Sustainable Food Technology

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d5fb00580a

Machine learning-based prediction of sensory quality in tea blends using a semi-trained panel assessment

Onkar Sarma [a] and Kavya Dashora*[ab]

Blending various types of teas with complementary biochemical profiles offers a promising path for enhancing the sensory quality of tea blends. Traditionally, blending is done by professional tea tasters, which makes it dependent on human sensitivity, time-consuming and difficult to scale. Thus, the prediction of tea's sensory quality and identification of its biochemical drivers by reducing the cost of experimentation and expediting the blending process could have tremendous techno-economic value. In this study, we curated a meta-dataset from our own experiments involving blends of four major varieties of Assam tea. Their key biochemical composition was analyzed, and sensory score was estimated using a lexicon-based descriptive technique with a semi-trained consumer-focused panel. Combining it with multiple machine learning models, we showed that while an increase in the (+)-C, TP, TR, and pH values led to a lower sensory score, higher protein/TP, TSS, organic acid, CAF, CAF/TP and TF/TR values resulted in an increased sensory score. Further, we adopted a game theory-based model agnostic interpretation technique called SHAP (Shapley Additive exPlanations) to identify features contributing to higher sensory scores and their relative significance. The key quality indicators were found to be the (+)-C, protein/TP, protein, TP, TSS, TR, citric acid, malic acid, and ascorbic acid contents. By integrating a consumer-centric evaluation with interpretable machine learning, we demonstrated how meta dataset, cutting-edge machine learning techniques, and model interpretability methods could be seamlessly integrated to reduce the number of experiments, minimize dependency on expert intuition, and enable automated quality assessment for developing superior tea blends.

Received 10th September 2025
Accepted 12th January 2026

DOI: 10.1039/d5fb00580a

rsc.li/susfoodtech

Sustainability spotlight

This study addresses the critical need for objective and efficient sensory evaluation in the tea industry by analyzing key biochemical features alongside consumer-driven sensory scores from a semi-trained panel. By leveraging artificial intelligence, this study reveals how each biochemical feature influences quality characteristics, thereby enabling manufacturers to optimize tea blends for superior taste. Adopting a semi-trained consumer panel as an alternative to expert tasters significantly reduces costs and time while analyzing consumers' perspectives about the product at the production level. This sustainable advancement aligns with UN SDGs 9 (industry, innovation, and infrastructure), 12 (responsible consumption and production), and 8 (decent work and economic growth), promoting innovation, resource efficiency, and inclusive economic practices.

1. Introduction

The state of Assam is the single largest tea growing region in India, producing about 606 million kg of tea annually, which is around 50.69% of India's total tea production.[1] Globally known for its crush tear curl (CTC) and orthodox black tea variants, other popular varieties of Assam tea include green tea, white tea, and oolong tea.[2] *Camellia sinensis* var. assamica or Assam tea is known for its distinct taste and flavor profile, resulting from its unique combination of biochemical attributes. Among these, flavanols, alkaloids, and phenolic acids are important bio-constituents that give tea its bitterness and astringency.[3] Amino acid and soluble sugar impart umami and sweet taste.[4,5] The blending of different teas with complementary biochemical profiles can enhance the overall sensory quality of tea blends. Traditionally, this process has mainly been explored through trial-and-error experiments by tea professionals based on their subjective judgments. However, this method is dependent on human sensitivity, time-consuming and therefore difficult to scale.

Tea blending has long been a practice in the tea industry to develop products with superior sensory qualities. Tea blending is performed by mixing two or more varieties and grades of tea from different regions or batches to maintain its sensory quality

[a]Agricultural-Nanotechnology Lab, Center for Rural Development and Technology, IIT Delhi, Delhi, 110016, India. E-mail: kdashora@rdat.iitd.ac.in
[b]Yardi School of Artificial Intelligence (Yardi ScAI), IIT Delhi, India

as per consumer demand.[6] Most studies have investigated the blending of black tea variants belonging to different grades. For instance, Xia fused images and spectral features to predict the sensory characteristics of different blends of black tea.[7] Ling employed a digital blending strategy in which NIR spectra of different blends of black tea were used to predict the sensory score.[8] Tie addressed the blending problem by developing hierarchical spatial clustering-based algorithms that achieved a universal, low-cost, and efficient blending program.[9] Turgut compared three NIR equipment for calibrating biochemicals with spectral data for the sensory prediction of black tea.[10] Recent studies, such as those by Chen and co-authors, have experimented with blending fresh tea leaves of three clonal varieties to manufacture black tea with improved flavor and aroma.[11] Most of these studies have investigated a particular tea type, and there is a lack of reports on blends made from differently processed teas, such as green tea, white tea, oolong tea, and black tea. The underlying biochemical properties responsible for each tea type's contribution to a blend's sensory attributes have not been explored thoroughly. Moreover, previous studies that employed machine learning (ML) for sensory prediction primarily relied on trained sensory panels to generate sensory data. Although such expert panels offer consistency and domain knowledge, they may not accurately reflect the preferences of everyday consumers. In contrast, an alternative and more scalable approach would be to train consumers directly to form a semi-trained panel and analyze the sensory profile of tea infusions using the Check-All-That-Apply method. This Semi Trained-Check-All-That-Apply (ST-CATA) approach is important in the context of the tea industry, where market success is largely influenced by consumer preferences, acceptance, and purchase behaviors. Accurately predicting the sensory quality and identifying its biochemical drivers can thus offer significant techno-economic benefits by reducing the experimental costs and accelerating the development of superior tea blends. However, this remains a challenge due to the inherent subjectivity of sensory evaluation and the complex, non-linear relationship between biochemical composition and human perception. To address this and make the process more objective, replicable, and aligned with market needs, there is a pressing need for a cost-effective, consumer-inclusive methodology that enables the automated assessment of tea quality.

This study aims to explore the effects of incorporating diverse aroma and flavor profiles from differently processed teas into a product and to assess consumer preferences. Various blends of differently processed teas were prepared and analyzed for key biochemical features. Then, a lexicon-based descriptive sensory evaluation technique was used, employing a semi-trained panel to assess the blend's sensory score. Finally, the ML models were trained to predict the sensory grades of the tea blends. This study aims to address the following key questions:

- What are the key biochemical features and their order of relevance in influencing tea sensory quality?
- How to integrate interpretable ML with sensory data obtained from a semi-trained panel to objectivize tea sensory assessments?
- What is the effect of differently processed teas on the overall sensory quality of a tea blend?

This research will aid tea producers in choosing blend ingredients and offer a scientific framework for consumers to identify high-quality tea. Furthermore, in the context of tea research and development, understanding the connections between biochemical content and the sensory qualities of tea can aid in developing superior blends and optimizing processing strategies to elevate the overall quality of processed tea.

2. Materials and methods

Caffeine (anhydrous ≥99%, HPLC grade), acetic acid, orthophosphoric acid (HPLC grade), ascorbic acid, acetonitrile (HPLC grade), methanol (HPLC grade), anthrone, Folin and Ciocalteu's phenol reagent, Bradford reagent, and sodium sulphate (Na_2SO_4) were procured from Merck, India. (+)-Catechin (≥98%, HPLC grade) was procured from CDH, India. EDTA, phenylmethanesulphonyl fluoride (PMSF), and sodium carbonate (Na_2CO_3) were obtained from Himedia Laboratories, India. Sulphuric acid, Tris HCL, β-mercaptoethanol, and organic acid standards (99% pure), ethyl acetate, *n*-butanol, and sodium bicarbonate ($NaHCO_3$) were obtained from SRL, India. The L-theanine standard (>98%) was obtained from the Tokyo Chemical Industry (TCI). The absorbance spectra of the samples were recorded using a UV-Vis spectrometer (Shimadzu 1800). An ultra-high-performance liquid chromatography (UHPLC) system (Waters Corporation, Milford, MA, USA) was used to separate and identify the compounds.

Four varieties of Assam tea (*Camellia assamica*), including green tea, white tea, oolong tea, and black tea (CTC), were procured from the Tea Auction Centre, Guwahati, Assam. These tea varieties included multiple Tocklai vegetative cultivars (developed at the Tocklai Tea Research Institute, Assam, India).

2.1. Sample preparation

Procured tea varieties were hand ground in a mortar and pestle to powder form and sieved through a mesh of 150-micron size. It was then mixed in the required proportions, as shown in Table 1, to develop 30 tea blends. To determine the blend proportions, an optimal mixture design was used by keeping constraints on the amount of each tea variety so that the effect of their variations is well captured. A quadratic model was used with a total of 30 runs, as depicted by the sample codes in Table 1. The Stat-Ease Design Expert 13 software was used to carry out the mixture design. The total weight of each blend was maintained at 4 g. All the tea samples were sealed and stored in the dark at 4 °C prior to chemical analyses and sensory evaluation.

2.2. Biochemical estimation

In this experiment, fifteen key taste influencing biochemical features, namely total soluble sugar (TSS), protein, total polyphenol (TP), (+)-catechin (C), caffeine (CAF), organic acids (malic acid, citric acid, ascorbic acid, oxalic acid, gallic acid, and succinic acid), L-theanine, theaflavin (TF), thearubigin (TR), and

Table 1 Proportions of the four varieties of processed tea in tea blends

Sample code	Green tea (gram)	White tea (gram)	Oolong tea (gram)	Black tea (gram)
01	0.5	0.5	2	1
02	1.5	0.5	1.5	0.5
03	1	2	0.5	0.5
04	1.5	1.5	0.5	0.5
05	1	0.5	2	0.5
06	1	0.5	1	1.5
07	0.5	1	1	1.5
08	1	1.5	1	0.5
09	0.5	0.5	0.5	2.5
10	1.5	0.5	0.5	1.5
11	0.5	1	0.5	2
12	1.5	1	0.5	1
13	0.5	0.5	1.5	1.5
14	1	0.5	1.5	1
15	0.5	2	0.5	1
16	1	1	1	1
17	0.5	1.5	1.5	0.5
18	0.5	0.5	1	2
19	1	0.5	0.5	2
20	1	1.5	0.5	1
21	1	1	0.5	1.5
22	0.5	1	2	0.5
23	0.5	2	1	0.5
24	0.5	1.5	1	1
25	2	0.5	1	0.5
26	2	1	0.5	0.5
27	1.5	1	1	0.5
28	2	0.5	0.5	1
29	1	1	1.5	0.5
30	1.5	0.5	1	1

pH were analyzed. The procedures for estimating the key indices are described briefly below:

2.2.1. Estimation of TSS. 0.3 g sample was extracted with 40 mL ultrapure water at 95 °C for 60 min. 1 mL of extract was filtered, diluted and then mixed with 5 mL anthrone reagent.[12] The solution was incubated at 90 °C for 17 min, and the absorbance was recorded at 620 nm.

2.2.2. Estimation of protein. 1.5 g of the sample was mixed with an equal volume of protein extraction buffer. The extract was centrifuged for 20 min at 6440×*g* and 4 °C, and 100 μL of the supernatant was mixed with 1.6 mL of Bradford reagent and incubated in the dark for 1 hour, followed by absorbance reading at 595 nm.[13]

2.2.3. Estimation of TP. TP was estimated using Folin-Ciocalteu's (FC) phenol reagent (2 mL, 0.2 M),.[14] The sample was extracted with methanol (70% v/v), then diluted with water, mixed with FC reagent and sodium carbonate (7% w/v), and then incubated in the dark for 60 min. The final absorbance spectra were recorded at 765 nm.

2.2.4. Estimation of C and CAF. C and CAF were estimated by applying the FSSAI method[15] using a UHPLC system. The UHPLC conditions followed those reported in ref. 16.

2.2.5. Estimation of organic acids and L-theanine. Organic acid and L-theanine contents were analyzed as per ref. 17 using the UHPLC system.

2.2.6. Estimation of TF and TR. TF% and TR% were analyzed according to ref. 18. Tea infusion was treated with ethyl acetate, saturated oxalic acid solution, $NaHCO_3$ (2.5%), and *n*-butanol in a separating funnel and absorbance was measured at 380 nm.

2.2.7. Estimation of pH. The pH of the tea infusions was measured using a pH meter (Labman LMPH-10).

2.3. Sensory estimation

Sensory assessment of the 30 blends was conducted using the ST-CATA method. Sensory attributes were selected as per the key flavor and taste factors present in Assam tea.[19,20] Representative tea consumers ($n = 20$) were screened and recruited from the IIT Delhi campus and the adjacent market complex after inquiring about their tea consumption history and frequency. Twenty assessors (13 males and 7 females, aged 18–57) received one and a half hours of sensory training by physical perception (tasting and sniffing) of references for each sensory attribute according to the methods mentioned previously.[21] Analysis was conducted in three sessions, where in each session (lasting approximately 30–40 min), 10 samples with a two-digit code were provided for the batch. There was a 15 min break between each session.

Samples were prepared for sensory analysis based on ISO standards for brewing.[22,23] Each blend (0.4 g) was placed in disposable fiber-made tea bags and placed in porcelain cups maintained at 50 °C, and 100 mL of boiled water was added. The total sensory score was evaluated based on five sensory-evaluation factors, totaling a score of 100, using the formula of Xiong:[24]

$$\text{Sensory score} = (\text{tea appearance} \times 25\%) + (\text{infusion color} \times 10\%) + (\text{aroma} \times 25\%) + (\text{taste} \times 30\%) + (\text{solubility} \times 10\%)$$

3. Dataset development

To thoroughly investigate the relationships between tea sensory attributes and their biochemical compounds, a comprehensive and diverse training dataset collected from various experimental scenarios is essential. Our dataset, Tea Biochemical and Sensory Dataset (TeaBioSens) contained a total of 600 experimental data points consisting of biochemical estimation and sensory score ($n = 20$) for each sample. Certain parameters, such as TSS, protein, TP, CAF, C, TF, TR, pH, organic acids, and L-theanine, were directly estimated from the experiments, while other factors such as TP/theanine, TF/TR, protein/TP, and CAF/TP were calculated. This dataset is continuously updated due to ongoing research on investigating highly influential parameters. As we continue to add more experimental parameters, TeaBioSens is strengthened to yield robust results.

3.1. Description of input parameters

A total of nineteen parameters were used as input features for the ML model classification. These parameters play a major role in the development and modulation of tea's sensory

characteristics. Therefore, understanding their effects and relevance can help in objectively achieving the targeted sensory quality of tea. These parameters can be broadly categorized as experimental and calculated parameters.

3.1.1. Experimental parameters. These parameters include C, TP, CAF, TSS, L-theanine, protein, organic acid, pH, TF, and TR. Catechins are the major phenolics in tea leaves, constituting up to 80% of the TP. Therefore, C has been considered an important quality attribute as it majorly influences tea sensory.[3] Tea polyphenol and CAF levels are influenced by leaf maturity, plant growth, metabolism, and influence tea sensory quality. TSS has been previously attributed to enhance the sweetness and mellowness of tea and L-theanine, contributing to umami and sweetness.[4,25] Among proteins, several peptides have been recognized to improve taste and enhance desirable sensory attributes, such as sweet and umami tastes. Citric acid, malic acid, ascorbic acid, oxalic acid, gallic acid, and succinic acid are majorly present in tea. These six organic acids were analyzed as research has concluded them as contributors to a variety of tastes, such as sourness, light umami, and gentle astringency in tea infusions, whereas pH is a prime indicator of sourness in tea infusions.[26] TF and TR are the key oxidation compounds found in black and oolong tea that influence color, briskness, mouthfeel and overall quality of tea liquor.[27] TF content positively affects the brightness and mouthfeel of tea infusion, whereas TR content can hamper liquor taste and brightness.[27]

3.1.2. Calculated parameters. Besides the key biochemical components, another four features, including CAF/TP, TP/theanine, protein/TP, and TF/TR, were fed into the models for the prediction of sensory quality. The ratio of CAF/TP is believed to balance tea taste by affecting the sensory threshold.[28] Additionally, numerous studies have observed the ratio of TP/theanine as a key indicator of freshness and briskness in tea.[4,29,30] The strong relation between amino acids and other compounds indicates the important role that the protein to polyphenol ratio might play in tea's sensory. Several studies have shown that the TF/TR ratio is an important biomarker of tea quality characteristics and positively influences tea infusion brightness.[27,31]

3.2. Modelling methodology

K-medoids was performed to classify the samples into two categories of low grade and high grade based on the overall sensory score. The optimal number of clusters for classifying the samples was based on the silhouette coefficient. In the TeaBioSens dataset, the highest silhouette coefficient was obtained for $k = 2$. Moreover, clustering into two groups based on overall sensory scores ensures that samples are best separated into two dominant sensory profiles, quality levels, or taste groups among consumers. This clearly distinguishes the driving biochemical features of the sensory profile of tea blends. K-Medoid is an unsupervised clustering method that identifies medoids based on the location of data points and progressively improves clustering results. This approach is more accurate in the classification of data where a clear boundary does not exist due to continuous scoring within a narrow range. The K-medoids clustering method has been applied in a similar scenario of sample classification based on sensory score.[14,32]

Four ML models: Extreme gradient boosting (XGB), support vector machine (SVM), logistic regression (LR), and multilayer perceptron (MLP) were calibrated and compared for their strength in classifying the samples into categories obtained from K-medoids. The methodology involved in the modelling and interpretation of the results is depicted in Fig. 1.

XGB constructs an ensemble of trees one after another, correcting the node's errors at every level to develop an optimized gradient boosting network. This process is repeated continuously for XGB to be able to capture the intricate correlations between quality metrics and progressively improve predictions. LR is typically used for classification and works on a supervised learning approach. It uses a sigmoid function to

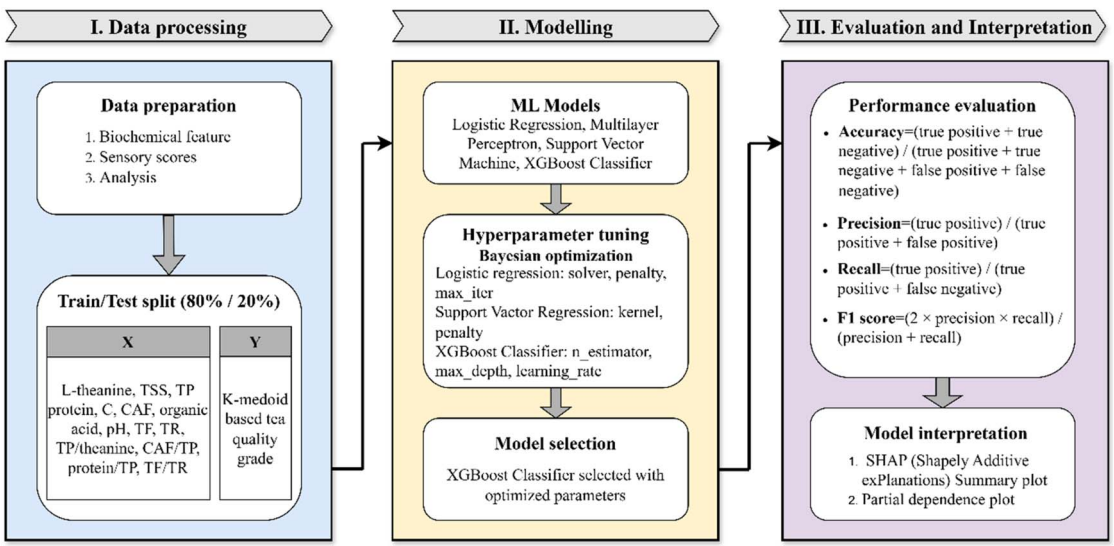

Fig. 1 Flow diagram depicting the methodology of dataset building and modelling for tea quality grade prediction.

map the output of a linear model to the probability of a specific category. SVM is suitable for performing qualitative classification problems by mapping data points on an optimal separating hyperplane in a high-dimensional space.[33] MLP is a type of artificial neural network consisting of at least one hidden layer. MLP is suitable for deciphering complex, nonlinear relationships with the capacity to handle noisy and diverse data.

Hyperparameters for each model, such as the number of trees, learning rate and maximum depth of trees for XGB, kernel and penalty parameters for SVM, solver, penalty, and iteration number for LR, were optimized using Bayesian optimization, which iteratively assessed various parameter combinations. Five-fold cross-validation was employed to obtain the final performance metrics for model evaluation. The performance of each model was evaluated based on four aspects: accuracy, precision, recall, and F_1 score. The model with the best performance was chosen for further analysis. The mathematical definitions of these performance metrics are as follows:

$$\text{Accuracy} = \frac{\text{true positive} + \text{true negative}}{\text{true positive} + \text{true negative} + \text{false positive} + \text{false negative}}, \quad (1)$$

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}, \quad (2)$$

where true positives and true negatives are correctly classified gestures, while false positives are misclassified gestures, *i.e.* assigned to an incorrect class.

$$\text{Recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}, \quad (3)$$

where false negatives are the actual gestures classified wrongly as a different gesture.

The F_1 score is the harmonic mean of the precision and recall metrics:

$$F_1 \text{ score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \quad (4)$$

4. Results and discussions

4.1. Sensory analysis and biochemical composition

This study analyzes the relevance of the specific biochemistry that influences the sensory quality of tea blends. Additionally, the relationship between biochemical and sensory factors was captured to predict the sensory grade of various blends. The overall sensory quality of the samples was represented by a continuous score, on which K-medoids clustering was applied to classify the blend samples. A total of thirty samples were classified into two categories of 17 high-grade samples (overall score of 72 to 85) and 13 low-grade samples (overall score of 60–68) (refer to SI).

The violin plots (Fig. 2) present the distribution pattern of L-theanine, protein, TSS, TP, C, CAF, total organic acid, pH, TF, and TR in both grades. The plot width shows the density of data along with a comparative demonstration of data distribution across both grades. The embedded box plot depicts minimum, first quartile (Q1), mean, third quartile (Q3), and maximum values. Points appearing outside the box were designated as "outliers".

There was a difference in the distribution pattern of L-theanine content among the low-grade and high-grade samples (Fig. 2a). Although the mean content of the two groups was similar, the high-grade samples exhibited a wider and higher distribution range. In high-grade samples, L-theanine was in the interquartile range (IQR) of 1.64–2.26 mg g^{-1} dry weight (DW) and an average of 1.96 mg g^{-1} DW. L-Theanine was in the IQR of 1.77–2.16 mg g^{-1} DW for the low-grade samples with an average of 1.95 mg g^{-1} DW. A similar range of 2–5 mg g^{-1} of L-theanine content is noted in commercial tea, and it further varies based on the growing location and processing conditions of the tea.[34,35]

The protein and TSS contents in the high-grade samples were higher than those of the low-grade samples. The IQR for protein was 0.07–0.091 mg g^{-1} DW with an average content of 0.086 mg g^{-1} DW in high-grade samples and 0.066–0.079 mg g^{-1} DW IQR with an average content of 0.072 mg g^{-1} DW in low-grade samples (Fig. 2b). TSS was in the IQR of 2.172–2.877 mg g^{-1} DW with an average of 2.574 mg g^{-1} DW in high-grade samples and IQR of 1.867–2.709 mg g^{-1} DW with an average TSS content of 2.246 mg g^{-1} DW in low-grade samples (Fig. 2c). A higher content of L-theanine, TSS, and protein indicates a fresh, mellow, and sweet taste, which improves the sensory quality of tea. Studies have reported high-grade tea containing more soluble sugar and L-theanine content, which is in line with this study's findings.[4,25] In our study, high-grade samples (samples 21, 23, and 24), which contained higher proportions of white tea, showed L-theanine levels of up to 2–2.7 mg g^{-1}. A higher amount of L-theanine in the high-grade samples must be contributed by amino acid-rich white tea. Similarly, higher TSS and protein were able to counterbalance the aversive taste and ultimately improve the taste score to make the high-grade sample.

There was an opposite trend observed for TP and C content among the samples. TP was in the IQR of 36.3–44.5 mg GAE per g DW (average of 39.9 mg GAE per g) in high-grade samples and 42.4–45.9 mg GAE per g (average of 45 mg GAE per g) in low-grade samples (Fig. 2d). C varied with an IQR of 0.17–0.26% (average of 0.23%) in the low-grade sample and an IQR of 0.12–0.18% with an average of 0.16% in the high-grade sample (Fig. 2e). The astringent and bitter taste of green tea generally comes from its higher polyphenol and C content. Therefore, the higher TP and C in the lower-grade samples were due to the higher proportions of green tea compared to oolong and black tea. Nonetheless, the distribution of C in high-grade samples is not normal and suggests the potential contribution of a higher C quantity by white tea. CAF in low-grade samples was in the IQR of 6.3–8.7% (average of 7.5%) and 6.6–7.8% (average of 7.3%) in high-grade samples. There was not much variation in the CAF content among both the grades (Fig. 2f). CAF content mainly depends on plant maturity and physiological processes. Therefore, this feature is not affected by the blending process

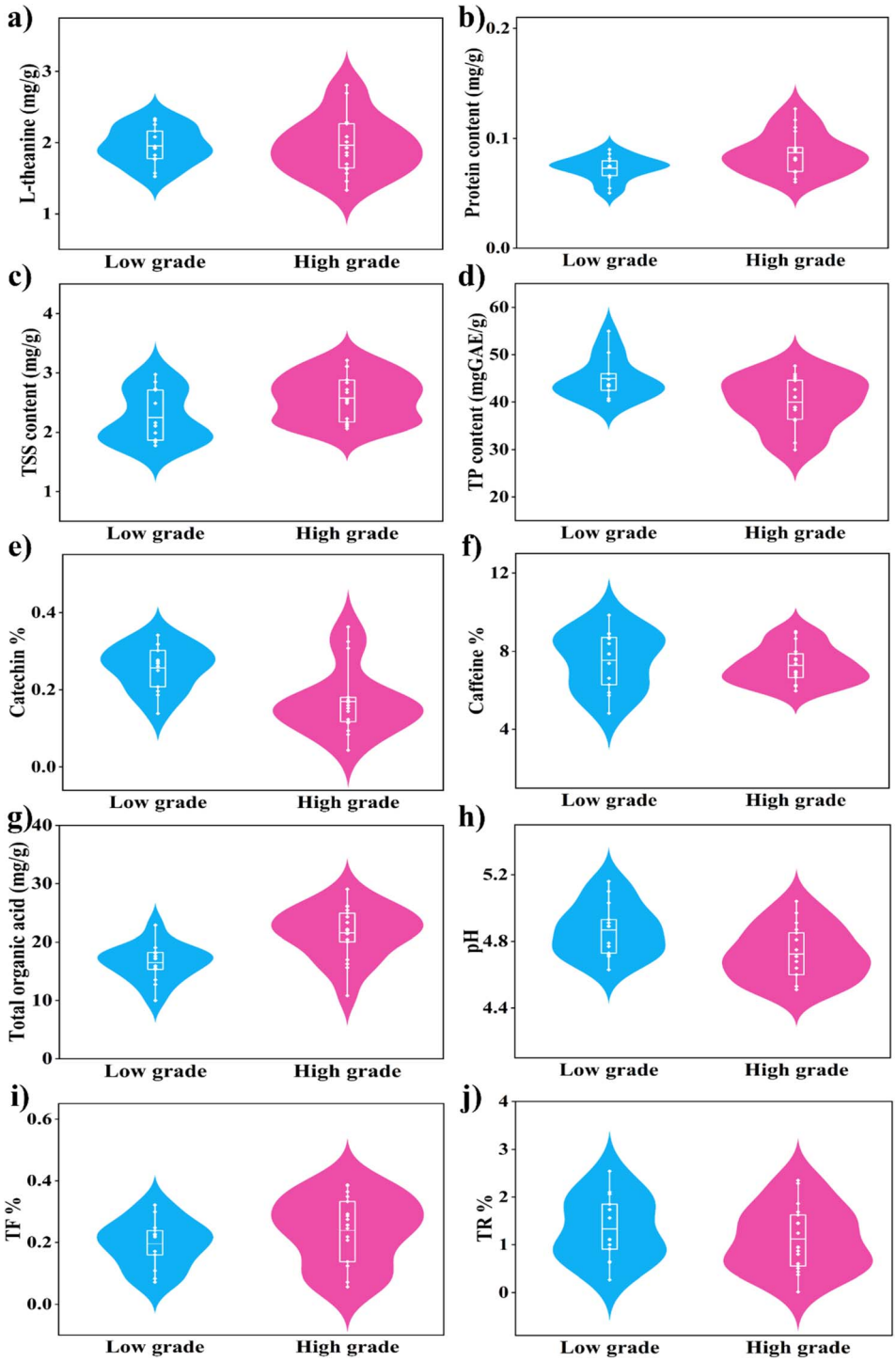

Fig. 2 Violin plots for the comparative demonstration of the biochemical composition of both sample grades. Content of (a) L-theanine, (b) protein, (c) total soluble sugar, (d) total polyphenol, (e) catechin, (f) caffeine, (g) total organic acid, (h) pH, (i) theaflavin, and (j) thearubigin depicted along with their average, interquartile range, and distribution pattern. TSS: total soluble sugar; TP: total polyphenol; TF: theaflavin; TR: thearubigin.

and maintains the ideal range for the tea to have favorable sensory characteristics.

The contents of citric acid, malic acid, ascorbic acid, oxalic acid, gallic acid, and succinic acid were summed to calculate the total organic acid content (Fig. 2g). For the lower grade, total organic acid was in the IQR of 15.3–18.2 mg g^{-1} with an average of 16.4 mg g^{-1}. For high grade, the IQR was 20–24.9 mg g^{-1} with a mean quantity of 21.5 mg g^{-1} DW. The pH was in an IQR of 4.73–4.93 (average of 4.86) for low-grade samples and an IQR of 4.6–4.85 (average of 4.72) for high-grade samples (Fig. 2h). The lower pH in high-grade samples was attributed to their higher organic acid content. Moreover, the pH of tea infusions

depends on the proton concentration of water, which directly measures the sour intensity of the tea infusion.[26] The average quantity of organic acid was present in the order of succinic acid (4.12 mg g^{-1}) > malic acid (3.87 mg g^{-1}) > citric acid (3.53 mg g^{-1}) > oxalic acid (2.91 mg g^{-1}) > ascorbic acid (2.54 mg g^{-1}) > gallic acid (2.35 mg g^{-1}). A similar trend of high succinic acid is found in tea, which provides the sour and light umami note, and citric, oxalic, and malic acids provide sour and gentle astringency to tea infusion.[26]

TF was in the IQR of 0.15–0.23% (average of 0.19%) for low-grade samples and in the IQR of 0.13–0.33% (average 0.23%) for high-grade samples (Fig. 2i). In an opposing trend, TR was in the IQR of 0.9–1.8% (average of 1.3%) for low-grade samples and an IQR of 0.5–1.6% (average of 1.1%) for high-grade samples (Fig. 2j). Depending on processing conditions, TF ranges from 0.5% to 2% of DW and TR ranges from 6% to 18% DW.[27] Contrastingly, TR content was very low in the sample blends in this study.

4.2. ML model selection and performance

Four machine learning models were trained using 19 biochemical features to assess their accuracy in classifying the tea blends into two categories: high grade and low grade. The performance of each model was evaluated based on four performance parameters that are displayed in Fig. 3 for ease of comparison. The XGB model had the highest precision, recall values, and F_1-score of 96%, 78%, and 0.86 for high-grade samples, and 79% precision, 96% recall, and 0.87 F_1-score for low-grade samples, respectively. XGB outperformed all models with an overall accuracy of 87%, followed by SVM with an accuracy of 86%. LR achieved an accuracy of 80%, but its basic structure limits its ability to effectively model non-linear relationships. On the contrary, MLP has a complex structure, and it can only learn meaningful patterns from a large dataset. In this case, MLP had the lowest accuracy of 77%. Clearly, the XGB model emerged as the top performing model with the highest values across all performance parameters and therefore was selected for further analysis.

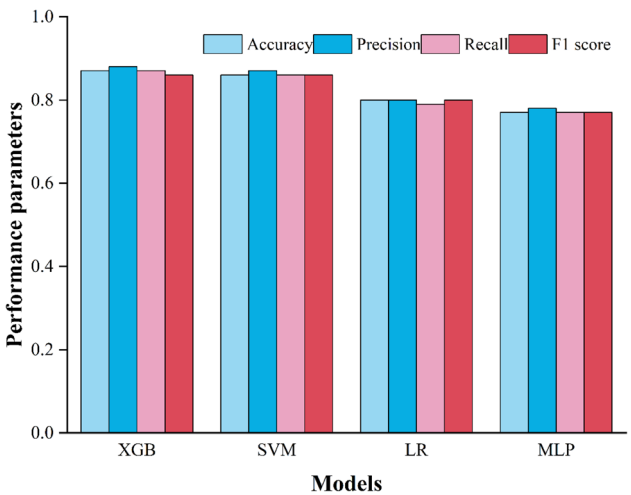

Fig. 3 Comparative demonstration of the performance parameters of the four ML models. The parameters are accuracy, the weighted average of precision, recall score, and F_1 score.

The performance of the models is summarized in Fig. 4 by the confusion matrices. Confusion matrices offer comprehensive insight into the performance of a classification task by comparing the model's predictions with actual values. They reveal whether a model is struggling with a particular class and how often it confuses one class with another. The predicted categories are denoted by columns and the actual categories by rows, as shown in Fig. 4. If a simple 80 : 20 train-test split is used, and a single confusion matrix is generated based on the test data, the matrix heavily depends on the specific test data chosen. This dependency introduces high variance, which can lead to an unreliable estimate of the model's overall performance. To mitigate this issue, a 5-fold cross-validation approach was employed. This produced 5 confusion matrices, which were then averaged and normalized to create a final confusion matrix. This approach reduces the variance that can arise from using a single test set, offering a more accurate reflection of the model's performance.

The SVM model closely mirrored XGB in terms of true positive and true negative, implying their efficiency in correctly predicting both the sample grades. MLP exhibited the highest false positive and false negative rates. The predictive performance for the models for lower-grade samples is higher than for high-grade samples. This indicates the noise and high subjective variations among sample scores, especially in the high-grade category. Moreover, high-grade samples might depend more on other features such as ester-to-non-ester catechin varieties, volatile organic compounds and flavonol glycosides. Inclusion of these features further improves the true positive or predictive performance for high grades. The reason for XGB's superior performance lies in its iterative process, which corrects its node's error at every level. SVM has also been effective in handling this dataset, managing noisy patterns, and identifying the hyperplane that maximizes the margin between these two classes. XGB was further considered for feature interpretation in this study.

4.3. Model interpretation using SHAP

SHAP explains how much each feature in a model contributes to a specific prediction, making complex models (like neural networks or ensemble methods) more transparent and interpretable. It considers all possible combinations of features to measure the average contribution of each feature to the prediction. The SHAP summary plot (Fig. 5) depicts the SHAP value for each feature according to their importance in the prediction of high-grade samples, where the most important feature is at the top, the second most important at the second place and so on. The points in the summary plot represent the Shapley values for a feature and an instance. Each point's color represents the value of a feature, ranging from low to high. Low values are represented by blue, while high values are represented by red. This color coding helps to illustrate whether a predicted class is influenced by a feature's low or high value. The location of these points along the *X*-axis signifies the

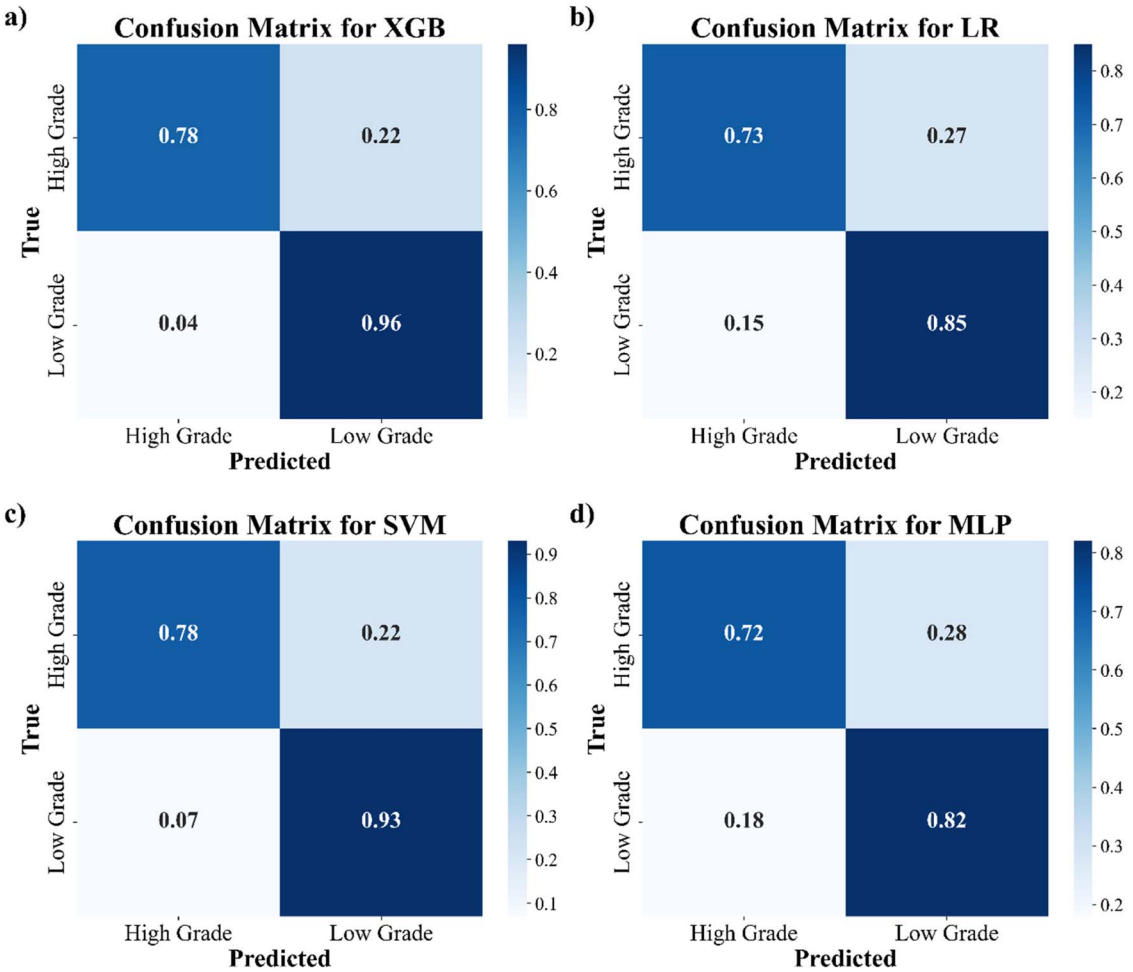

Fig. 4 Comparative confusion matrices for (a) XGB, (b) LR, (c) SVM, and (d) MLP in tea grade classification. The *X*-axis of every matrix denotes the true categories, and the *Y*-axis denotes the predicted categories. The number in each cell displays the normalized count of predictions that refers to the sample proportion for each category. The colorbar on the right side of the matrix displays the color shade of each cell depending on the value inside.

Shapley value of a feature, with a higher positive value indicating a greater influence on predicting high-grade tea.

The SHAP value in Fig. 5 indicates that protein/TP, TSS, citric acid, ascorbic acid, oxalic acid, gallic acid, CAF, CAF/TP, and TF/TR were positively related to the model prediction of high grade. An elevated level of these features, denoted by red color, corresponds to a high-grade sample, as evidenced by a positive SHAP value. Protein/TP ratio is an important feature that demonstrated a positive relation with tea sensory score, as depicted by the SHAP summary plot (Fig. 5). A higher protein/TP means more protein content, which can eventually bind with polyphenols to alter their structure and influence how they interact with taste receptors.[36] The mode of interaction between protein and polyphenol depends on several factors, such as temperature, their structure, and, most importantly, the protein/TP ratio. Polyphenols, such as EGCG, saturate the interaction sites in proline-rich proteins by binding when the protein/TP ratio is low. Then, polyphenols bridge the saturated soluble protein molecules to form a colloid; finally, haze formation takes place to negatively influence tea appearance.[37] It is important to note that these interactions are highly dependent on several abiotic factors, including tea brewing time, infusion temperature, pH, and ionic strength, whose interactive effect makes it a complex and dynamic system requiring in-depth exploration to explain their influence on tea sensory phenomena.[4,38] TSS is a vital component that adds sweetness to tea infusion.[39] Moreover, a variety of soluble sugars continuously interact and form complexes with other biochemicals to develop the unique taste of tea. This is the reason why TSS was observed to have a positive relation with sensory scores in our model.

Organic acids showed a positive association with the sensory quality of tea. Succinic acid and gallic acid have been reported as umami-enhancing constituents in tea liquor, where they can intensify the umami perception of amino acids.[40] In addition, citric, ascorbic, malic, and succinic acids act as natural antioxidants that can decrease the pH of the infusion and limit the formation of hydrogen peroxide, thereby contributing to an overall improvement in the taste profile of the tea liquor.[41] The extent of influence of different organic acids on the sensory of tea blends was citric acid > malic acid > ascorbic acid > oxalic acid > gallic acid > succinic acid.

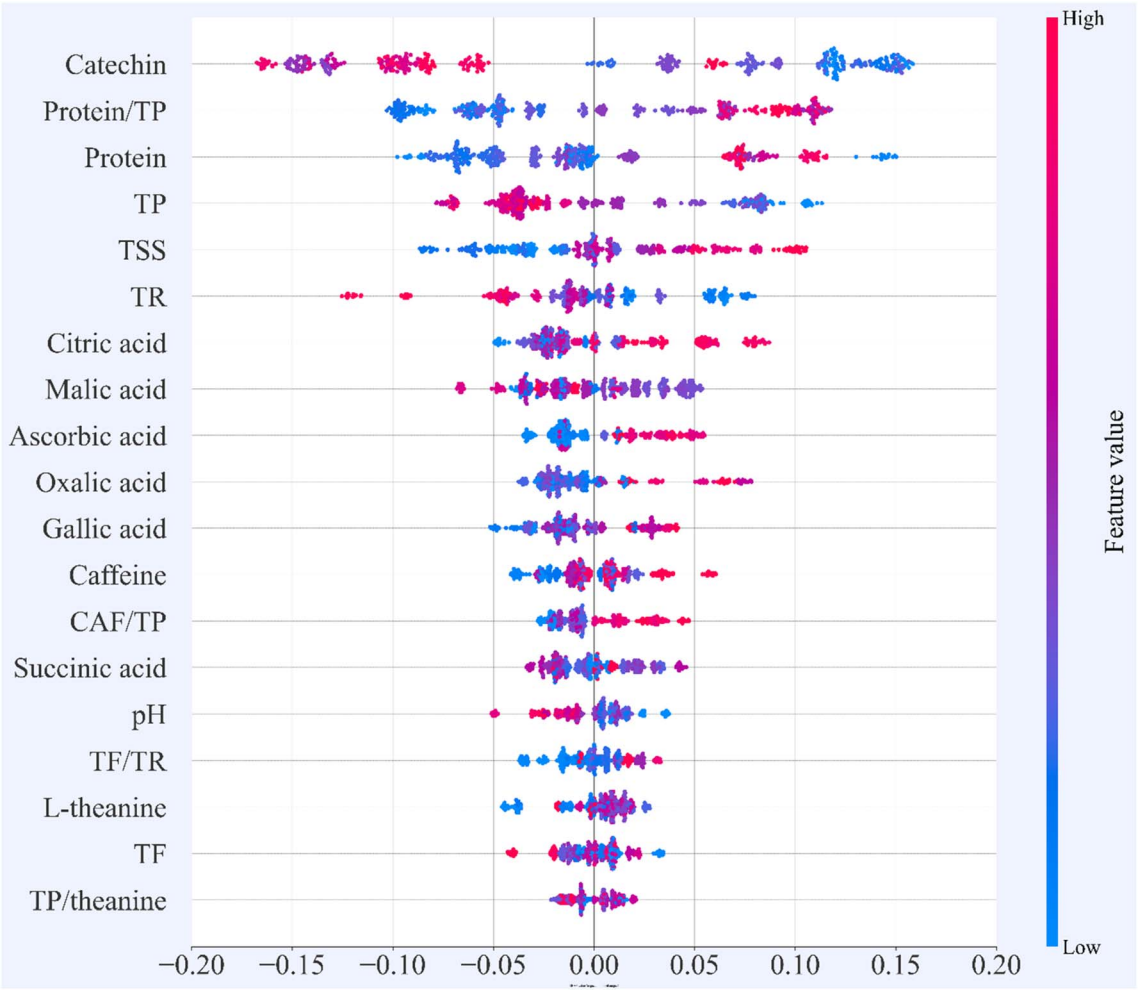

Fig. 5 SHAP global summary plot with feature contribution based on the XGB model. The features are arranged vertically in the order of significance, with the *y*-axis displaying their importance. The location of the data points along the *x*-axis reflects the degree of their influence on the model. A value of zero on the *x*-axis signifies no effect, while positive values that extend to the right and negative values that extend to the left indicate supportive and obstructive effects, respectively. Additionally, the color of each data point for a given feature indicates its level, whether high or low.

CAF is an alkaloid that imparts bitterness and was found to improve the sensory quality of tea blends. The fact that white tea contains more CAF[42] supports our observation of a positive correlation between CAF and high-grade samples that had higher proportions of white tea. CAF and TP are among the principal metabolites that form the distinct sensory characteristics of tea. Their ideal ratio is crucial to a balanced note of both bitterness and astringency in high-grade tea. Caffeine forms complexes with polyphenols through hydrogen bonding. This interaction prevents polyphenols from binding to salivary proteins, ultimately enhancing the overall taste by improving umami flavors and reducing bitterness.[43] Given that green tea contains more TP and black tea has more CAF, adequate proportions of both green and black tea can improve sensory quality by contributing a balanced CAF/TP ratio in a blend. The relationships among TF, TR, their ratio, and tea sensory quality are well established. TF influences not only the appearance and taste of tea but is also a critical determinant of market value.[27] CAF in combination with TF contributes to the characteristic briskness of tea liquor.[44] The samples with high sensory scores had a positive relationship with the TF/TR ratio, which is in line with several prior findings.[27,31,45]

The SHAP summary plot depicts C, TP, TR, and pH as negatively related features, and C as the most important feature in the classification of tea grades. This is because C is one of the primary compounds in tea that contributes to its intense and aversive taste. A high-grade sample should have a balanced C to amino acid ratio so that the unpleasant taste from C can be balanced by umami or sweet notes.[30] It is also clear from Fig. 5 that higher TP content corresponds to a negative SHAP value, *i.e.* low-grade sample. TP had a prominent effect on model output, and Shapley interprets this feature as having a depleting effect on the sensory quality of tea blends. TR, as a highly polymerized pigment, is primarily responsible for the depth and intensity of liquor color. The total TR content was found to have an opposing relationship with sensory quality, which is in accord with the findings of Ngure.[46] In addition, the high-grade

sample had a lower pH compared to the low-grade ones due to their higher organic acid content.

The relation of protein, malic acid, succinic acid, L-theanine, TF, and TP/theanine with sensory was not well-defined in the SHAP summary plot. These features demonstrated low and high values along both positive and negative directions along the *X*-axis. Generally, protein is related to the good sensory characteristics of tea infusion. However, high protein can negatively affect tea's appearance at high temperatures.[47] This could be attributed to the low feature value depicted in Fig. 5 at the extreme right. Higher contents of citric acid, ascorbic acid, oxalic acid, and gallic acid were observed in the higher grade sample. However, malic acid and succinic acid demonstrated a complex relationship with a mix of high and low feature values corresponding to low grade (Fig. 5). The concentration of malic acid increases as tea freshness decreases.[48] This explains the high value of this feature associated with the model prediction of low grade due to a decrease in freshness. Additionally, a study has reported a decrease in malic acid levels by 85.8% and an increase in succinic acid content by 8.42-fold after fermentation.[49] Hence, blends that involve black and oolong teas tend to have low malic acid but higher succinic acid content. Usually, succinic acid improves tea taste by imparting sourness and enhancing umami notes; however, different proportions of tea varieties contribute to varying levels of succinic acid, with no direct relationship to the sensory quality of tea blends.

L-Theanine accounts for 40–70% of the total free amino acids in tea and is a primary contributor to its freshness and umami taste.[50] Besides flavor, L-theanine acts as a crucial precursor for aroma formation during tea processing, making it a key quality indicator for premium teas.[51] Minimal variations were found in the L-theanine content among the tea blends, which suggests that tea processing conditions have little impact on this metabolite. However, factors such as growing location and conditions could affect the synthesis and accumulation of L-theanine in tea leaves.[50] Such confounding factors and the limited variation in L-theanine content across tea blends make it difficult for the model to clearly visualize its influence in the SHAP summary plot.

TF content did not emerge as a significant predictor of blend quality. TF is a reddish-yellow and bright red pigment, which mainly contributes to liquor strength, brightness, and the characteristic golden ring.[44] Different proportions of TF fractions (for example, galloyl TF) with different astringency levels are believed to form the basis of variation in the perception of astringency and umami notes among tea varieties.[52] In this study, high TF feature values for blends predicted as high grade most likely reflect their richer appearance and higher brightness although TF simultaneously hampered taste due to its low astringent threshold.[31]

High amino acid content and a lower TP level correspond to less bitterness and therefore a favorable taste although the relationship is highly non-linear. TP to theanine ratio has been linked to a high sensory score with fresh, soft, and more brisk infusion characteristics. Nonetheless, the TP/theanine ratio demonstrated a mixed nature with sensory quality in this study. This shows that the proportions of TP and L-theanine do not directly influence the sensory quality of tea. There can be other interacting features, such as acidic glycoproteins; flavonols, such as kaempferol, myricetin, and quercetin; alkaloids, such as theobromine and theophylline; and volatile compounds, that were not considered in this study. In general, numerous studies have correlated a low TP/theanine ratio to favorable tea infusion characteristics, such as freshness and mellowness.[4,29,30] Besides being a taste indicator of tea infusion, TP/theanine content is also a crucial parameter in judging tea variety.[53] Therefore, the TP/theanine ratio can be used as an indicator of the proportions of tea varieties in a blend.

SHAP can reveal only the extent to which each feature contributes to the model's predictions. However, this may not always clarify the true causes of specific outcomes. For instance, if a model misinterprets the effect of TP on tea sensory attributes due to confounding variables, like TSS (given that TP and TSS are complex positively influenced, while TP has a negative relation with sensory scores), a SHAP summary plot might misleadingly suggest that higher TP levels are linked to better sensory scores. This finding contradicts the results of experimental studies. Therefore, it is essential to combine SHAP's interpretations with experimental knowledge because SHAP alone is not suitable for identifying the actual causes behind specific events.

Further elaboration on the relation between biochemical features and tea grade has been done by the SHAP dependence plot (Fig. 6) with the top nine features. The reason for focusing only on the top-ranked nine features is to gain clear insight into the model behavior. It also helps center the narratives around the features that have the highest relevance to tea sensory and avoids less relevant details. The SHAP dependence plot displays feature values on the *X*-axis and their corresponding SHAP values on the *Y*-axis for various data points. This approach enables the interpretation of the importance of a feature and its interactions with other features as their values change. This technique captures the actual scenario of how changes in the feature value are related to the model's prediction. The SHAP dependence plot visualizes how a feature affects the outcome and how this effect can change depending on the context and interactions with other features. The scatter plots were color coded as red for positive outcome, *i.e.* high-grade classification and blue for negative outcome *i.e.* low-grade classification. This highlights the influence of a particular feature on the SHAP value of a target outcome. The red trendline depicting data fit is based on third order polynomial regression. The density and sparseness of the plotted histograms represent the accuracy of this analysis, with dense areas depicting a more accurate prediction. Histograms of the *X*-axis and *Y*-axis are on the top and right sides of the diagram, respectively.

Tea catechin showed a complex relation with sensory score as confirmed by the SHAP dependence plot (Fig. 6a). Non-ester type catechins, such as +(−) C, contribute less bitterness and improve tea flavor compared to ester type catechins, which are abundant in green tea.[30] The vertical spread in the values of C in the dependence plot suggested an interactive effect with polysaccharide, protein and other biochemicals in tea. Furthermore, an increase in C content after 0.8 (Fig. 6a) indicated an

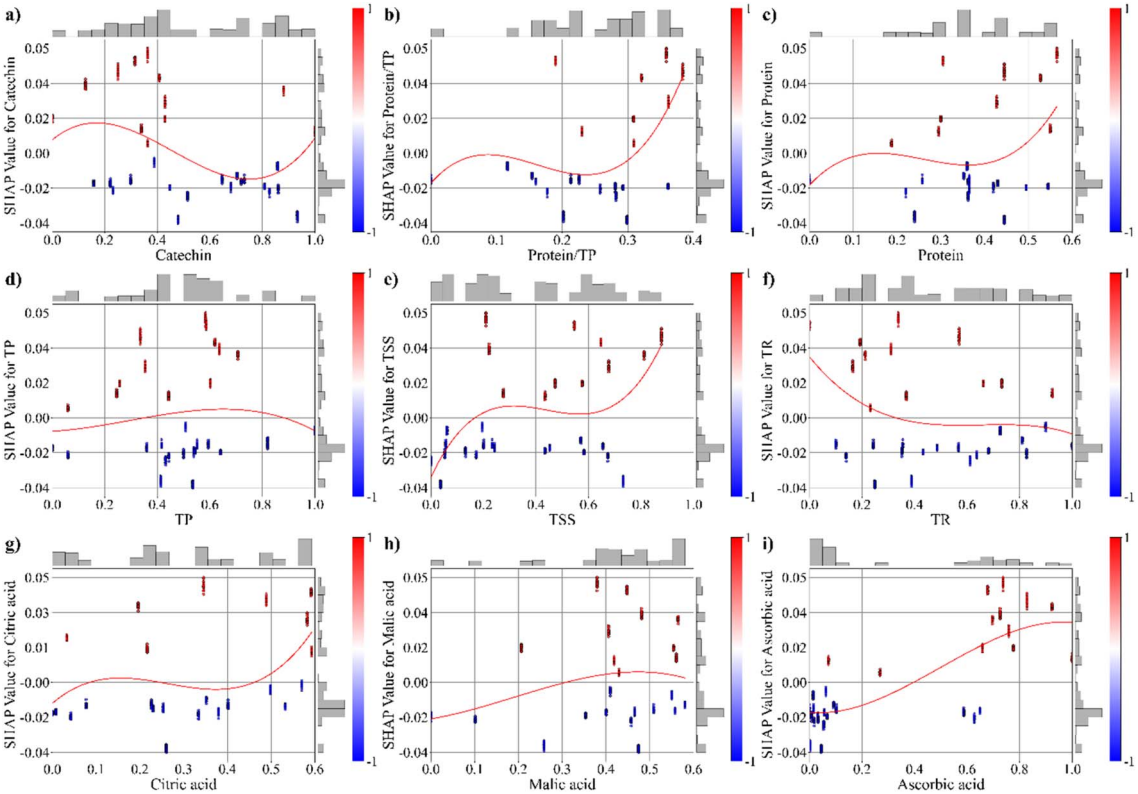

Fig. 6 SHAP dependence plot of 9 important features: (a) catechin, (b) protein/TP, (c) protein, (d) TP, (e) TSS, (f) TR, (g) citric acid, (h) malic acid, and (i) ascorbic acid. The *X*-axis displays the feature values, while the *Y*-axis represents the corresponding SHAP values. The dots in the scatter plot visually map the connection between the respective features and the model's predictions of high-grade (red dots) and low-grade (blue dots) samples. The red trendline is based on 3rd order polynomial regression.

interesting observation, which was not clear from the SHAP summary plot (Fig. 5). This could be attributed to several reasons, such as a high non-ester-to-ester catechin ratio, tea particle size, and high water extract. High-grade samples in this study had a low ratio of green to white/black tea varieties, which implies a balanced ratio of esterified to non-esterified catechin, and therefore a higher sensory score with a pleasant aftertaste. Studies have suggested that tea processing conditions, such as hot air drying and drum rolling at low temperatures, induce the isomerization of EGC, EC, ECG, and gallic acid by polyphenol oxidase and other hydrolases to convert *cis*-catechins to *trans*-catechins to ultimately form CG, C, and GCG, resulting in a lower bitterness index.[54] This could be a reason for the high-grade sample to have high C and a lower esterified-catechin to simple-catechin ratio, which promotes its mellow and brisk taste. In a recent attempt, EGC content in tea blends was found to be increased with a smaller particle size (100–120 μm) although it induced less bitterness compared to the whole leaf tea.[55] In our study, blending was done after grinding and sieving the tea samples through a mesh size of 150 μm. These particles inside the tea bag rapidly release C, improving its content without compromising its taste. The rise in C after 0.8 (Fig. 6a) can also be attributed to its rapid release along with other water-soluble carbohydrates and theanine from smaller-sized particles. The resultant water extract could have synergistically helped mask bitterness and aversive taste from the total catechin.[56]

High protein/TP was related to a higher sensory index. Dependence plot shows a slight decrease throughout 0.1–0.25, followed by a sharp rise in sensory score along with protein/TP value, as depicted in Fig. 6b. The ratio of protein to polyphenol in tea dictates the nature of the interaction between these two compounds and their subsequent effects on sensory properties. Studies have suggested that polyphenols form bonds with proteins through multisite and multidentate interactions, resulting in less TP content.[57] EGCG and ECG are the main tea polyphenols that induce astringency by reducing the lubrication of salivary proteins by interacting with them.[36] However, externally added proteins can alter such polyphenolic properties to reduce bitterness and astringency in tea. The concentrations of EGCG and ECG were found to be reduced by 70–80% in the presence of proteins, such as casein and albumin.[57] Therefore, a higher ratio of protein to TP in tea overpowers the puckering sensation of polyphenols and improves the overall sensory quality, as depicted in Fig. 6b. However, excess protein in tea can lead to cream formation and negatively impact the appearance of tea at higher temperatures.[47]

Protein also exhibited a positive relation with tea sensory, except for a slight fall across the 0.2–0.4 range (Fig. 6c). Proteins containing L-amino acid residues, such as Val–Phe and Val–Gly–Val, impart umami and sweet taste.[58] The kokumi peptides contain γ-glutamyl peptides, which are known to intensify sweet and umami taste.[59] The umami taste of processed tea is

mainly attributed to the presence of pyroglutamic acid, pyroglutamyl peptides, and some other peptides, particularly aspartic acid, serine, threonine, and α-glutamyl-di and tri peptides.[60] Besides taste improvement, proline and hydroxyproline in tea also participate in Stecker degradation to form pyrrole, which provides tea with a good flavor and aroma.[61] Studies have also mentioned polyphenol–protein–polysaccharide interactions that can significantly influence the sensory, functional, and nutritional properties of the food system.[62]

Flavonoids represent the major polyphenol class that contributes to the sensory quality of tea. Flavonoids can be subclassed into several categories, of which flavanols are primarily catechins.[3] Catechins form the bulk of the TP in tea and provide the characteristic astringency and bitterness of tea infusions. The negative effect of TP on the sensory quality of tea blends is apparent from the SHAP summary plot (Fig. 5) and dependence plot (Fig. 6d). The vertical spread of SHAP values throughout 0.4–0.6 shown in Fig. 6d indicates model prediction of high grade and low grade at similar TP values. This underscores the complex relationship that tea sensory has with TP content. TP alone negatively affects tea sensory, but the complex formation between TP and TSS or polysaccharide-polyphenol conjugates helps develop a more rounded, mellow, and thicker taste profile by masking sharp and harsh taste.[63] Beyond catechins, TF and TR are another class of polyphenols that affect the appearance and taste of tea blends. Tea polyphenols also include phenolic acids, hydrolysable tannins, flavonols and their glycosyl derivatives, which impart fresh sour and bitter notes, colour stability, and general strength to the tea liquor.[3]

TSSs, such as monosaccharides and disaccharides, are the key compounds that are closely related to sweetness and therefore correlate to a higher sensory score (Fig. 6e). TSSs also reduce bitterness and astringency perception of catechins, alkaloids, and flavonol glycosides by increasing the best estimate threshold for those compounds.[64] Tea leaves naturally contain glucose, fructose, maltose, oligofructose and other soluble sugars whose content and composition change during tea processing.[5] Therefore, differently processed tea can contribute to varying TSS levels to enhance the overall sweetness and mellow flavor of tea blends.

TR showed an inverse relation with sensory quality (Fig. 6f). TR is responsible for the mouthfeel of tea liquor, but a high content of TR can decrease the brightness and taste of tea infusion.[27] The percentage of TR in tea blends generally varied from 0.5% to 1.8% depending on the ratio of tea varieties, pH, and drying temperatures, with a stronger dependence on the conditions of fermentation time and temperature of black and oolong tea in blends. TR mainly consists of the high-molecular-weight oxidation products of catechins, which are oligomers and polymers. Beyond that, the degradation products of theacitrin-like intermediate compounds may also contribute to the color of tea infusion.[31] It has been reported that both TR and TF are positively related to the aftertaste of astringency, while TR is negatively related to bitterness.[65] On cooling, the interaction between TF, TR, and CAF may form cream, thereby causing discoloration, precipitation, and loss of tea liquor stability, which seriously affects visual appeal, flavor, and color.[31] This further substantiates the findings of this study: TR has a negative relation with sensory quality.

Citric acid (Fig. 6g) and ascorbic acid (Fig. 6i) exhibited a positive association with the sensory quality of the tea blends. Malic acid also displayed a direct link with the SHAP score but slightly fell after 0.5 (Fig. 6h). Citric acid and ascorbic acid are established contributors to the sour taste of tea infusions. Moderate levels of sour compounds can enhance taste fullness, while excessive concentrations may have a negative effect on the perceived quality.[48] The SHAP analysis also showed a generally positive contribution of organic acids to sensory quality. Although malic and succinic acids had a mixed effect in the SHAP summary plot (Fig. 5), some studies have mentioned that succinic acid, along with gallic acid, can enhance the umami taste contributed by amino acids.[40] Additionally, citric, malic, ascorbic, and oxalic acids have been documented to positively contribute to taste while being negatively associated with turbidity and cream formation in tea liquor.[66] The citric acid curve displayed a plateau across the 0.1–0.4 range (Fig. 6g), suggesting potential interactive effects with other biochemical constituents. Up to a certain concentration, citric and malic acids interact to enhance the perceived sweetness of sugars, and sweet components suppress the initial perception of sourness from organic acids and reduce sourness sensitivity.[67] This complex interplay between organic acids and soluble sugars can be used as an effective strategy for optimizing the sensory characteristics of tea blends. It is important to note that the highly spread and deviated (in the *y* direction) SHAP values for all features suggest that these biochemical compounds do not independently influence sensory characteristics. One possible explanation for this is the varying levels of these metabolites found in the different tea varieties used in the blends. These varieties differ in terms of processing methods, raw materials, and biochemical composition, which allows blends to have the same C, protein, or TSS levels but different sensory scores. This observation also suggests probable noise and instrumental errors during data acquisition. Furthermore, the TeaBioSens dataset is not comprehensive, and the narrow gap in sensory scores between the two grades may have restricted the ability of the XGBoost model to accurately capture the true relationship.

Our study specifically focuses on tea blends using four varieties of Assam tea, but our methodology of combining experimental data with interpretable ML can be easily adapted to other tea blends. The underlying framework can capture complex, non-linear interactions between biochemical content and sensory scores, which are similar across different tea blends. However, it depends on the availability of sufficient experimental data, which is essential to train ML models reliably. This is why we emphasize the need for community datasets to enable the generalization of predictive models across any tea blends.

5. Conclusions

In this study, we introduced a novel framework for characterizing the biochemical-sensory relationships of tea blends

by utilizing our dataset, a robust ML algorithm, and an interpretable ML pipeline designed to investigate the relative and absolute effects of various biochemical features on the sensory quality of tea blends. This study has uniquely integrated ST-CATA sensory assessment and a state-of-the-art explainable ML approach to elaborate on the complex relationship between biochemical compounds and the sensory quality of tea blends.

The XGB model successfully captured the multivariate relationships between experimental biochemical features, calculated features, and the sensory scores of tea blends. The XGB model was selected for its superior performance metrics, while the relative importance of input features was determined by SHAP, a game-theoretic methodology that provides interpretability for any ML model output. The top nine important features influencing the sensory quality were C, protein/TP, protein, TP, TSS, TR, citric acid, and ascorbic acid contents, with all having higher and proportionally increasing impacts on sensory scores except C, TP, and TR.

A limitation of the TeaBioSens dataset is its small sample size and non-inclusion of failed experiments and certain features, such as volatile organic compounds, catechin subtypes, and the ratio of esterified to non-esterified catechin. These might contribute to the challenge of obtaining conclusive insights into the effects of features, like succinic acid, L-theanine, TF, and TP/theanine, on sensory scores through SHAP analysis. Presently, the dataset has been supplemented with data from our experimental findings. Further extension of the dataset by including factors such as harvesting season, processing parameters and other secondary metabolites can shed more light on the feature's complex role in tea sensory modulation. This article encourages industry professionals and researchers to evaluate the effectiveness of TeaBioSens and add more features to this public dataset to further strengthen and generalize the model predictability for other tea blends. Considering the importance of consumer preferences and acceptance in the tea market, engaging a semi-trained consumer panel using the Check-All-That-Apply method allows for practical data collection from the consumer's perspective. Inclusion of more volunteers in ST-CATA analysis will overcome the limitations of this small dataset, allow ML models to better learn the subjective variations in sensory scores, and further improve their predictive performance. This method eliminates the need for professional tea tasters to evaluate every tea grade and batch, significantly reducing costs for the tea industry.

Author contributions

K. D.: conceptualization, supervision, resources, review and editing. O. S.: conceptualization, formal analysis, data curation, methodologies, writing of original drafts, review and editing.

Conflicts of interest

The authors declare that there are no competing interests.

Data availability

Data for this article, including sample, biochemical features, and sensory score are available at zenodo at https://doi.org/10.5281/zenodo.17921536.

Supplementary information (SI): "Tea Biochemical & Sensory Dataset (TeaBioSens)", is a meticulously curated compilation of data including important biochemical features that influences tea sensory quality. This dataset consists thirty different tea blends, made from four major processed tea variety, their biochemical content, and their sensory score obtained from a semi-trained panel using the lexicon based quantitative descriptive technique. TeaBioSens comprises a total of 600 data points. Features such as TSS, protein, TP, CAF, (+)-C, TF, TR, pH, citric acid, malic acid, ascorbic acid, oxalic acid, gallic acid, succinic acid, and L-theanine were directly estimated from experiments, and ratio of TP/theanine, TF/TR, protein/TP, and CAF/TP were calculated. See DOI: https://doi.org/10.1039/d5fb00580a.

Acknowledgements

The authors thankfully acknowledge IIT Delhi and Central Research facility (IIT Delhi) for UPLC analysis.

References

1 TRA, *TRA Vision 2030*, 2024.
2 S. Thakur, P. Kumar and N. Gupta, *J. Food Compos. Anal.*, 2025, **144**, 107683.
3 X.-Q. Zheng, Y. Nie, Y. Gao, B. Huang, J.-H. Ye, J.-L. Lu and Y.-R. Liang, *J. Food Compos. Anal.*, 2018, **67**, 29–37.
4 Y. Zhang, X. Chen, D. Chen, L. Zhu, G. Wang and Z. Chen, *Food Res. Int.*, 2025, **203**, 115796.
5 W. Li, Z. Zhang, R. Chen, L. Sun, X. Lai, Q. Li, M. Hao, S. Zhang, Q. Li, S. Sun and Z. Chen, *Food Funct.*, 2025, **16**, 3707–3720.
6 R. C. Gogoi, *Two Bud*, 2014, **61**, 53–56.
7 Z. Xia, Q. Zhou, S. Yang, F. Song, Z. Li, J. Wang, C. Ling and C. Song, *Food Res. Int.*, 2025, **202**, 115563.
8 C. Ling, L. Huang, Y. Bian, X. Lu, Y. Lin, Q. Zhou, F. Song, Z. Li, J. Teng and C. Song, *Infrared Phys. Technol.*, 2025, **146**, 105739.
9 J. Tie, W. Chen, C. Sun, T. Mao and G. Xing, *Clust. Comput.*, 2019, **22**, 6059–6068.
10 S. S. Turgut, J. A. Entrenas, E. Taşkın, A. Garrido-Varo and D. Pérez-Marín, *Food Control*, 2022, **142**, 109260.
11 W. Chen, J. Zan, L. Yan, H. Yuan, P. Wang, Y. Jiang and H. Zhu, *Foods*, 2025, **14**, 941.
12 X. Gong, L. Li, L. Qin, Y. Huang, Y. Ye, M. Wang, Y. Wang, Y. Xu, F. Luo and H. Mei, *Forests*, 2022, **13**, 1629.
13 M. M. Bradford, *Anal. Biochem.*, 1976, **72**, 248–254.
14 L. Lu, L. Wang, R. Liu, Y. Zhang, X. Zheng, J. Lu, X. Wang and J. Ye, *Food Chem.*, 2024, **441**, 138341.
15 FSSAI, *FSSAI Manual of Methods of Analysis of Foods-Beverages: Tea, Coffee and Chicory - Reg*, 2023.

16 H. Deka, T. Barman, J. Dutta, A. Devi, P. Tamuly, R. Kumar Paul and T. Karak, *J. Food Compos. Anal.*, 2021, **96**, 103684.
17 M. G. Narayanappa, H. Kaipa, A. Chinapolaiah, K. Upreti, A. P. M. Gowda, D. C. Manjunathagowda, H. H. Venkatachalapathi, S. H. Shekharappa and L. A. Narayanashetty, *3 Biotech*, 2024, **14**, 65.
18 H. Zou, T. Lan, Y. Jiang, X.-L. Yu and H. Yuan, *Foods*, 2024, **13**, 3718.
19 A. Parveen, C.-Y. Qin, F. Zhou, G. Lai, P. Long, M. Zhu, J. Ke and L. Zhang, *Horticulturae*, 2023, **9**, 1253.
20 S. Tongsai, K. Jangchud, A. Jangchud, B. Tepsongkroh, S. Boonbumrung and W. Prinyawiwatkul, *Int. J. Food Sci. Technol.*, 2022, **57**, 3116–3127.
21 N. Alexi, E. Nanou, O. Lazo, L. Guerrero, K. Grigorakis and D. V. Byrne, *Food Qual. Prefer.*, 2018, **64**, 11–20.
22 ISO6658, *Sensory Analysis—Methodology—General Guidance*, 2017.
23 *Sensory Analysis—Selection and Training of Sensory Assessors*, 2023.
24 Y. Xiong, H. Liao, H. Liao, X. Song, C. Ma and Y. Huang, *Foods*, 2025, **14**, 1552.
25 M. Sun, F. Yang, W. Hou, S. Jiang, R. Yang, W. Zhang, M. Chen, Y. Yan, Y. Tian and H. Yuan, *Molecules*, 2022, **27**, 3562.
26 X. Zhang, X. Du, Y. Li, C. Nie, C. Wang, J. Bian and F. Luo, *Food Sci. Nutr.*, 2022, **10**, 2071–2081.
27 A. Ghosh, B. Tudu, P. Tamuly, N. Bhattacharyya and R. Bandyopadhyay, *Chemom. Intell. Lab. Syst.*, 2012, **116**, 57–66.
28 J. Moreira, J. Aryal, L. Guidry, A. Adhikari, Y. Chen, S. Sriwattana and W. Prinyawiwatkul, *Foods*, 2024, **13**, 3580.
29 Y.-N. Zhang, J.-F. Yin, J.-X. Chen, F. Wang, Q.-Z. Du, Y.-W. Jiang and Y.-Q. Xu, *Food Chem.*, 2016, **192**, 470–476.
30 Y.-Q. Xu, Y.-N. Zhang, J.-X. Chen, F. Wang, Q.-Z. Du and J.-F. Yin, *Food Chem.*, 2018, **258**, 16–24.
31 P. Long, K. Rakariyatham, C.-T. Ho and L. Zhang, *Trends Food Sci. Technol.*, 2023, **133**, 37–48.
32 X. Jiang, X. Cao, Q. Liu, F. Wang, S. Fan, L. Yan, Y. Wei, Y. Chen, G. Yang, B. Xu, Q. Wu, Z. Xu, H. Yang and X. Zhai, *Food Res. Int.*, 2025, **211**, 116455.
33 D. A. Pisner and D. M. Schnyer, in *Machine Learning*, Elsevier, 2020, pp. 101–121.
34 Y. R. Liang, Z. S. Lio, Y. R. Xu and Y. L. Hu, *J. Sci. Food Agric.*, 1990, **53**, 548.
35 R. Thippeswamy, A. Martin and L. R. Gowda, A reverse phase high performance liquid chromatography method for analyzing neurotoxin β-N Oxalyl-L-α, β-diaminopropanonic acid in legume seeds, *Food Chem.*, 2006, 1290–1295.
36 E. Jöbstl, J. O'Connell, J. P. A. Fairclough and M. P. Williamson, *Biomacromolecules*, 2004, **5**, 942–949.
37 P. Bandyopadhyay, A. K. Ghosh and C. Ghosh, *Food Funct.*, 2012, **3**, 592.
38 L. Yang, X. Luo, Q. Wang, M. Liu, J. Yan, C. Wang, Y. Xian, K. Peng, K. Liu and B. Jiang, *Front. Nutr.*, 2025, **12**, 1587413.
39 M. Wong, S. Sirisena and K. Ng, *J. Food Sci.*, 2022, **87**, 1925–1942.
40 S. Kaneko, K. Kumazawa, H. Masuda, A. Henze and T. Hofmann, *J. Agric. Food Chem.*, 2006, **54**, 2688–2694.
41 H. Aoshima and S. Ayabe, *Food Chem.*, 2007, **100**, 350–355.
42 L. Paiva, C. Rego, E. Lima, M. Marcone and J. Baptista, *Antioxidants*, 2021, **10**, 183.
43 Z. Zhou, M. Ou, W. Shen, W. Jin, G. Yang, W. Huang and C. Guo, *Food Chem.*, 2024, **460**, 140753.
44 R. S. Senthil Kumar, N. N. Muraleedharan, S. Murugesan, G. Kottur, M. P. Anand and A. Nishadh, *Food Chem.*, 2011, **129**, 117–124.
45 C. Dong, J. Li, J. Wang, G. Liang, Y. Jiang, H. Yuan, Y. Yang and H. Meng, *Spectrochim. Acta, Part A*, 2018, **205**, 227–234.
46 F. M. Ngure, J. K. Wanyoko, S. M. Mahungu and A. A. Shitandi, *Food Chem.*, 2009, **115**, 8–14.
47 S. Banerjee and J. Chatterjee, *J. Food Sci. Technol.*, 2015, **52**(6), 3158–3168.
48 H. Chen, F. Yu, J. Kang, Q. Li, H. K. Warusawitharana and B. Li, *Molecules*, 2023, **28**, 2339.
49 S. Li, X. Gong, H. L. Zhong, H. Huang and J. Huang, *Mycosystema*, 2014, **33**, 713–718.
50 H. Cheng, J. Zheng, L. Tu, L. Chen, W. He, Y. Wang, Z. Liu and P. Xu, *J. Adv. Res.*, 2025, DOI: 10.1016/j.jare.2025.10.002.
51 X. Guo, C.-T. Ho, W. Schwab, C. Song and X. Wan, *Food Chem.*, 2019, **280**, 73–82.
52 X. Qin, J. Zhou, C. He, L. Qiu, D. Zhang, Z. Yu, Y. Wang, D. Ni and Y. Chen, *Food Chem.: X*, 2023, **19**, 100809.
53 L. Y. Yang, *A record of Chinese clonal tea varieties*, Shanghai Scientific & Technical Publishers, 2014.
54 X. Lu, Y. Lin, Y. Tuo, L. Liu, X. Du, Q. Zhu, Y. Hu, Y. Shi, L. Wu and J. Lin, *Foods*, 2023, **12**, 4334.
55 D. Li, Y. Zhang, R. Tamura, T. Nakajima and Y. Caballero, *Food Nutr. Sci.*, 2023, **14**, 1043–1056.
56 J. Hu, Y. Chen and D. Ni, *LWT–Food Sci. Technol.*, 2012, **45**, 8–12.
57 G. Ziyatdinova, A. Nizamova and H. Budnikov, *Food Anal. Methods*, 2011, **4**, 334–340.
58 J. Xue, P. Liu, G. Guo, W. Wang, J. Zhang, W. Wang, T. Le, J. Yin, D. Ni and H. Jiang, *LWT–Food Sci. Technol.*, 2022, **156**, 113010.
59 M. Kuroda, Y. Kato, J. Yamazaki, Y. Kai, T. Mizukoshi, H. Miyano and Y. Eto, *J. Agric. Food Chem.*, 2012, **60**, 7291–7296.
60 C. J. Zhao, A. Schieber and M. G. Gänzle, *Food Res. Int.*, 2016, **89**, 39–47.
61 P.-C. Kuo, Y.-Y. Lai, Y.-J. Chen, W.-H. Yang and J. T. Tzen, *J. Sci. Food Agric.*, 2011, **91**, 293–301.
62 H. Xue, J. Feng, Y. Tang, X. Wang, J. Tang, X. Cai and H. Zhong, *Chem. Biol. Technol. Agric.*, 2024, **11**, 95.
63 S. Deng, T. Zhang, S. Fan, H. Na, H. Dong, B. Wang, Y. Gao, Y.-Q. Xu and X. Liu, *Food Chem.: X*, 2024, **23**, 101726.
64 H. Oh and M. K. Kim, *J. Sens. Stud.*, 2021, **36**(3), e12653.
65 K. Wang, Q. Chen, Y. Lin, S. Li, H. Lin, J. Huang and L. Zhonghua, *Food Sci. Technol. Res.*, 2014, **20**, 639–646.
66 Y.-Q. Xu, X.-Y. Zhong, J.-F. Yin, H.-B. Yuan, P. Tang and Q.-Z. Du, *Food Chem.*, 2013, **139**, 944–948.
67 R. Zhou, J. Zhu, Y. Li, Y. Hua, Y. Niu, J. Zhang, Z. Xiao and L. Zhao, *Food Biosci.*, 2025, **74**, 107915.