Intelligent Manufacturing and Cloud Computing
I.S. Jesus and K. Wang (Eds.)

doi:10.3233/ATDE241363

Machine Learning-Based Prediction of Flue-Cured Tobacco Quality Using Chemical Composition Analysis: A Case Study in Sichuan Province, China

Jun Qiu[a†], Rui Bie[a,b,e†], Zhiying Wang[a,b], Jingxian Sun[c], Xiaojie Li[c], Dongmei Jin[d] and Jianmin Cao[a*]

[a] *Tobacco Research Institute of CAAS, Qingdao 266101, China*
[b] *Graduate School of CAAS, Beijing 100081, China*
[c] *China Tobacco Shandong Industry Co., Ltd., Jinan 250013, China*
[d] *Sichuan Tobacco Quality Supervision and Testing Station, Chengdu 610041, China*
[e] *HONG TA Liaoning Tobacco Co., LTD, Shenyang 110003, China*
ORCiD ID: Jianmin Cao https://orcid.org/0000-0002-9179-7726

Abstract. Background: In this study, we aimed to establish an efficient and accurate machine -learning model for evaluating tobacco quality based on its chemical composition. To achieve this, we gathered a dataset comprising 188 tobacco samples taken from the production area of Sichuan, China in 2021. Four machine learning algorithms—Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Gradient Boosting Classifier (GBC)—were used to establish a predictive model for assessing tobacco quality. The study focused on comparing the predictive performance of these models and exploring the upper limit of prediction accuracy using genetic algorithm (GA) hyperparameter optimization. Additionally, the SHAP value model interpretation framework was introduced to provide a comprehensive global interpretation and conduct feature dependency analysis. Results: The results showed that model accuracy ranked as RF>GBC> KNN>SVM, with the GA-RF machine learning model achieving a prediction accuracy of 86.8%. SHAP values identified seven important characteristic indexes affecting Sichuan tobacco leaf quality, highlighting RF as the optimal classifier for predicting flue-cured tobacco quality in Sichuan Province. Conclusions: The GA-RF model constructed in this study effectively identifies Sichuan tobacco leaf quality. These findings offer novel insights and data support for the application of machine learning algorithms in tobacco fields and tobacco leaf quality evaluation.

Keywords: Sichuan tobacco, SHAP value, Random forest, Machine learning, Quality evaluation

† These authors contributed equally to this work
* Corresponding Author: Jianmin Cao, caojianmin@caas.cn

1. Introduction

Despite the adverse health effects of tobacco, it remains a vital economic crop worldwide, making the objective evaluation of tobacco leaf quality practically significant. The chemical composition of tobacco influences the taste and aroma of its products[1], and understanding this link is crucial for optimizing tobacco quality. Recent advancements in machine learning have led to its application in various aspects of the tobacco industry, including origin identification [2], grading[3], aroma quality prediction[4, 5]. However, there is a scarcity of studies examining the relationship between chemical composition and tobacco quality using these algorithms.

Machine learning, a key area of artificial intelligence, excels in analyzing multidimensional data without subjective biases[6], providing a more objective reflection of tobacco quality compared to traditional sensory evaluations. Yet, the "black-box" nature of some efficient models poses challenges in understanding prediction paths and identifying key predictors. Thus, the present study aimed to identify and establish a reliable predictive model for the quality of flue-cured tobacco, using four machine learning algorithms, namely, Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Gradient Boosting Classifier (GBC), and analyze the content of eight major chemical components.

In addition, the study aimed to compare the predictive performances of four models and used a genetic algorithm (GA) to optimize the best model's hyperparameters. It introduced the SHAP (SHapley Additive exPlanations) framework to address the models' inherent "black-box" nature, enhancing interpretability and allowing for a more scientific understanding of critical features. This approach improves the alignment between evaluation outcomes and sensory quality, providing innovative insights and robust data support for applying machine learning algorithms in the tobacco industry and assessing tobacco leaf quality.

2. Materials and Methods

2.1. Tobacco samples

Tobacco samples were collected from 94 sites across five major tobacco-producing areas in Sichuan, China, with each site providing one upper (B2F) and one middle (C3F) tobacco sample, totaling 188 samples.

2.2. Sensory quality evaluation

The China National Tobacco Corporation (CNTC) in Sichuan Province enlisted sensory evaluation experts to assess the sensory quality of tobacco, using a 9-point scoring system to calculate total quality scores based on weighted indicators. (Table 1).

Table 1. Indicators of sensory quality and their weights

Aroma Characteristics (0.600)			**Smoke characteristics (0.150)**			**Taste characteristics (0.250)**		
Aroma quality	Aroma quantity	Offensive odor	Fineness	Softness	Roundness	Irritation	Dryness	After-taste
0.250	0.250	0.100	0.050	0.050	0.050	0.100	0.050	0.100

2.3. Chemical composition measurements

Conforming to CNTC tobacco industry standards, measurements were conducted for conventional chemical components (soluble total sugars, reducing sugars, total alkaloids such as nicotine, total nitrogen, starch, proteins, potassium, and chloride), polyphenols, alkaloids, polyhydric acids, higher fatty acids, free amino acids, as well as mono- and disaccharides and polyols, the latter prepared following Ghfar et al.'s method.[7]

2.4. Data processing

2.4.1. Derivatization of chemical composition

This study explores the cross-linked information in tobacco leaves to enhance variable utility and clarify the nonlinear relationship between chemical and sensory qualities. We used automatic feature derivation, combining 13 selected tobacco chemical indicators (e.g., total sugars, alkaloids, nitrogen, starch, proteins) into pairwise combinations to create deeper features. Data were standardized to mg/g and aggregated using three relationships ("divide," "add," "subtract") to generate 312 new features.

2.4.2. Normalization of raw data

Due to variations in dimensions and data distribution disparities, standardization preprocessing was crucial to minimize scale differences' impact on spatial distance measurements in analysis. Z-score normalization transformed features to a mean of 0 and standard deviation of 1.

2.4.3. Model evaluation

A confusion matrix evaluated model prediction accuracy using four metrics: Accuracy, recall, precision, and F1-score, which serve as performance evaluation indicators (Figure 1).

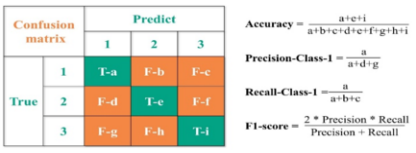

Figure 1. Evaluation principle of the confusion matrix. Lowercase letters in the green region along the diagonal represent true predictions and those in the orange region represent false predictions.

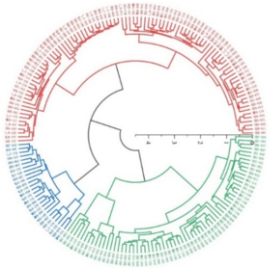

Figure 2. Hierarchical cluster analysis pedigree of 188 samples of Sichuan flue-cured tobacco.

2.4.4 Model construction and interpretation

Four machine learning models, namely RF, SVM, KNN, and GBC, were built using the pycaret 2.3.10 package in Python 3.9.7. To ensure the stability of the models, 5-fold cross-validation was employed. This involved dividing the entire dataset into five equal parts, where each part was used for testing, and the remaining four parts were used for training. This process was repeated five times to ensure that each part had been tested. The average accuracy of the five iterations represented the final accuracy of the models. The GeatPy 2.7.0 toolbox was utilized for model GA hyperparameter optimization[8]. Additionally, the SHAP values were calculated based on the SHAP 0.41.0 package for feature importance analysis.

3. Results and Discussion

3.1. Sample hierarchical clustering

In this study, 188 Sichuan flue-cured tobacco samples were analyzed for sensory quality. To avoid subjective bias in manual grading, hierarchical clustering was used to objectively classify the samples based on sensory multivariate variables. The samples were divided into three grades: grade 1 (good), grade 2 (medium), and grade 3 (general), as shown in Figure 2.

Tobacco leaves of different quality grades were labeled for classification modeling. Table 2 summarizes sensory scores: overall average was 67.80 points. Specifically, 66 samples (35%) were "good" (avg. 70.21), 96 samples (51%) were "medium" (avg. 67.04), and 26 samples(14%) were "ordinary" (avg. 64.49).

Table 2. Clustering information table of 188 samples of Sichuan flue-cured tobacco

Label	Quality grade	Number of samples	Sensory score	Max	Min	CV
1	Grade1 (good)	66	70.21±1.23	73.62	68.68	1.75%
2	Grade2 (medium)	96	67.04±0.89	68.57	65.50	1.32%
3	Grade3 (general)	26	64.49±0.70	65.45	62.67	1.08%
Total		188	67.80±2.20	73.62	62.67	3.25%

3.2. Selection of key quality features

3.2.1 Created feature selection

The method of deriving new features from feature indicators enables rapid and extensive feature construction. While it can generate new features strongly correlated with tobacco leaf quality, it may also produce unimportant or low-content features that intermingle with important or high-content ones, leading to confusion during feature selection. Additionally, irrelevant features may persist, complicating their removal through traditional statistical methods, such as data dimensionality reduction and feature selection—a critical preprocessing step in machine learning. This step effectively reduces redundancy, enhances learning accuracy, and improves result comprehensibility[9]. To address the high-dimensional problem, we first employed the RF algorithm in this study to assess the derived features. This algorithm evaluates each feature's importance for modeling by measuring the average reduction in impurity caused by the feature. Importantly, it is unaffected by inter-feature scale and multicollinearity[10]. Moreover, the algorithm ranks feature importance and provides visual output, reducing interpretation costs and achieving the goal of selecting optimal features.

Figure 3A shows the importance ranking of the division of composite features based on the RF algorithm. The results indicated that 10 indicators, namely starch/total alkaloids (as nicotine), starch/total alkaloids, total mono- and disaccharides/total nitrogen, total mono- and disaccharides/total alkaloids, soluble total sugars/total alkaloids, soluble total sugars/total higher fatty acids, soluble total sugars/total nitrogen, total higher fatty acids/total alkaloids (as nicotine), total alkaloids (as nicotine), reducing sugars, and total nitrogen/proteins are the most important division composite features for predicting the quality gradient in this model.

However, the analysis revealed that, even after feature selection, some features with severe collinearity remained. These features were further reduced to ensure the model's accuracy. For instance, only the starch/total alkaloids (as nicotine) with higher importance were retained, whereas other starch/total alkaloids were discarded. Similarly,

only the total mono- and disaccharides/total nitrogen with higher importance were maintained, and the total mono- and disaccharides/total alkaloids were removed. The same approach was followed for soluble total sugars/total alkaloids and soluble total sugars/total nitrogen, and only those with higher importance were retained.

Figure 3B shows the importance ranking of the addition and subtraction of composite features based on the RF algorithm. The results indicate that 10 indicators, namely soluble total sugars + proteins, starch + reducing sugars, soluble total sugars + reducing sugars, starch-total alkaloids, starch +soluble total sugars, starch + total mono- and disaccharides, total polyhydric acids + soluble total sugars, total polyols + total alkaloids, soluble total sugars, total mono- and disaccharides, and soluble total sugar-starch, are the most important addition and subtraction composite features for predicting the quality gradient in this model. Regarding the three collinear features, starch + reducing sugars, starch + soluble total sugars, and starch + total mono- and disaccharides, only the starch + reducing sugars of higher importance were retained.

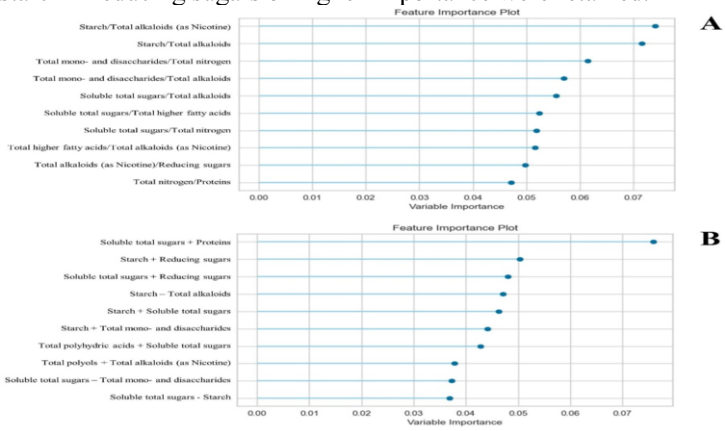

Figure 3. Feature importance plot.

Note: A, feature importance plot of division composite features; feature importance plot ofadditive and subtractive composite features.

Table 3. Feature tool-generated datasets of newly created features

New feature, mg/g	Relation of aggregation	New feature, mg/g	Relation of aggregation
Starch/Total alkaloids (as nicotine)	divide_numeric	Soluble total sugars + Proteins	add_numeric
Total mono- and disaccharides/Total nitrogen	divide_numeric	Starch + Reducing sugars	add_numeric
Total mono- and disaccharides/Total alkaloids	divide_numeric	Soluble total sugars + Reducing sugars	add_numeric
Soluble total sugars/Total higher fatty acids	divide_numeric	Total polyhydric acids + Soluble total sugars	add_numeric
Total higher fatty acids/Total alkaloids (as nicotine)	divide_numeric	Total polyols + Total alkaloids (as nicotine)	add_numeric
Total alkaloids (as nicotine)/Reducing sugars	divide_numeric	Starch - Total alkaloids	subtract_numeric
Total nitrogen/Proteins	divide_numeric	Soluble total sugars - Total mono- and disaccharides	subtract_numeric
		Soluble total sugars - Starch	subtract_numeric

Based on the importance ranking of division and addition-subtraction composite features, a total of seven division composite indicators and eight addition-subtraction

composite indicators were selected as the derived feature dataset for the subsequent modeling process, as shown in Table 3.

3.2.2. Original feature selection

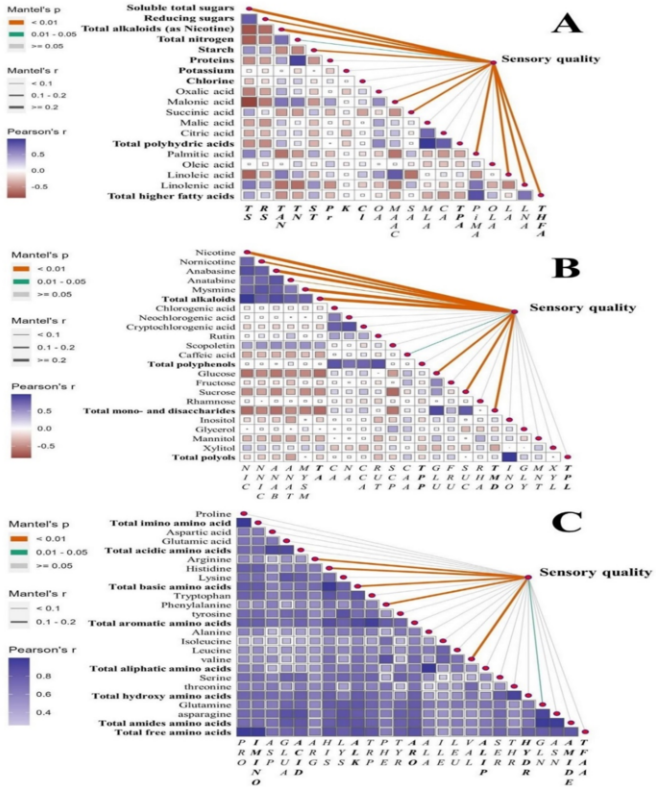

Figure 4. Mantel test of original chemical sensory quality evaluation indexes.

To avoid introducing unnecessary features into the model from the original high-dimensional chemical index dataset, key index features needed to be selected. Unlike Pearson analysis, which quantifies the correlation between two continuous variables individually, the Mantel test treats sensory multivariate variables as one distance matrix and each chemical variable as another matrix. This allows for regression analysis between the two distance matrix forms. Additionally, the R-package LinkET was employed for data visualization, representing bidirectional correlation strength with line thickness and indicating the significance of 999 substitution tests through line segment color. As depicted in Figure 4A, several components displayed significant correlations with sensory quality. These included four conventional chemical components (soluble total sugars, reducing sugars, total alkaloids as nicotine, and starch), two polyhydric acid components (malonic acid and succinic acid), and four higher fatty acid components (palmitic acid, linoleic acid, linolenic acid, and total higher fatty acids). As shown in Figure 4B, except for anatabine and Mantel's r -value, the correlation index between other alkaloid indexes and sensory quality was greater than 0.1. Notably, Mantel's p-value for the total amount of glucose, sucrose, and total mono- and disaccharides in relation to sensory quality was less than 0.01. Conversely, chemical components like polyphenols and polyols exhibited weak correlations with sensory quality and lacked a

significant relationship. In Figure 4C, among the 17 free amino acid indexes and eight total amino acid indexes, three basic amino acid indexes (arginine and histidine), one aromatic amino acid index (phenylalanine), and one aliphatic amino acid index (valine) displayed a close association with sensory quality. In summary, 23 indicators exhibited varying degrees of strong correlation with sensory quality, making them pivotal chemical indicators affecting Sichuan flue-cured tobacco. Consequently, these indicators were incorporated into the feature datasets (Section 2.2.1) and employed alongside their variables in the final entry model. By screening the quality index features, redundancy was eliminated to obtain the optimal feature subset, simplifying subsequent modeling and enhancing the model's accuracy and prediction potential.

3.3. Performance analysis of different prediction models

Algorithm choice hinges on data type and size, with no universal solution due to varying principles. Performance can differ even with the same dataset, necessitating multiple trials for optimal results. Hyperparameter selection also impacts model performance. To streamline, we used pycaret for model training and optimization, reducing workload and enhancing prediction accuracy and model versatility. Considering the Sichuan flue-cured tobacco dataset, we selected RF, SVM, KNN, and GBC for hyperparameter tuning and performance comparison using 5-fold cross-validation. Table 4 shows significant variations in prediction results among algorithms, with RF standing out at 78.2% accuracy and superior recall, precision, and F1-score, despite slower processing due to large datasets. Thus, RF was chosen as the classifier for predicting Sichuan flue-cured tobacco quality.

Table 4. Evaluation index scores of different prediction models

Model	Accuracy	Recall	Precision	F1-score	Time/s
RF	0.782	0.630	0.758	0.759	0.178
SVM	0.704	0.521	0.643	0.669	0.026
KNN	0.740	0.545	0.668	0.698	0.022
GBC	0.746	0.545	0.680	0.704	0.264

3.4. Establishment of the RF model

3.4.1 GA optimization

The setting of hyperparameters is crucial for enhancing model prediction stability and generalization ability. To avoid the subjectivity of manual tuning, GA is employed to optimize and adjust the hyperparameters of the RF model. GA is a type of random search algorithm that simulates the evolutionary principles of natural selection in biology, showing favorable performance for highly complex nonlinear problems[11]. It facilitates a leap from local to global optima, enhancing model performance significantly. Therefore, GA's efficient spatial search ability was chosen to determine the optimal RF model and the optimal combination of hyperparameters. GA sets the initial population to 100, a maximum number of iterations to 50, the error accuracy to 1*10-6, and the coding mode to real integer coding. Several hyperparameters can affect RF model performance. Optimizing these hyperparameters is a complex task, and it is challenging to determine the optimal combination by considering only a single parameter[12]. After preliminary debugging, six hyperparameters were selected for global optimization of the RF model. The results of the optimization process are listed in Table 5. The remaining hyperparameters were set to their default values.

Table 5. RF hyperparameter value range and GA optimization results

RF hyperparameter	Value range	GA optimization results
n_estimators	[1,150]	21
max_features	[10,38]	27
max_depth	[1,30]	13
min_samples_split	[2,20]	11
min_samples_leaf	[1,10]	2
min_impurity_decrease	[0.00,0.20]	0.00

3.4.2. Model evaluation

The training set served to assess the learning ability of the model, whereas the test set was used to evaluate its generalization capability. Through processing, analysis, and learning from the data in the training set, machine learning algorithms enable models to make predictions on unknown samples[13]. In this study, 38 random samples out of 188 were designated as the test set, with the remaining 150 used for training. To maintain class balance, the test set reflected the dataset's quality grade proportions. Table 6 shows the GA-RF prediction model's evaluation metrics, with a discriminatory accuracy of 86.8%. Recall, precision, and F1-scores for grades 1, 2, and 3 were 0.846-0.900-0.800, 0.917-0.857-0.800, and 0.880-0.878-0.800, respectively. Compared to the single RF model (78.2% accuracy, 0.630 recall, 0.758 precision, and 0.759 F1-score), the GA-RF model significantly outperformed. GA's integration enhanced the model's deep information extraction, improving overall effectiveness and performance. Thus, the GA-RF model's superiority was demonstrated, showcasing GA's crucial role in model optimization and intricate data pattern extraction.

The optimized GA-RF model, designed to avoid overfitting and boost generalization, delivered overall accurate predictions. Figure 5 shows minor misclassifications across tobacco leaf quality grades, with five samples (two grade 1, two grade 2, and one grade 3) miscategorized due to similar chemical compositions among adjacent grades, causing slight deviations in model judgment.

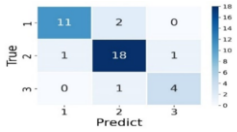

Figure 5. Confusion matrix of the GA-RF prediction model

Table 6. Evaluation index scores of the GA-RF prediction model

Quality grade	Recall	Precision	F1-score	Number of samples
Grade 1 (good)	0.846	0.917	0.880	13
Grade 2 (medium)	0.900	0.857	0.878	20
Grade 3 (general)	0.800	0.800	0.800	5
Accuracy		0.868		38

3.5. Key Feature Analysis of SHAP Values

3.5.1 Comprehensive feature analysis

Besides evaluation indicators, reliability of a model is also judged by transparency and explainability. The SHAP value, rooted in game theory, enhances tree model interpretation with rich visualizations for global and personalized feature insights[14]. Figure 6 shows the global feature importance using mean absolute SHAP values. The top 10 features affecting model predictions were ranked. Beeswarm plots for each quality grade illustrate the positive and negative contributions of each feature index. Here, each

color point indicates a sample's feature index level, with the SHAP value showing its contribution magnitude to the grade prediction[15]. As shown in Figure 6A, in the characteristic importance analysis of the GA-RF model, starch + reducing sugars contributed the most to the correct classification of Sichuan flue-cured tobacco quality; other important chemical indexes such as histidine, total basic amino acids, and total mono- and disaccharides/total nitrogen were also found, but their importance was significantly lower than that of starch + reducing sugars. As shown in Figure 6B and C, the indexes of starch + reducing sugars, histidine, total basic amino acids, and total mono- and disaccharides/total nitrogen had significant effects on the first and second quality grades. High starch + reducing sugars and total mono- and disaccharides/total nitrogen contents inclined predictions toward the first quality grade. Elevated histidine and total basic amino acids content suggested a likelihood of second quality grade prediction. As shown in Figure 6D, when the sucrose content was low, it had a positive contribution to the determination of the third-quality grade, while the trend of mysmine, malonic acid, nornicotine, and other indicators was the opposite.

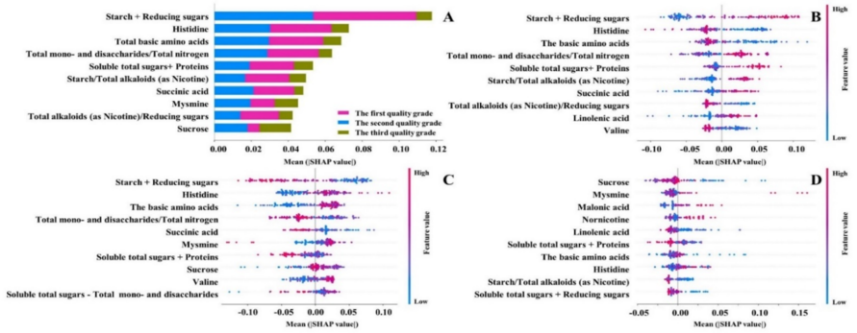

Figure 6. Global interpretation of the GA-RF prediction model based on SHAP value.

Note: A, summary plot; B, first quality grade warmth; C, second quality grade warmth; D, third-quality grade warmth.

3.5.2. Feature dependence analysis

To delve deeper into how each characteristic index affects model predictions, seven key indices—starch + reducing sugars, histidine, total basic amino acids, total mono- & disaccharides/total nitrogen, starch/total alkaloids (nicotine), total alkaloids (nicotine), reducing sugars, and sucrose—were chosen for feature-dependent analysis. These indices significantly contribute to overall interpretation and quality grading. Three-dimensional scatter plots were created for different quality grades, showing feature value ranges on the X-axis and corresponding SHAP values on the Y-axis. Since the tree model is unaffected by inter-feature dimension and multicollinearity[16], importing all indicators with original data allows the SHAP value to determine contribution degrees and help identify suitable intervals for feature indicators.

Figure 7A–C reveals the SHAP values for starch + reducing sugars in the first and second quality grades, reaching 0.100 and 0.075, respectively, while the prediction performance for the third-quality grade was weaker. The results indicate that when starch + reducing sugars exceeded 350 mg/g, most tobacco leaves improved in quality. Conversely, when the content fell below 325 mg/g, the tobacco leaves tended to be of the second quality grade. Starch and reducing sugars are the primary carbohydrates in flue-cured tobacco plants. Starch accumulates during growth in the field and is subsequently degraded during baking and curing, typically breaking down into small molecules of water-soluble sugars such as glucose, maltose, and other reducing

sugars[17,18]. Studies suggest that full conversion of macromolecular substances like starch benefits tobacco leaf drying, aroma enhancement, and quality improvement after roasting[19]. This indicates that there is a certain intensity of metabolic conversion between starch and reducing sugars in complex physiological activities, as well as a suitable dynamic balance relationship of content, which reflects the importance of the accumulation and decomposition of sugars in high-quality Sichuan tobacco leaves to a certain extent.

As shown in Figure 7D–I, histidine content ranging from 0.050 mg/g to 0.075 mg/g positively impacts the first quality grade of tobacco leaves. Slight deviations upwards lean towards the second grade, while extreme deviations in either direction favor the third grade. The optimal total basic amino acid content for the highest quality falls between 0.075 mg/g and 0.100 mg/g. Deviations suggest a nonlinear relationship, impacting sensory quality. Excessive free amino acids have been linked to the formation of harmful nitrogen-containing substances, such as hydrogen cyanide (HCN) and ammonia (NH3), during flue gas combustion, thereby affecting the quality and safety of cigarette smoke[20, 21, 1, 22]. In Sichuan flue-cured tobacco, histidine comprises 82.5% of total basic amino acids.

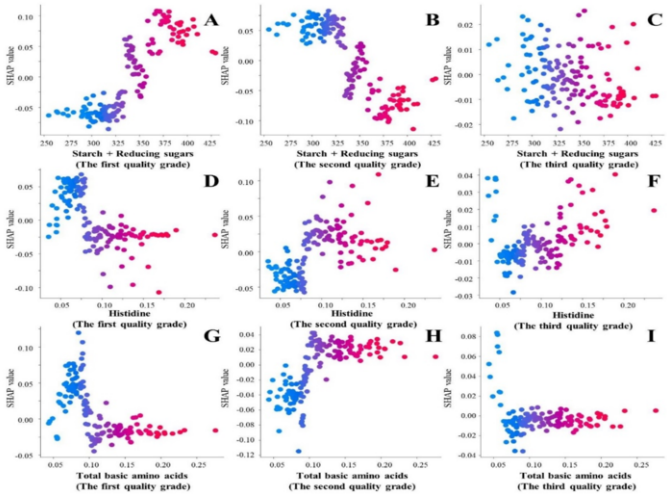

Figure 7. Dependence plot of starch + reducing sugars, histidine, and total basic amino acids among the three quality grades.

As shown in Figure 8, the contributions of sugar base characteristics, such as total mono- and disaccharides/total nitrogen, starch/total alkaloids (as nicotine), total alkaloids (as nicotine), and reducing sugars, to the model were also large. When the ratio of total mono- and disaccharides/total nitrogen was more than 12, and the ratio of starch/total alkaloids (as nicotine) was more than 2.5, the tobacco leaf tended to be of the highest quality grade. When the total alkaloid (as nicotine)/reducing sugar ratio was lower than 0.08, that is, the sugar-alkali ratio was higher than 12.5, the probability of judging the sample as the first quality grade was greater. When the ratio of total alkaloids (as nicotine) to reducing sugars exceeded 0.08, or the ratio of total mono- and disaccharides to total nitrogen was lower than 12, along with starch to total alkaloids ratio less than 2.5, the probability of categorizing the samples as quality grades 2 and 3 was higher.

Carbon and nitrogen metabolism are crucial for flue-cured tobacco growth,

influencing leaf compound composition and cigarette quality. Carbon metabolism converts inorganic carbon to sucrose, eventually forming carbohydrates like starch. Nitrogen metabolism involves amino acid-based protein synthesis or nitrogen conversion to nitrogenous substances[23]. Sichuan, situated in Southwest China's interior, features varying terrain, a complex climate, and diverse soil types. These unique ecological conditions in Sichuan promote sugar accumulation. high-quality flue-cured tobacco in the region boasts elevated levels of soluble total sugars and reducing sugars compared to the national average. Tobacco carbohydrates, along with nitrogen compounds, are key aroma precursors [19, 22, 24]. The sugar-alkali ratio (reducing sugars/total alkaloids) of Chinese tobacco leaves, typically ranging from 10 to 15 for high-quality leaves, affects tobacco flavor and internal quality, balancing vitality and aroma intensity.

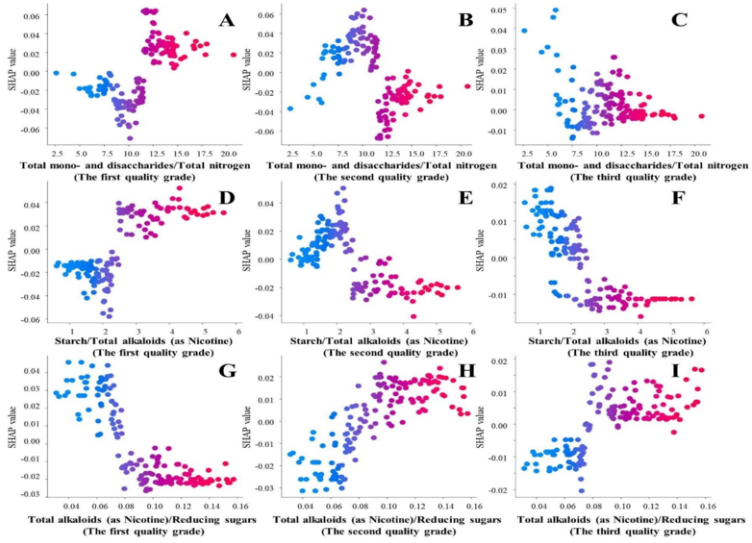

Figure 8. Dependence plot of total mono- and disaccharides/total nitrogen, starch/total alkaloids (as nicotine), total alkaloids (as nicotine)/reducing sugars among the three quality grades.

For the third-quality grade with a low discrimination accuracy of the GA-RF prediction model, the SHAP value was positioned to play a key role in the prediction of sucrose. According to the SHAP values in Figure 9A–C, the ranking of the contribution of sucrose to model quality grade prediction is as follows: 3> 2> 1 quality grade; when the sucrose content is less than 30 mg/g, the sample is more inclined to predict the third-quality grade. Sucrose, the main non-reducing sugar in tobacco leaves, reacts with amino compounds such as amino acids in non-enzyme-catalyzed Maillard reactions. This not only generates a series of important flavor substances in smoke, such as pyrazines, furans, and pyrroles, but also produces intermediate products, such as Amadori rearrangement products[19, 26]. Furthermore, sugars are also precursors of some characteristic flavor compounds during the tobacco preparation process. Some studies have shown that spraying a sucrose solution at a certain mass concentration before curing is beneficial for increasing the aroma of tobacco, alleviating green spots, and improving the sensory quality of cigarettes[24, 26, 27].

In the comprehensive feature analysis, SHAP values identified seven key indicators affecting Sichuan tobacco leaf quality: starch, reducing sugars, histidine, alkaline amino acids, total mono- and disaccharides/total nitrogen, starch/total alkaloids (nicotine), and sucrose. Starch, reducing sugars, and histidine significantly impact first- and second-

grade quality, while sucrose affects third-grade quality. Chemical composition and ratios closely relate to leaf quality, but no single indicator can fully represent it. Future studies will use multivariate statistical methods to explore these indicators across quality grades, establishing suitable ranges and enhancing the quality evaluation system.

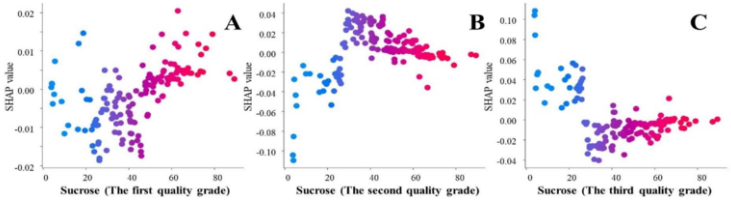

Figure 9. Dependence plot of sucrose among the three quality grades.

While a predictive model using the GA-RF algorithm has been developed, comprehensive validation and debugging are incomplete. The current dataset, limited to a specific Sichuan region and time period, may affect the model's predictive accuracy in different contexts. Thus, broadening data collection and refining the model through ongoing research is crucial.

4. Conclusion

In this study, model performance comparison, construction, and optimization were performed using various machine learning algorithms. In the case of unbalanced sample classification, the accuracy of the prediction model under 5-fold cross-validation was as follows: RF>GBC>KNN>SVM; the RF model performed the best in the evaluation metrics of recall, precision, and F1-score. Therefore, RF was identified as the best classifier model for predicting the quality of Sichuan tobacco plants. The six hyperparameters of the RF model underwent adjustment and optimization through GA, resulting in the identification of the optimal hyperparameter combination. The GA-RF prediction model achieved an impressive accuracy of 86.8% for Sichuan tobacco, successfully distinguishing between various tobacco quality grades. Furthermore, the incorporation of SHAP values offered a deeper insight into the model, enhancing its interpretability. This development holds significant value in advancing our scientific understanding of Sichuan tobacco's essential characteristic components and determining appropriate content ranges.

Acknowledgements

This work was funded by the Science and Technology Project of Sichuan Province of the CNTC (SCYC202118) and the Science and Technology Innovation Project of CAAS (ASTIP-TRIC07).

References

[1]. Rodgman A and Perfetti T, The Chemical Components of Tobacco and Tobacco Smoke, 2nd edn. CRC press (2013).

[2]. Wang D and Yang SX, Broad learning system with Takagi-Sugeno fuzzy subsystem for tobacco origin identification based on near infrared spectroscopy. Appl Soft Comput 134 (2023).
[3]. Marcelo MCA, Soares FLF, Ardila JA, Dias JC, Pedo R, Kaiser S, Pontes OFS, Pulcinelli CE and Sabin GP, Fast inline tobacco classification by near-infrared hyperspectral imaging and support vector machine-discriminant analysis. Anal Methods 11:1966–1975 (2019).
[4]. G DL, Somsubhra C, Bin L, C WD, Prithwiraj D and Jacob GC, Non-destructive prediction of nicotine content in tobacco using hyperspectral image-derived spectra and machine learning. J Biosyst Eng 47:106–117 (2022).
[5]. He C, Chen R, Ren K, Zhao G, He C, Hu B, Zou C, Jiang Y and Chen Y, A predictive model for the sensory aroma characteristics of flue-cured tobacco based on a back-propagation neural network. SV Appl Sci 2 (2020).
[6]. Gupta P, Sharma A and Jindal R, Scalable machine-learning algorithms for big data analytics: a comprehensive review. WIREs Data Min & Knowl 6:194-214 (2016).
[7]. Ghfar AA, Wabaidur SM, Ahmed AYBH, Alothman ZA, Khan MR and Al-Shaalan NH, Simultaneous determination of monosaccharides and oligosaccharides in dates using liquid chromatography-electrospray ionization mass spectrometry. Food Chem 176:487-492 (2015).
[8]. Song Q, Zhang C, Wu Y, Feng K, Guo H and Gu H, Multi-objective optimization of method of characteristics parameters based on genetic algorithm. Ann Nucl Energy 194 (2023).
[9]. Zaman EAK, Mohamed A and Ahmad A, Feature selection for online streaming high-dimensional data: A state-of-the-art review. Appl Soft Comput 127 (2022).
[10]. Gregorutti B, Michel B and Saint-Pierre P, Correlation and variable importance in random forests. Stat Comput 27:659-678 (2017).
[11]. Fogel DB, An introduction to simulated evolutionary optimization. IEEE Trans Neural Netw 5:3-14 (1994).
[12]. Raji ID, Bello-Salau H, Umoh IJ, Onumanyi AJ, Adegboye MA and Salawudeen AT, Simple deterministic selection-based genetic algorithm for hyperparameter tuning of machine learning models. Appl Sci Basel 12 (2022).
[13]. Han Z, Zhao J, Leung H, Ma KF and Wang W, A review of deep learning models for time series prediction. IEEE Sens J 21:7833-7848 (2021).
[14]. Wang D, Thunell S, Lindberg U, Jiang L, Trygg J and Tysklind M, Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods. J Environ Manage 301 (2022).
[15]. Tseng P-Y, Chen Y-T, Wang C-H, Chiu K-M, Peng Y-S, Hsu S-P, Chen K-L, Yang C-Y and Lee OK-S, Prediction of the development of acute kidney injury following cardiac surgery by machine learning. Crit Care 24 (2020).
[16]. Breiman L, Random forests. Mach Learn 45:5-32 (2001).
[17]. Song Z, Li T and Gong C, A review on starch changes in tobacco leaves during flue-curing. Front Agric China 3:435-439 (2009).
[18]. Yamaguchi N, Suzuki S and Makino A, Starch degradation by alpha-amylase in tobacco leaves during the curing process. Soil Sci Plant Nutr 59:904-911 (2013).
[19]. Banozic M, Jokic S, Ackar D, Blazic M and Subaric D, Carbohydrates-key players in tobacco aroma formation and quality determination. Molecules 25 (2020).
[20]. Im HS, Rasouli F and Hajaligol M, Formation of nitric oxide during tobacco oxidation. J Agric Food Chem 51:7366-7372 (2003).
[21]. Kibet JK, Khachatryan L and Dellinger B, Molecular products from the thermal degradation of glutamic acid. J Agric Food Chem 61:7696-7704 (2013).
[22]. Schmeltz I and Hoffmann DJCR, Nitrogen-containing compounds in tobacco and tobacco smoke. Chem Rev 77:295-311 (1977).
[23]. Dinkeloo K, Boyd S and Pilot G, Update on amino acid transporter functions and on possible amino acid sensing mechanisms in plants. Semin Cell Dev Biol 74:105-113 (2018).
[24]. Roemer E, Schorp MK, Piade J-J, Seeman JI, Leyden DE and Haussmann H-J, Scientific assessment of the use of sugars as cigarette tobacco ingredients: A review of published and other publicly available studies. Crit Rev Toxicol 42:244-278 (2012).
[25]. Mitsui K, David F, Tienpont B, Sandra K, Ochiai N, Tamura H and Sandra P, Analysis of the reaction products from micro-vial pyrolysis of the mixture glucose/proline and of a tobacco leaf extract: Search for Amadori intermediates. J Chromatogr A 1422:27-33 (2015).
[26]. Li T-X, Shi F-C, Li P-H, Luo C and Li D-L, A roasting method with sugar supplement to make better use of discarded tobacco leaves. Food Sci Technol 42 (2022).
[27]. Li N, Yu J, Yang J, Wang S, Yu L, Xu F and Yang C, Metabolomic analysis reveals key metabolites alleviating green spots under exogenous sucrose spraying in air-curing cigar tobacco leaves. Sci Rep 13:1311 (2023).