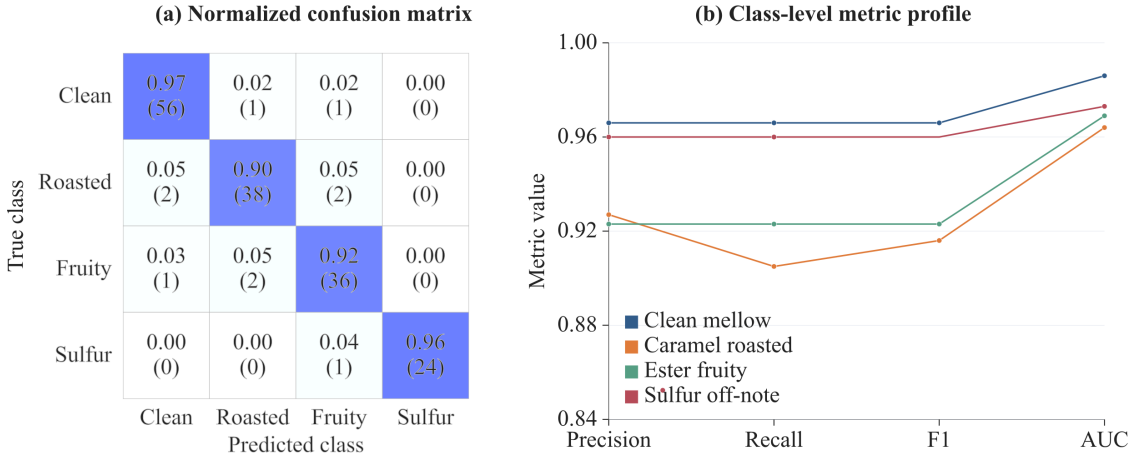


Figure 4. Classification behavior of the proposed model: (a) normalized confusion matrix; (b) class-level precision, recall, F1, and one-vs-rest AUC

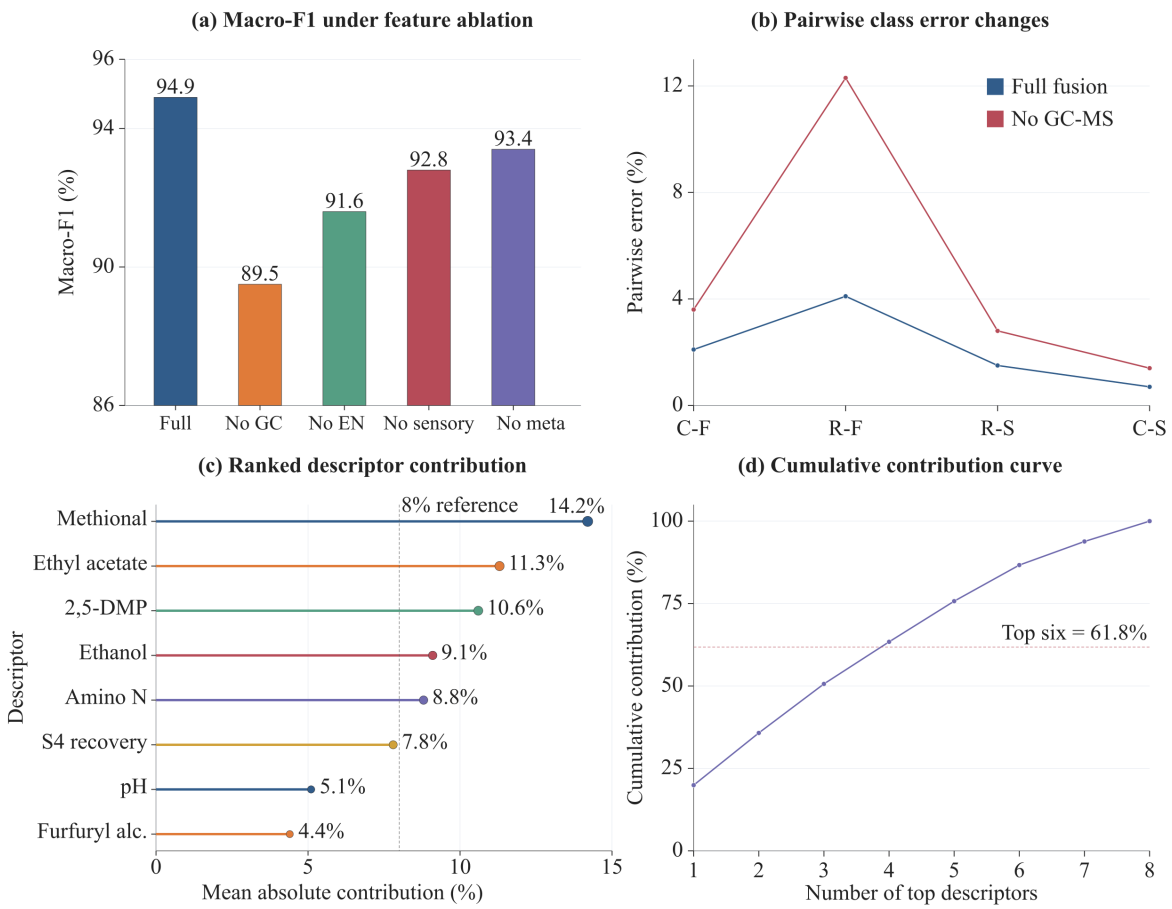


Figure 5. Feature ablation and interpretability analysis: (a) macro-F1 after removing feature groups; (b) pairwise class error changes; (c) ranked SHAP contributions; (d) cumulative contribution of dominant descriptors

4.2. Ablation Study and Small Sample Behavior

The ablation experiment removed one feature group at a time while maintaining identical splits and hyperparameters. Feature-group ablation helps distinguish a stable aroma signature from a result dominated by one correlated production variable [22]. Figure 5(a) shows that removing GC-MS, electronic-nose, sensory, and process-metadata features reduced macro-F1 by 5.4, 3.3, 2.1, and 1.5 percentage points, respectively. The 5.4-point volatile-feature loss was accompanied by a decline in caramel roasted recall from 90.5% to 81.0%, identifying compound-level information as the main source of roasted-class specificity. Figure 5(b) shows that roasted-fruity error rose from 4.1% to 12.3% without GC-MS descriptors, a threefold increase that was larger than the change in any other class pair. Figure 5(c) ranks methional at 14.2%, ethyl acetate at 11.3%, and 2,5-dimethylpyrazine at 10.6% contribution, followed by ethanol, amino nitrogen, and S4 recovery slope. Figure 5(d) shows that the top six descriptors jointly accounted for 61.8% of accumulated contribution while no single descriptor exceeded 14.2%, indicating concentration at the feature-family level without dependence on one measurement.

The label-efficiency experiment compared five-, ten-, and twenty-shot training per class with full-label training. Compact representations are expected to be most useful when labels are scarce but feature channels are abundant [23]. Figure 6(a) shows that TabM-CatBoost achieved macro-F1 values of 89.8%, 92.7%, 94.0%, and 94.9% across the four regimes, compared with 83.7%, 88.9%, 90.6%, and 91.1% for CatBoost. The proposed model recovered 3.8 of its 5.1 percentage-point five-shot deficit by the ten-shot setting. Few-shot aroma classification remains sensitive to minority-class sampling [24]; across the repeated five-shot folds, CatBoost missed seven sulfur off-note batches, whereas the fused model missed three. Figure 6(b) shows that expected calibration error decreased from 0.087 to 0.041 under full-label training and from 0.142 to 0.078 under five-shot training. The larger calibration gains in the five-shot setting indicates that fusion improved both ranking accuracy and probability reliability.

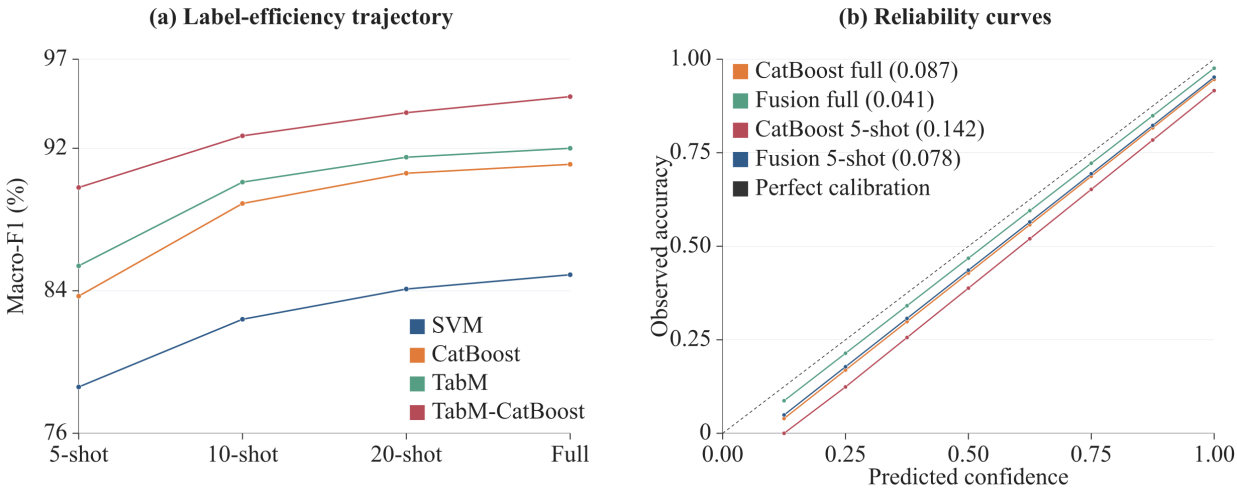


Figure 6. Small-sample performance and calibration: (a) macro-F1 under five-shot, ten-shot, twenty-shot, and full-label settings; (b) reliability curves and expected calibration error

The fold-level variance also changed after feature fusion. Across ten repeated five-fold runs, the proposed model had a macro-F1 standard deviation of 1.12 percentage points, whereas CatBoost alone had 2.38 percentage points and TabM alone had 1.94 percentage points. In the sulfur off-note class, recall ranged from 88.0% to 100.0% for the proposed model and from 72.0% to 96.0% for CatBoost. Small-sample food classification studies frequently report this type of minority-class instability when odor-active sulfur compounds are present at low concentration [25]. The reduced variance suggests that the neural fusion stage did not simply increase model capacity; it regularized multiple weak measurements into a representation that gave the boosting stage a more consistent boundary. The same pattern appeared in probability dispersion. For correctly classified samples, the median predicted probability of the selected class was 0.91 for TabM-CatBoost, 0.84 for CatBoost, and 0.82 for TabM. For misclassified samples, the median probability was 0.57, showing that most errors were low-confidence decisions rather than overconfident failures. This behavior is useful in practice because low-confidence cases can be routed to an expert panel without forcing the model to make an irreversible production decision.

The ablation results show that the four input channels provide complementary rather than interchangeable evidence. GC-MS descriptors are chemically specific but sparse and may fall below the detection limit. Electronic-nose descriptors provide dense kinetic responses but cannot directly identify the responsible compound family. Sensory scores are intuitive but costly and less chemically specific, whereas process metadata constrains plausible aroma trajectories

without uniquely determining them. Removing GC-MS, electronic-nose, sensory, and metadata features changed the predicted labels of 18, 12, 9, and 6 batches, respectively. Only three batches changed under every ablation, confirming that the channels did not encode the same decision signal. Eleven of the 18 volatile-sensitive batches belonged to the roasted or fruity classes and had full-model probability margins below 0.20. Five early-stage clean mellow batches were most sensitive to electronic-nose removal. These patterns indicate that channel agreement reduces uncertainty, whereas disagreements remain concentrated in batches with small posterior margins.

4.3. Robustness and Aroma Indicator Interpretation

Robustness was evaluated through controlled sensitivity tests in which electronic-nose descriptors and GC-MS intensities were perturbed with zero-mean Gaussian noise at relative amplitudes of 2%, 5%, and 10%. These perturbations approximate increasing measurement uncertainty but do not reproduce every mechanism of humidity, sensor aging, or cross-instrument drift [26]. Figure 7(a) shows that TabM-CatBoost declined from 94.9% to 92.7% macro-F1 at 10% perturbation, compared with declines from 91.1% to 87.6% for CatBoost and from 84.9% to 78.8% for the support vector machine. Figure 7(b) attributes the largest recall losses to different channels: GC-MS perturbation reduced caramel roasted recall by 3.9 points, whereas electronic-nose perturbation reduced clean mellow recall by 3.1 points. Figure 7(c) shows a median decision margin of 0.41 at 10% noise, and Figure 7(d) shows that sulfur off-note recall remained 92.0%.

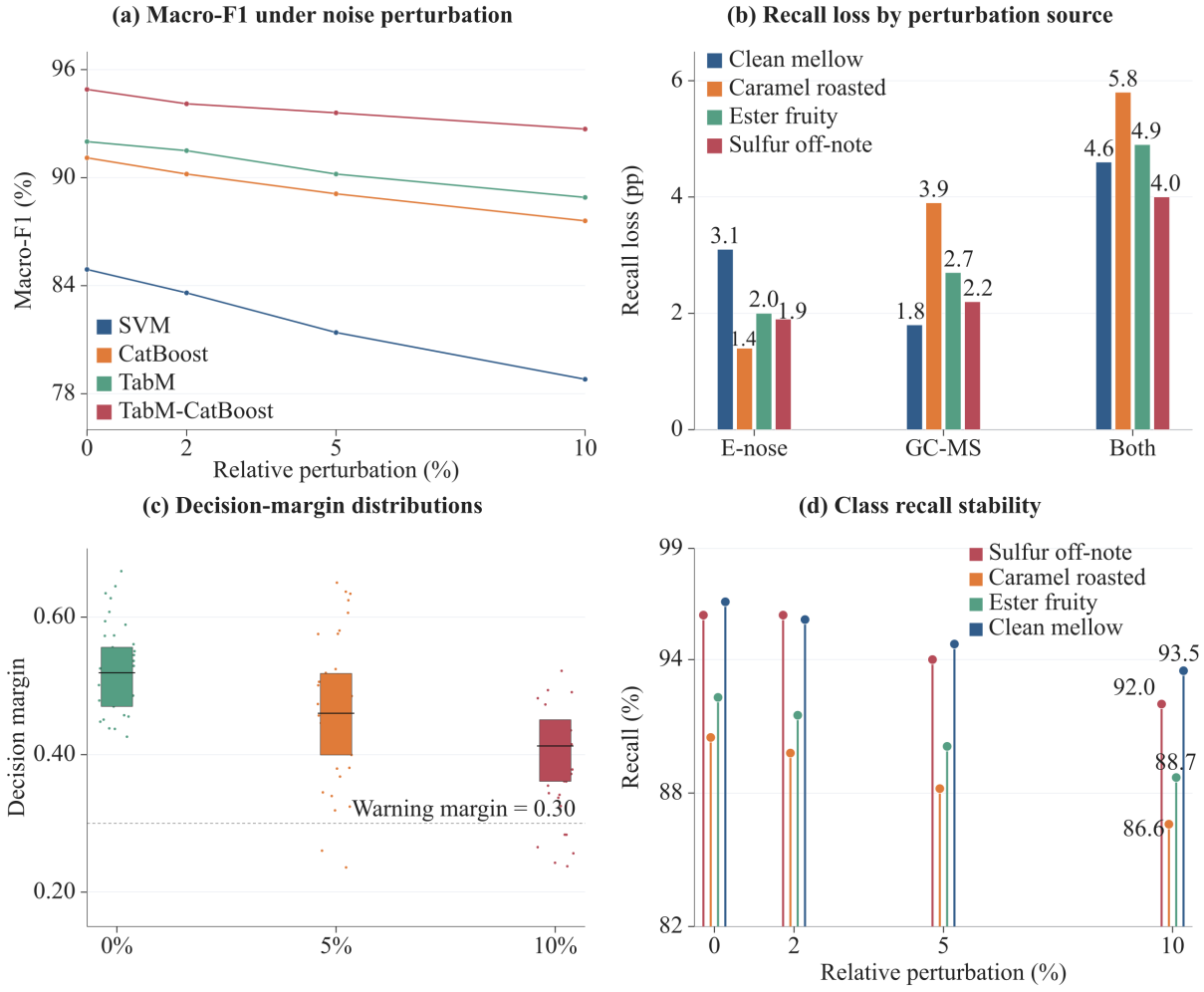


Figure 7. Robustness under instrumental perturbation: (a) macro-F1 curves under noise levels; (b) class recall loss by perturbation source; (c) decision-margin distributions; (d) sulfur off-note recall stability

External-season validation used 31 batches from a later production cycle that was excluded from hyperparameter selection. Seasonal validation is relevant because microbial succession and raw-material composition can change volatile-release intensity under nominally identical processing conditions [27]. TabM-CatBoost achieved 93.5% accuracy and 92.8% macro-F1 on this set, whereas CatBoost achieved 89.7% accuracy and 88.4% macro-F1. The largest shift occurred in ester fruity batches, where ethyl acetate and isoamyl alcohol intensities increased. Nevertheless, only two fruity batches were misclassified as clean mellow. The external batches also had a lower median amino nitrogen value than the internal folds, which could shift clean mellow probabilities upward. The result supports the value of joint

features, although the limited external sample does not establish generalization across factories, instruments, or regional product styles.

Aroma-indicator interpretation linked model behavior to plausible fermentation chemistry without treating the classifier as a mechanistic kinetic model. Targeted volatile studies have associated methional and sulfur derivatives with off-notes, pyrazines with roasted expression, and acetate esters with fruity expression [28]. In the present model, these compound families interacted with sensor kinetics and process metadata rather than acting as isolated thresholds. Figure 8(a) shows a fused-space silhouette value of 0.61 for sulfur off-note and 0.54 for caramel roasted. The tighter sulfur cluster is consistent with its 92.0% recall under strong perturbation, while the lower roasted silhouette explains the persistent overlap with ester fruity batches. Figure 8(b) shows a prototype distance of 1.92 between sulfur off-note and clean mellow, but only 1.18 between caramel roasted and ester fruity. The 0.74 distance difference quantifies why the latter pair accounted for most residual confusion. High-temperature finishing increased roasted probability only when pyrazine descriptors were elevated, and mixed starter fermentation increased fruity probability only when ethyl acetate or phenethyl alcohol exceeded the fold median, limiting the influence of metadata proxies.

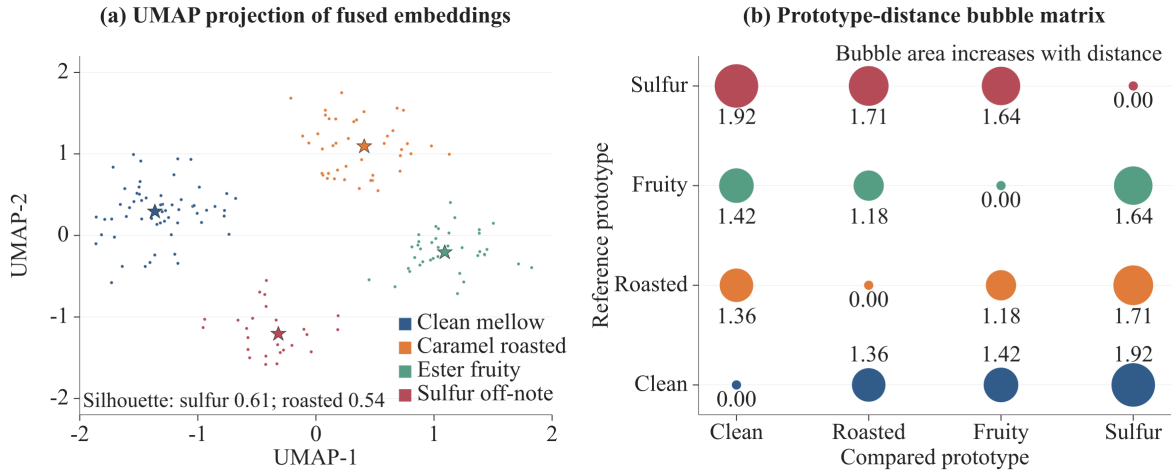


Figure 8. Embedding interpretation and prototype geometry: (a) UMAP projection of fused aroma embeddings; (b) class-prototype distance profile in TabM space

Risk-sensitive thresholding is appropriate when missing an undesirable fermentation state is more costly than requesting an additional verification test [29]. At a threshold of 0.70, the retrospective decision simulation assigned high-confidence labels to 71.3% of batches by day 45 and 89.6% by day 75. It identified 24 of 25 sulfur off-note batches and generated three additional warnings. Batch-level support also requires staged access to measurements and low computational latency [30]. Inference required 8.6 ms per batch, although sample preparation and instrumental acquisition remain the dominant operational costs. With electronic-nose and physicochemical descriptors available on day 45, macro-F1 was 87.2%; adding metadata increased it to 88.6%, GC-MS descriptors available on day 75 increased it to 94.1%, and sensory scores produced 94.9%. The simulated review rule reduced expert review load by 46.2% while retaining 24 of 25 off-note alerts.

5. Conclusion

This study developed a small-sample TabM-CatBoost model that integrates electronic-nose responses, GC-MS volatile profiles, physicochemical indices, sensory descriptors, and process metadata to classify four soy sauce fermentation aroma categories.

1. The proposed model achieved 95.6% accuracy, 94.9% macro-F1, 95.1% balanced accuracy, and a macro-AUC of 0.973 on 164 fermentation batches. Under five-shot training, it retained a macro-F1 of 89.8%, exceeding CatBoost and TabM by 6.1 and 4.4 percentage points, respectively.

2. Feature fusion produced its largest improvement for the minority sulfur off-note class, whose F1 score increased by 6.8 percentage points over CatBoost. Removing GC-MS, electronic-nose, sensory, and process-metadata features reduced macro-F1 by 5.4, 3.3, 2.1, and 1.5 percentage points, respectively. Under 10% instrumental perturbation, TabM-CatBoost retained a macro-F1 of 92.7%, whereas CatBoost declined to 87.6%.

3. The principal limitation is that the dataset was obtained from pilot and workshop-scale fermentation environments and therefore does not fully represent differences among factories, raw-material origins, starter cultures, instruments, and regional product styles. Nevertheless, the two-stage architecture provides an effective innovation for small-sample aroma analysis by combining compact neural feature fusion with leakage-resistant ordered boosting and interpretable feature contributions.

The framework provides a practical basis for label-efficient fermentation supervision. Future work should prioritize multi-factory and multi-instrument validation, uncertainty-aware active learning for selecting new sensory labels, and temporal models that predict aroma trajectories before the end of fermentation.

6. Declarations

6.1 Author Contributions

Slobodan Ćorović contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Darko Cvetković contributes to validation, analysis, investigation, data collection, draft preparation. Mei-hsien Chi contributes to conceptualization, validation, analysis. All authors have read and agreed with the manuscript before its submission and publication.

6.2 Funding and Acknowledgments

This research received no specific financial support from any funding agency.

6.3 Institutional Review Board Statement

Not applicable.

6.4 Informed Consent Statement

Not applicable.

6.5 Data Availability

Not applicable.

6.6 Declaration of Competing Interest

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

6.7 Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors utilized Grammarly for linguistic polishing, logical organization and draft revision. The generative AI tool was not involved in research design, data collection, data analysis, or the equation of core conclusions. All outputs generated by AI were reviewed, revised and verified manually by the authors. The authors take full responsibility for all content of this manuscript.

7. References

[1] Chen, X. C., Meng, F. B., Tian, H. L., Chen, A. X., Wang, X. B., & Li, Y. C. (2024). Dynamic changes in metabolomics and flavoromics during industrial scale fermentation of Chinese traditional soy sauce. Food Bioscience, 59, 103940. https://doi.org/10.1016/j.fbio.2024.103940

[2] Gao, X., Han, X., Tang, Y., Song, H., Meng, Q., Chen, R., Yu, J., Yang, P., & Yu, Z. (2024). Exploring changes in key aroma-active compounds of soy sauce due to different fermentation processes. Journal of Food Composition and Analysis, 133, 106393. https://doi.org/10.1016/j.jfca.2024.106393

[3] Kan, Q., Cao, L., He, L., Wang, P., Deng, G., Li, J., Fu, J., Huang, Q., Ho, C. T., Li, Y., Xie, C., Cao, Y., & Wen, L. (2025). Tracing the change of the volatile compounds of soy sauce at different fermentation times by PTR-TOF-MS, E-nose and GC–MS. Food Chemistry: X, 25, 102002. https://doi.org/10.1016/j.fochx.2024.102002

[4] Tan, Q., Wu, Y., Li, C., Jin, J., Zhang, L., Tong, S., Chen, Z., Ran, L., Huang, L., & Zuo, Z. (2024). Characterization of Key Aroma Compounds of Soy Sauce-like Aroma Produced in Ferment of Soybeans by Bacillus subtilis BJ3-2. Foods, 13(17), 2731. https://doi.org/10.3390/foods13172731

[5] Guo, Q., Peng, J., & He, Y. (2024). A Systematic Comparative Study on the Physicochemical Properties, Volatile Compounds, and Biological Activity of Typical Fermented Soy Foods. Foods, 13(3), 415. https://doi.org/10.3390/foods13030415

[6] Ji, C., Petchkongkaew, A., van Ruth, S., Wu, D., & Elliott, C. (2024). The crucial importance of soy sauce authenticity: Global trade, adulteration risks, and analytical challenges. Trends in Food Science & Technology, 152, 104666. https://doi.org/10.1016/j.tifs.2024.104666

[7] Wu, Z., Chao, J., Tang, H., Liu, T., Jiang, L., & Liu, Y. (2024). Characterization of key aroma-active compounds in different types of Douchi based on molecular sensory science approaches. Food Chemistry: X, 21, 101170. https://doi.org/10.1016/j.fochx.2024.101170

[8] Wang, X., Dang, C., Liu, Y., Ge, X., Suo, R., Ma, Q., & Wang, J. (2024). Effect of indigenous Saccharomyces cerevisiae strains on microbial community successions and volatile compounds changes during Longyan wine fermentation. Food Bioscience, 57, 103595. https://doi.org/10.1016/j.fbio.2024.103595

[9] Singh, R., Nisha, R., Naik, R., Upendar, K., Nickhil, C., & Deka, S. C. (2024). Sensor fusion techniques in deep learning for multimodal fruit and vegetable quality assessment: A comprehensive review. Journal of Food Measurement and Characterization, 18(9), 8088–8109. https://doi.org/10.1007/s11694-024-02789-z

[10] Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., & Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. Nature, 637(8045), 319–326. https://doi.org/10.1038/s41586-024-08328-6

[11] Fan, Y., & Waldmann, P. (2024). Tabular deep learning: a comparative study applied to multi-task genome-wide prediction. BMC Bioinformatics, 25(1), 322. https://doi.org/10.1186/s12859-024-05940-1

[12] Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., & Harmouch, H. (2025). The effects of data quality on machine learning performance on tabular data. Information Systems, 132, 102549. https://doi.org/10.1016/j.is.2025.102549

[13] Silvey, S., & Liu, J. (2024). Sample Size Requirements for Popular Classification Algorithms in Tabular Clinical Data: Empirical Study. Journal of Medical Internet Research, 26, e60231. https://doi.org/10.2196/60231

[14] Uddin, S., & Lu, H. (2024). Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data. PLOS ONE, 19(4), e0301541. https://doi.org/10.1371/journal.pone.0301541

[15] Villaizán-Vallelado, M., Salvatori, M., Carro, B., & Sanchez-Esguevillas, A. J. (2024). Graph Neural Network contextual embedding for Deep Learning on tabular data. Neural Networks, 173, 106180. https://doi.org/10.1016/j.neunet.2024.106180

[16] Wang, A. X., Chukova, S. S., Simpson, C. R., & Nguyen, B. P. (2024). Challenges and opportunities of generative models on tabular data. Applied Soft Computing, 166, 112223. https://doi.org/10.1016/j.asoc.2024.112223

[17] Liakos, K. G., Athanasiadis, V., Bozinou, E., & Lalas, S. I. (2025). Machine Learning for Quality Control in the Food Industry: A Review. Foods, 14(19), 3424. https://doi.org/10.3390/foods14193424

[18] Luo, H., Akkermans, S., & Van Impe, J. F. M. (2025). Demystifying food flavor: Flavor data interpretation through machine learning. Food Chemistry, 483, 144000. https://doi.org/10.1016/j.foodchem.2025.144000

[19] Zhou, C., Yu, Y., Ai, J., Song, C., Cui, Z., Zhou, Q., Zhao, S., Huang, R., Ao, Z., Peng, B., Chen, P., Feng, X., Li, D., & Liu, Y. (2025). Fruit wines classification enabled by combing machine learning with comprehensive volatiles profiles of GC-TOF/MS and GC-IMS. Food Research International, 204, 115890. https://doi.org/10.1016/j.foodres.2025.115890

[20] Li, S., Li, T., Han, Y., Yan, P., Li, G., Ren, T., Yan, M., Lu, J., & Qiu, S. (2024). Machine learning discrimination and prediction of different quality grades of sauce-flavor baijiu based on biomarker and key flavor compounds screening. Food Chemistry: X, 24, 101877. https://doi.org/10.1016/j.fochx.2024.101877

[21] Gil, M., Rudy, M., Duma-Kocan, P., & Stanisławczyk, R. (2025). Electronic Sensing Technologies in Food Quality Assessment: A Comprehensive Literature Review. Applied Sciences, 15(3), 1530. https://doi.org/10.3390/app15031530

[22] Peveler, W. J. (2024). Food for Thought: Optical Sensor Arrays and Machine Learning for the Food and Beverage Industry. ACS Sensors, 9(4), 1656–1665. https://doi.org/10.1021/acssensors.4c00252

[23] Schreurs, M., Piampongsant, S., Roncoroni, M., Cool, L., Herrera-Malaver, B., Vanderaa, C., Theßeling, F. A., Kreft, Ł., Botzki, A., Malcorps, P., Daenen, L., Wenseleers, T., & Verstrepen, K. J. (2024). Predicting and improving complex beer flavor through machine learning. Nature Communications, 15(1), 2368. https://doi.org/10.1038/s41467-024-46346-0

[24] Heng, Y., Zhou, Y., Nguyen, D. H., Nguyen, V. D., & Jiao, M. (2025). An electronic nose drift compensation algorithm based on semi-supervised adversarial domain adaptive convolutional neural network. Sensors and Actuators B: Chemical, 422, 136642. https://doi.org/10.1016/j.snb.2024.136642

[25] Sun, J., Zheng, H., Diao, W., Sun, Z., Qi, Z., & Wang, X. (2025). Prototype-Optimized unsupervised domain adaptation via dynamic Transformer encoder for sensor drift compensation in electronic nose systems. Expert Systems with Applications, 260, 125444. https://doi.org/10.1016/j.eswa.2024.125444

[26] Wei, G., Xu, Y., Lv, X., Jiao, S., & He, A. (2024). An Adaptive Drift Compensation Method Based on Integrated Dual-Channel Feature Fusion for Electronic Noses. IEEE Sensors Journal, 24(16), 26814–26824. https://doi.org/10.1109/jsen.2024.3412679

[27] Dennler, N., Drix, D., Warner, T. P. A., Rastogi, S., Casa, C. D., Ackels, T., Schaefer, A. T., van Schaik, A., & Schmuker, M. (2024). High-speed odor sensing using miniaturized electronic nose. Science Advances, 10(45), eadp1764. https://doi.org/10.1126/sciadv.adp1764

[28] Cai, M., Xu, S., Zhou, X., & Lu, H. (2024). Electronic Nose Humidity Compensation System Based on Rapid Detection. Sensors, 24(18), 5881. https://doi.org/10.3390/s24185881

[29] Wei, J., Li, Z., Kuang, J., Yan, Z., Wang, L., Lin, Y., Du, J., Li, K., Wang, Y., & Yang, L. (2024). Microbial community succession and changes of volatile compounds in the fermentation process of bamboo shoots. Food Microbiology, 124, 104618. https://doi.org/10.1016/j.fm.2024.104618

[30] Tang, A., & Peng, B. (2024). Metatranscriptomics reveals microbial community function succession and characteristic flavor formation mechanisms during black rice wine fermentation. Food Chemistry, 457, 140428. https://doi.org/10.1016/j.foodchem.2024.140428