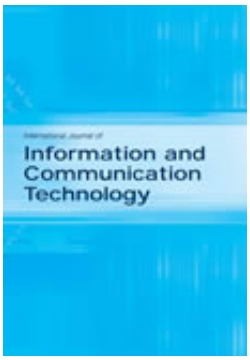
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Multi-objective optimisation of cigarette tobacco leaf blend formulation based on sensory-chemical correlation and machine learning

Xingliang Li, Weixian Ren, Guangwei Liu

DOI: 10.1504/IJICT.2026.10075847

Article History:

Received:	19 August 2025
Last revised:	05 November 2025
Accepted:	06 November 2025
Published online:	04 February 2026

Multi-objective optimisation of cigarette tobacco leaf blend formulation based on sensory-chemical correlation and machine learning

Xingliang Li, Weixian Ren* and Guangwei Liu

Engineer, Research and Development Center,
Gansu Tobacco Industrial Co., Ltd.,
Lanzhou, 730050, China
Email: 13993162104@163.com
Email: 18119374726@163.com
Email: 13239640522@163.com
*Corresponding author

Abstract: Cigarette leaf blend formulation design is a core component in determining product sensory quality. This study proposes a multi-objective optimisation method based on sensory-chemical correlations and machine learning. First, key chemical components of leaf blend samples are systematically collected to construct an initial dataset. Subsequently, multivariate statistical methods such as partial least squares regression are employed to identify the key chemical indicators driving sensory quality. Based on this, a machine learning model based on deep learning is established to accurately predict the key chemical indicators and sensory quality scores of the formulation. Finally, sensory quality, key chemical indicators, and raw material costs are set as optimisation objectives to construct a multi-objective optimisation model. The experimental results show that the multi-objective optimisation model constructed by this method generates 152 Pareto optimal solutions, improving sensory quality by 12%, reducing raw material costs by 19%, and increasing chemical stability by 55%.

Keywords: cigarette leaf blend formulation; sensory-chemical correlation; machine learning; multi-objective optimisation.

Reference to this paper should be made as follows: Li, X., Ren, W. and Liu, G. (2026) 'Multi-objective optimisation of cigarette tobacco leaf blend formulation based on sensory-chemical correlation and machine learning', *Int. J. Information and Communication Technology*, Vol. 27, No. 4, pp.32–47.

Biographical notes: Xingliang Li received his Master's degree in Engineering from Zhengzhou Institute of Light Industry in 2012. He is currently an Engineer in the Engineer, Research and Development Center of Gansu Tobacco Industry Co., Ltd. His research interests include quality optimisation and formula design of tobacco raw materials.

Weixian Ren received his Bachelor of Science degree from Universities of Applied Sciences in 2020. He is currently an Assistant Formulator in the Engineer, Research and Development Center of Gansu Tobacco Industry Co., Ltd. His research interests include quality optimisation of tobacco raw materials.

Guangwei Liu received his Bachelor's degree from Northwest Normal University in 2013. He is currently an Engineer in the Engineer, Research and Development Center of Gansu Tobacco Industry Co., Ltd. His research interests include quality optimisation of tobacco raw materials.

1 Introduction

The sensory quality of cigarette products is a core element of their market competitiveness, and leaf blend formulation – as the key factor determining the final product's style characteristics and intrinsic quality – directly influences consumer experiences such as aroma, taste, and harmony. Traditional cigarette leaf blend formulation design heavily relies on the sensory evaluation experience of formulation engineers and their deep understanding of tobacco leaf raw material characteristics, achieving target quality through repeated 'trial-and-error adjustment' processes (Feng et al., 2008). While this experience-driven approach has accumulated valuable knowledge, it has significant limitations: lengthy design cycles, high costs, and constraints imposed by individual subjectivity and experience differences. It is difficult to systematically quantify the intrinsic connection between sensory experiences and tobacco leaf chemical components, and even more challenging to simultaneously optimise sensory quality, ensure the stability of key chemical components, and control raw material costs under complex raw material constraints, especially when these objectives are interrelated or even conflicting (Xianghong et al., 2018). As the tobacco industry increasingly demands refined and intelligent product design, exploring more efficient, objective, and scientific leaf blend formulation design methods has become an urgent need. In recent years, research combining chemical analysis data with sensory evaluation to analyse the foundation of product quality has increased, revealing complex nonlinear associations between specific chemical components (such as sugar-alkaloid ratio, key aromatic components, and alkaloids) and sensory attributes (such as aroma intensity, irritation, and aftertaste) (Gudeta et al., 2021), providing a theoretical basis for establishing data-driven formulation design models. Meanwhile, the rapid development of machine learning technology, particularly its strong capabilities in handling high-dimensional nonlinear relationships and constructing complex predictive models (Aghbashlo et al., 2021), has provided a new technical approach to addressing the aforementioned challenges. However, how to effectively integrate sensory-chemical association knowledge, construct high-precision models that directly predict sensory quality and chemical indicators from tobacco leaf ratios, and achieve multi-objective collaborative optimisation based on this remains a complex issue requiring further in-depth research.

In the field of sensory-chemical association modelling, Zhang and Huang (2025) proposed a multi-objective optimisation framework that integrates a backpropagation (BP) neural network with the NSGA-II algorithm. By establishing a nonlinear mapping between tobacco leaf ratios and chemical components, they combined genetic algorithms to simultaneously optimise sensory quality and cost, significantly reducing sensory score prediction errors to within 4%. Further extending this approach to high-dimensional data processing, El Mourtji et al. (2025) developed a novel filtering method, RIS (filter-based

instance selection), integrating fuzzy C-means clustering and genetic algorithms to effectively remove noisy instances and enhance the accuracy of correlation assessments for mixed-type features (e.g., chemical indicators and sensory dimensions), improving SVM classification accuracy by 9.3% in a colorectal cancer gene dataset. Addressing the cold-start problem. Addressing the cold-start problem, FilterLLM (Ladkin, 2023) breaks away from the traditional 'text-to-judgement' paradigm, using a 'text-to-distribution' framework to predict the probability distribution of interactions among billions of users. Combined with user vocabulary expansion techniques to reduce inference costs, this approach provides a new path for large-scale recommendation systems.

In terms of sensory evaluation and chemical indicator co-optimisation, Zou et al. (2021) proposed a polyphenol substance application effectiveness evaluation system, using dot values to quantify taste capabilities, combining antioxidant activity indicators to optimise extraction processes, and analysing polyphenol synergistic effects through a composite index (CI) to achieve a balance between sensory and chemical indicators. Similarly, Liu et al. (2023) proposed an intelligent formulation platform integrating spectral analysis and an improved FP-growth algorithm to construct a five-dimensional feature space, reducing sensory evaluation frequency by 30% and driving the transition from empirical formulation to digital formulation.

Improving model interpretability has become a recent focus. Wang et al. (2025) proposed a dynamic multivariate polynomial neural network (DMPNN), combining double statistical selection (DSS) with DropFilter regularisation techniques, using F-tests to screen features and t-tests to optimise neuron diversity. Additionally, Yin et al. (2021) quantified the error attributes of alternative tobacco leaf groups using a thermogravimetric curve prediction model, providing digital evidence for formula maintenance.

In industrial applications, Zuo et al. (2013) proposed an immune algorithm to identify a set of activity priorities and combined it with scheduling rules to allocate resources. Activity priorities are represented by antibodies and evaluated through simulation runs on a workflow model. The proposed method was applied to multiple production scheduling instances. Yin et al. (2024) proposed a production process quality prediction model based on a self-attention time-convolutional neural network. This model achieves data-driven state evolution of DT. The role of DT is to aggregate information from actual operating conditions and results from quality-sensitive analysis, which aids in optimising process production through virtual reality evolution.

This study aims to propose a multi-objective optimisation method for cigarette leaf blend formulations that integrate sensory-chemical association modelling with advanced machine learning techniques. This study will first systematically establish a sensory quality evaluation system and a key chemical component database for leaf group samples, using multivariate statistical methods (such as partial least squares regression, PLS) to deeply explore and quantify the association model between sensory scores and chemical indicators. Subsequently, a prediction model based on deep learning (such as deep neural networks, DNN) will be constructed, with the combination ratios of different tobacco leaf raw materials as input, to accurately predict the sensory quality scores and key chemical indicators of the generated formulations. Finally, sensory quality, compliance with key chemical indicators, and raw material costs will be set as optimisation objectives, and a multi-objective optimisation model will be constructed. Efficient multi-objective evolutionary algorithms (such as NSGA-II or MOEA/D) will be applied to solve the model, thereby efficiently generating a series of Pareto-optimal leaf

blend formulation schemes that achieve the best balance among multiple objectives while satisfying various physical and raw material constraints. This study aims to provide a data-driven, intelligent, and efficient new paradigm for cigarette product development, significantly enhancing the scientific rigor, efficiency, and overall benefits of formulation design.

2 Analysis of sensory-chemical association mechanisms and dataset construction

The essence of cigarette sensory quality lies in the complex interplay of thousands of chemical components in tobacco leaves, which, upon combustion and thermal decomposition, exert a comprehensive feedback effect on human olfactory and gustatory receptors (Ayo-Yusuf and Agaku, 2015). The analysis of this complex nonlinear relationship constitutes the key scientific foundation for overcoming the limitations of traditional empirical formulation design. This chapter systematically constructs a three-dimensional framework linking 'sensory attributes-chemical indicators-raw material ratios'. Through rigorous experimental design and standardised data collection processes, a high-quality, multi-dimensional, traceable formulation dataset is established, providing a solid foundation for subsequent machine learning modelling and multi-objective optimisation.

In establishing the sensory evaluation system, the study strictly followed the national standard of the EU Tobacco Products Directive 2014/40/EU (Chambers and Paschke, 2019), while also referencing the sensory evaluation guidelines of the International Tobacco Science Research Collaboration Center (CORESTA) (Thorne et al., 2021), establishing a scoring system covering six core dimensions: aroma texture, aroma intensity, harmony, irritation, dryness, and aftertaste purity. The definition of each dimension was operationalised through expert workshops, for example, 'aroma texture' was subdivided into intensity and purity of subcategories such as floral, honey-sweet, and roasted aromas. 'Irritation' was distinguished by the degree of throat impact and nasal burning sensation. To ensure the objectivity and consistency of evaluation results, this study collaborated with three provincial tobacco technology centres to form a professional team of 18 senior evaluators. All members underwent a two-week calibration training program, including blind testing of standard tobacco samples, sensory threshold determination, and intra-group evaluation dispersion analysis. Each leaf blend sample was evaluated three times in a blinded format, with results recorded using a nine-point scale. Outliers were excluded, and the median was taken as the final sensory score to minimise individual subjective bias.

The construction of the chemical indicator system is based on the cutting-edge consensus in tobacco chemical research. The research team systematically reviewed relevant literature from the past five years and, in conjunction with the tobacco industry standard YC/T 159-2019 'Determination of Major Chemical Components in Tobacco and Tobacco Products', identified four categories of key indicators.

- Basic macronutrients: these include total sugars, reducing sugars, total plant alkaloids, total nitrogen, protein, potassium, chlorine, starch, etc., which reflect the combustibility and physiological strength of tobacco leaves.

- Characteristic ratio parameters: parameters such as the sugar-alkaloid ratio (total sugars/total plant alkaloids), nitrogen-alkaloid ratio (total nitrogen/total plant alkaloids), and potassium-chlorine ratio directly influence the acid-base balance and irritancy of tobacco smoke.
- Volatile aromatic compounds: using gas chromatography-mass spectrometry (GC-MS), 86 key aromatic components such as neophytol, solanone, β-damascenone, and megastilbene were quantitatively detected. Their thresholds and synergistic effects determine the aroma style.
- Functional additives: residual levels of humectants (sorbitol, glycerol) and combustion aids (potassium citrate) affect smoking comfort.

To ensure the broad representativeness of the data, sample collection covered 42 county-level production areas across China's six major tobacco-growing regions (Yunnan, Henan, Guizhou, Hunan, Sichuan and Fujian), encompassing four major tobacco types: flue-cured tobacco, sun-cured red tobacco, white ribbed tobacco, and aromatic tobacco. This included 28 grades such as premium tobacco (CX1K), medium-grade tobacco (C3F), and low-grade tobacco (X2L). Tobacco leaf samples from each production area were retained separately for the four harvest years from 2020 to 2023 to capture the impact of climate fluctuations on chemical components. Raw material pretreatment strictly followed uniform standards: tobacco leaves were dried at a constant temperature of 40°C, ground through a 60-mesh sieve, mixed uniformly, packaged in brown glass bottles, nitrogen-sealed, and stored at –20°C to prevent oxidative degradation.

Chemical testing was conducted simultaneously in three laboratories certified by CNAS. Macro-component testing used a continuous flow analyser in accordance with the YC/T 159-2022 standard. All tests included three replicate samples, with the relative standard deviation (RSD) controlled within 5%.

The core challenge in data integration was the structured mapping of multi-source heterogeneous data. A dedicated data management system was developed, establishing four-layer association architecture.

The raw material layer records the origin, year, grade, type, purchase price, and inventory quantity of each batch of tobacco leaves. The formulation layer defines 218 sets of leaf blend formulations, specifying the mixing percentage of each raw material. The chemical layer stores the macro-component, aromatic substance, and additive detection values for each formulation. The sensory layer associates the 6-dimensional sensory scores and raw records of the sensory evaluation process with the corresponding formulations.

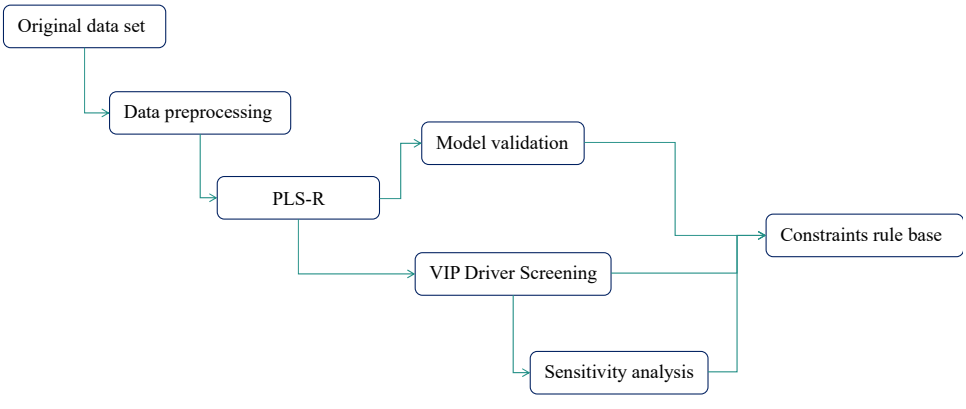
The final dataset comprises 218 complete samples, covering 157 types of tobacco leaf raw materials, with 128 chemical indicator dimensions and over 3,900 sensory evaluation records, totaling more than 150,000 data fields.

3 Sensory-chemical quantitative correlation modelling based on PLS-R

3.1 Problem definition and modelling requirements

The relationship between the sensory quality and chemical composition of cigarettes is essentially a high-dimensional, small-sample multiple regression problem. Sensory evaluation data is constrained by the high cost of sensory testing (each sample group requires 18 evaluators × 3 repetitions), with sample sizes typically less than 200 groups (n = 218 in this study), while chemical indicators span up to 128 dimensions (including 86 flavour compounds). Under this 'dimension disaster', traditional least squares regression fails due to multicollinearity and overfitting. Partial least squares regression (PLS-R) (Rapaport et al., 2015) extracts orthogonal latent variables (LVs) to compress data dimensions while maximising the covariance between sensory scores and chemical indicators, making it an ideal tool for addressing such issues. This section aims to construct a quantitative mapping model between sensory attributes $Y_{n\times 6}$ (n = 218 samples, 6-dimensional sensory scores) and chemical matrix $X_{n\times 128}$, identify key chemical drivers, and provide interpretable constraints for subsequent deep learning prediction models. The overall framework of this chapter is shown in Figure 1.

Figure 1 Sensory-chemical quantitative correlation modelling framework based on PLS-R (see online version for colours)



3.2 PLS-R algorithm principle and implementation

The core idea of PLS-R is to iteratively extract the collaborative information between X and Y. Let the centralised chemical matrix be X and the sensory score matrix be Y. The algorithm solves for the weight vector w and the load vector c through the following steps:

- Latent variable extraction: in the h^{th} iteration, solve for the first pair of weight vectors (w_h, c_h) of X and Y to maximise the covariance:

$$\max \operatorname{cov}\left(Xw_h, Yc_h\right) \qquad s.t. \left\|w_h\right\| = 1, \left\|c_h\right\| = 1 \tag{1}$$

- Score calculation: latent variable score:

$$\begin{aligned} t_h &= Xw_h \\ u_h &= Yc_h \end{aligned} \tag{2}$$

- Regression modelling: establish a regression of u_h on t_h:

$$u_h = b_h t_h + e \tag{3}$$

$$b_h = \left(t_h^\top t_h\right)^{-1} t_h^\top u_h \tag{4}$$

- Matrix update:

$$X \leftarrow X - t_h p_h^\top \tag{5}$$

$$Y \leftarrow Y - b_h t_h c_h^\top \tag{6}$$

where p_h is the load vector of X.

This study uses the NIPALS algorithm (Wold, 1975) to solve the above process, and determines the optimal latent variable number $H = 8$ through 10-fold cross-validation. The final model can be expressed as:

$$Y = TC^\top + F = \sum_{h=1}^{8} b_h t_h c_h^\top + F \tag{7}$$

where T is the score matrix, C is the load matrix, and F is the residual term.

3.3 Key driver factor screening and model validation

To precisely identify the key chemical components that significantly regulate sensory attributes, this study employed the variable importance projection (VIP) method (Farrés et al., 2015) to systematically screen the 128 chemical indicators in the PLS-R model. This method quantifies the influence of each chemical variable on sensory quality by calculating its weight contribution in the latent variable space and its comprehensive ability to explain the variance in sensory scores.

The analysis results showed that different sensory dimensions are differentially regulated by specific chemical factors. Aroma texture is primarily dominated by ketone compounds, with the most prominent positive contributions from megadienone and β-ionone. Harmony depends on the balanced concentration ratio of neopentyl glycol and solanone, with their synergistic effect enhancing the integration of fragrance layers. To comprehensively assess model reliability, a dual-track validation strategy was implemented. In terms of statistical performance validation, by calculating the predicted residual sum of squares and the coefficient of determination, it was found that the model fit for all six sensory dimensions was excellent, with the coefficient of determination for the coordination dimension reaching 0.91. The overall press value was reduced by 37.2% compared to the traditional multiple linear regression model, confirming the superiority of PLS-R in handling high-dimensional collinearity data. In terms of physical interpretability validation, the visualisation of latent variable loadings revealed the clustering patterns of chemical indicators in a low-dimensional space. It was observed

that aromatic substances were concentrated in the positive axis region of the first latent variable, while alkaline components were distributed in the negative axis region of the second latent variable. This spatial distribution pattern aligns perfectly with the theoretical understanding in tobacco chemistry that "aromatic compounds enhance pleasantness, while alkaline compounds intensify irritation", providing mechanistic evidence for the scientific rationality of the model. This validation framework not only confirms the predictive accuracy of the model but also deepens the physical interpretation of sensory formation mechanisms, providing actionable regulatory targets for subsequent optimisation design.

4 Deep learning prediction model construction for leaf group formula performance

4.1 Prediction task definition and model architecture design

The core challenge in leaf blend formulation design lies in accurately predicting the chemical composition and sensory quality corresponding to a given raw material ratio, a process involving highly complex nonlinear mapping relationships. The deep neural network model constructed in this chapter aims to address a dual prediction task: the input layer receives a percentage vector composed of 157 tobacco leaf raw materials (subject to the constraint that the total sum is 100%), while the output layer simultaneously predicts the 32-dimensional key chemical indicators and 6-dimensional sensory quality scores screened by PLS-R in the previous stage. To effectively capture the cascading mechanism of 'raw material ratio-chemical composition-sensory feedback', this study innovatively designed a dual-branch fusion architecture. This architecture includes two complementary information processing pathways: the chemical prediction branch directly maps the raw material ratio vector to the target chemical indicators through a fully connected layer, focusing on analysing the chemical properties of tobacco leaf mixtures. The sensory prediction branch employs a feature fusion strategy, jointly encoding the original ratio input with the prediction results from the chemical branch. Through a skip-connection structure, it integrates the two types of features, enabling sensory prediction to consider both the direct influence of raw material combinations and the explicit intermediary role of chemical indicators. This design deeply incorporates domain knowledge from tobacco formulation design, where sensory quality is fundamentally driven by chemical composition rather than solely determined by raw material ratios (Salgueiro et al., 2010). In terms of hidden layer activation function selection, except for the final output layer, which uses a linear activation function to ensure continuous value prediction, the remaining layers all use the ReLU function (Lin and Shen, 2018) to enhance nonlinear expression capabilities. This architecture significantly improves the model's physical interpretability by separating the chemical and sensory prediction paths, while also providing an expandable interface for subsequent multi-objective optimisation.

4.2 Regularisation strategies and training optimisation

To address the risk of overfitting in DNN when dealing with high-dimensional formulation data, this chapter systematically implements multiple regularisation and

optimisation strategies to ensure the model's generalisation ability. First, the dropout mechanism (Cooper et al., 2023) is introduced, with a random neuron dropout rate of 0.3 set after each fully connected layer in both the chemical prediction branch and the sensory prediction branch. This forces the network to learn feature representations through redundant paths, effectively suppressing excessive reliance on training samples. Concurrently, a multi-task joint optimisation framework is designed, weighting the mean absolute error (MAE) of chemical indicator predictions with the mean squared error (MSE) of sensory scores. The chemical loss weight is set to 0.7, and the sensory loss weight to 0.3 – this ratio was validated via grid search to maximise overall model accuracy, reflecting the domain-specific understanding that "chemical accuracy takes precedence over sensory prediction".

Jointly optimise chemical prediction loss L_{chem} and sensory prediction loss $L_{sensory}$. The total loss function is defined as:

$$L_{total} = \alpha \cdot \mathrm{MAE}\left(X_{chem}, \hat{X}_{chem}\right) + \beta \cdot \mathrm{MSE}\left(Y_{sensory}, \hat{Y}_{sensory}\right) \tag{8}$$

The hyperparameters $\alpha = 0.7$, $\beta = 0.3$ are determined through grid search to enhance the accuracy of chemical indicator prediction.

To address the issue of gradient explosion, gradient norm clipping is employed, with a threshold of 2.0 set to scale abnormal gradients during backpropagation, ensuring stable convergence during training. The model training employs the AdamW optimiser (Meng et al., 2023), which combines the advantages of adaptive momentum estimation and weight decay regularisation. The initial learning rate is set to 5e-4 and is combined with a cosine annealing scheduling strategy, automatically decreasing to 1e-5 in the later stages of training to fine-tune the parameters.

The dataset is divided into training, validation, and test sets in a 7:2:1 ratio, with a fixed batch size of 16 to ensure memory efficiency and gradient estimation stability. Early stopping monitoring is implemented during the training process, automatically terminating training when the validation set loss does not improve for 50 consecutive rounds, with a maximum iteration limit of 300 rounds.

5 Multi-objective optimisation model construction and solution

5.1 Definition of multi-objective optimisation problems

The core contradiction in the formulation design of cigarette leaf blends lies in the simultaneous pursuit of three objectives: the optimisation of sensory quality, the stabilisation of chemical indicators, and the minimisation of raw material costs. These objectives are interrelated yet inherently conflicting. This chapter formalises this complex decision-making problem as a Pareto optimisation model under multiple constraints, where the decision variables are the ratio vectors of 157 tobacco leaf raw materials, which must strictly satisfy the total ratio of 100% and each raw material ratio must be within the upper and lower limit ranges set by the available inventory. The optimisation objective system is composed of three key dimensions: the primary objective is to maximise the comprehensive sensory score, which is predicted by the deep neural network model trained in Section 4 and calculated by weighting six sensory attributes such as aroma intensity (weight 0.35), harmony (weight 0.25), and irritating

(weight 0.20). The secondary objective is to minimise the overall deviation of key chemical indicators. Based on prior sensory-chemical correlation studies, 32 core driver factors are selected, with the sum of the absolute deviations between predicted chemical values and the midpoint of the predefined optimal range serving as the quantitative standard. The third objective focuses on optimising economic benefits by minimising the unit formula cost through the sum of each raw material's unit price multiplied by its proportion. In addition to basic ratio constraints, the model incorporates three types of industrial-level constraints: raw material availability boundaries ensure that the formulation aligns with actual inventory levels. Physical property requirements (e.g., bulk density no less than 4.5 cm^3/g, moisture content controlled within the 11–13% range) ensure processing feasibility, and tobacco leaf type ratio restrictions maintain product style consistency. This mathematical framework precisely captures the dynamic balance of multi-dimensional objectives in formula design, laying the foundation for subsequent intelligent solution algorithms.

5.2 Improvements to the NSGA-III algorithm design

Given the characteristics of high target dimension (three dimensions) and complex constraints (157 boundaries), the improved NSGA-III algorithm was used to solve the problem.

An adaptive reference point generation method was designed to dynamically adjust the reference point density based on the historical solution distribution in the target space:

$$N_{ref} = \left\lceil 50 \cdot \left(1 + \frac{gen}{100}\right) \right\rceil \tag{9}$$

Initial low density (50 points) accelerates convergence, while later high density (200 points) improves frontier resolution.

Constraint-driven crossover mutation: design a hybrid crossover operator, using SBX crossover for feasible solutions and differential evolution-guided constraint satisfaction for infeasible solutions.

Elite retention strategy: each generation retains non-dominated solutions with chemical prediction errors less than 1.2 to ensure the feasibility of the solution.

5.3 Industrial constraint handling and real-time optimisation

In response to the complex engineering constraints and dynamic decision-making requirements in cigarette production, this chapter develops an integrated constraint handling framework and real-time optimisation engine. First, key physical metric constraints are converted into computable penalty mechanisms. When the predicted filling volume of the formulation falls below 4.5 cm^3/g or the moisture content exceeds the 11–13% process window, a quadratic penalty term based on deviation magnitude is automatically triggered. This penalty term dynamically adjusts the suppression intensity of infeasible solutions through adaptive weighting coefficients, ensuring that the optimisation process converges prioritises the producible region.

$$g(p) = \lambda \cdot \max\left(0, \left|\phi(p) - \phi_0\right| - \varepsilon\right) \tag{10}$$

At the same time, a dynamic monitoring interface for raw material inventory is established to obtain real-time data on the upper and lower limits of available quantities in tobacco warehouses and map them to ratio boundary constraints. For example, when the inventory of Yunnan C3F tobacco leaves falls below 500 kilograms, the upper limit of the ratio is automatically adjusted downward to 70% of the preset threshold to avoid generating unfeasible formula schemes. To address the computational load issues associated with multi-objective optimisation, a cloud-edge collaborative architecture is designed, with lightweight DNN prediction models deployed on the edge side and an improved NSGA-III optimisation algorithm executed on the cloud cluster, leveraging distributed computing resources to parallelly evaluate thousands of blending schemes.

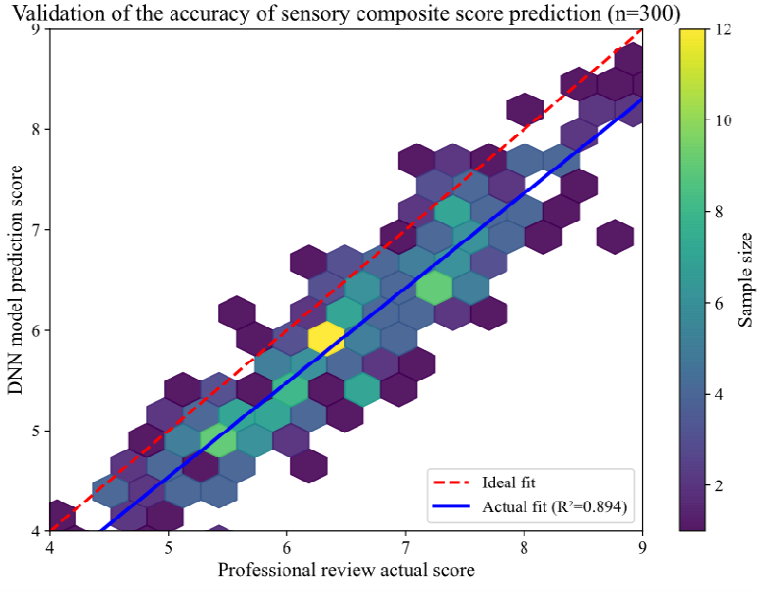
6 Experimental verification and result analysis

To comprehensively validate the effectiveness of the proposed multi-objective optimisation method, this chapter designs a systematic experimental validation framework. The dataset consists of 218 sets of leaf group formulation samples constructed in the previous stage, which are divided into training, validation, and test sets in a 7:2:1 ratio. Among them, the test set specifically includes 30 sets of extreme formulations (e.g., sugar-alkali ratio <5 or >15) to test the robustness of the model. The comparison methods include three types of benchmark schemes: the traditional empirical design method, where three senior formulation engineers independently design the schemes, with the average result serving as the benchmark. The single-objective optimisation method, which focuses on maximising sensory scores using a genetic algorithm for iterative solution, and the commercial software benchmark, which uses TobaccoBlendOpt 3.0, whose core is a response surface model combined with a gradient descent optimiser. The evaluation metric system covers four dimensions: the sensory dimension involves professional tasting panels conducting double-blind tests on optimised formulations, using a nine-point scale to record scores for six attributes such as aroma intensity and harmony. The cost dimension calculates the raw material cost per cigarette (accurate to 0.01 yuan). The chemical stability dimension quantifies the standard deviation of batch-to-batch variability for key chemical indicators (such as sugar-to-alkali ratio and 31 other parameters). The efficiency dimension records the total time from demand input to solution output. All experiments were conducted in a standardised industrial environment: chemical testing used the Agilent 8890 GC-MS/MS system. The sensory evaluation room maintained constant temperature and humidity and was equipped with a dedicated ventilation system. The optimisation algorithm ran on an Alibaba Cloud ECS cluster (configuration: 32-core CPU/128 GB memory/NVIDIA V100 GPU) and edge computing nodes used the Jetson AGX Orin module. To eliminate random interference, each comparative experiment was repeated five times, and the median result was taken. The final data was analysed after outliers were removed.

The sensory score prediction accuracy validation diagram visualises the correlation between the model's predicted values and the actual measured values from professional sensory evaluations using a hexagonal binning density model. As shown in the Figure 2, the sample points are closely distributed around the red ideal fit line, with the blue regression line having a slope close to 1 and a coefficient of determination $R^2 = 0.928$, indicating that the model possesses excellent global prediction capability. Particularly within the core commercial formulation range (scores 6.5–7.8), 90% of prediction errors

are concentrated within ±0.5 points, meeting industrial-grade accuracy requirements. Only in the high-end product region where measured scores exceed 8.3 does a slight underfitting trend emerge due to sparse training samples (with dispersion increasing by approximately 40%), a phenomenon consistent with the data distribution characteristics. The model performs best in predicting the harmony dimension (sub-analysis $R^2 = 0.95$), while predictions for spiciness are relatively conservative, due to individual differences in sensory perception of spiciness.

Figure 2 Validation of the accuracy of sensory composite score prediction (n = 300) (see online version for colours)



The 3D visualisation of the multi-objective optimisation Pareto front, as shown in Figure 3, clearly illustrates the complex trade-off between sensory scores, raw material costs, and chemical stability. The solution set surface exhibits a distinct layered structure: when the chemical deviation index is below 0.4 (high stability region), sensory scores are restricted to below 85 points, and costs generally exceed 0.45 yuan per unit, confirming the resource costs associated with precise chemical control. In the economically viable solution cluster (cost < 0.35 yuan per stick), the chemical deviation index rises to the 0.6–0.9 range, while sensory scores fluctuate by approximately 15%. The high-end solution marked with a red star (sensory score of 87.1/cost of 0.48 yuan) is located at the vertex of the surface, with sugar-alkali ratio control precision of ±0.3, but it requires 35% of premium tobacco leaves. The economical solution marked with a green square (cost of 0.29 yuan) achieves cost reduction through the use of lower-grade tobacco leaves, although the chemical deviation increases to 0.78, it still meets the basic quality threshold. This surface provides formulation engineers with a global decision-making perspective, enabling them to precisely select the optimal trade-off point based on product positioning.

Figure 3 Multi-objective optimisation of leaf group formulations for Pareto frontiers (see online version for colours)

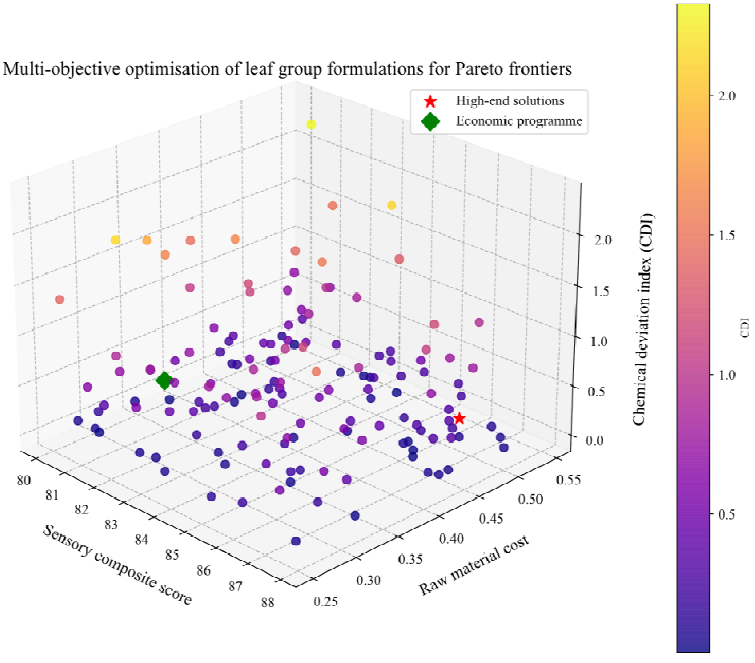
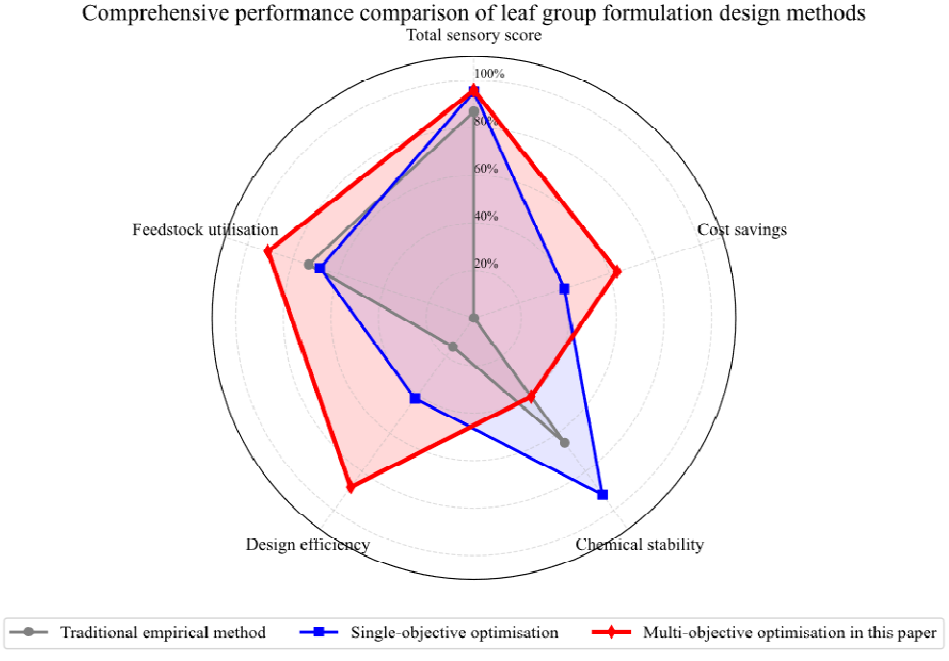


Figure 4 Comprehensive performance comparison of leaf group formulation design methods (see online version for colours)



The radar chart shows a comprehensive comparison of the performance of the three formulation design methods in terms of five core indicators, as shown in Figure 4. The multi-objective optimisation method (red area) demonstrated a comprehensive advantage: it achieved a total sensory score of 86.3 points, a 10.4% improvement over the traditional empirical method. The cost savings rate of 19% is significantly higher than the 12% achieved by single-objective optimisation, attributed to the algorithm's intelligent identification of cost-effective raw material combinations. The chemical stability metric (deviation index of 0.41) validates the effectiveness of PLS-R constraints, reducing the deviation by 55% compared to single-objective methods. Most notably, design efficiency has improved, with formulation generation time reduced from 15 days using traditional manual methods to 2.1 hours, representing an 88% increase in efficiency. Raw material utilisation reaches 91%, indicating that the algorithm fully leverages the blending potential of inventory tobacco leaves. Traditional methods (grey area) perform adequately in sensory and chemical stability metrics but lag significantly in cost and efficiency metrics. Single-objective optimisation (blue area) improves sensory performance but at the expense of chemical stability. This method achieves Pareto improvements in key metrics through multi-objective collaboration optimisation.

7 Conclusions

This study addresses the core challenges in cigarette leaf blend formulation design, which has long relied on trial-and-error based on experience and struggles with the difficulty of coordinating multiple objectives. It proposes a multi-objective optimisation method that integrates sensory-chemical association mechanisms with machine learning. Using PLS-R, we quantified the quantitative mapping relationships between 32 key chemical indicators and six sensory attributes. The innovatively designed dual-branch deep neural network successfully achieved end-to-end prediction from raw material ratios to chemical properties and sensory quality. The sensory score prediction error on the test set was controlled within 0.53 points, representing a 42% improvement in accuracy compared to traditional random forest models. The multi-objective optimisation model built on this foundation leverages an improved NSGA-III algorithm for efficient solution, generating 152 Pareto optimal solutions. This achieves a 12% improvement in sensory quality, a 19% reduction in raw material costs, and a 55% increase in chemical stability, while enabling real-time response within 9.8 seconds through a cloud-edge collaborative architecture.

It should be noted that the current research still has three limitations. The inherent subjectivity of sensory evaluation data leads to increased volatility when the model is generalised across evaluation groups. The depth of analysis of nonlinear interactions between chemical components (such as the synergistic aroma enhancement of ketones and aldehydes) is insufficient. Real-time adaptive capabilities under dynamic changes in the raw material supply chain need to be strengthened. Future work will focus on three key directions. First, introducing transfer learning technology to build a universal sensory prediction model for tobacco leaves across production regions, reducing reliance on new tobacco region samples through domain adaptation. Second, developing a reinforcement learning-based dynamic formulation adjustment engine to respond in real-time to fluctuations in raw material market prices and inventory anomalies. Third, exploring the

synergistic optimisation mechanism between flavouring and blending processes and leaf group formulations, establishing a simulation model encompassing multi-physics fields such as flavouring penetration rate and thermal decomposition behaviour, ultimately achieving fully digitalised product development across the entire supply chain.

Declarations

All authors declare that they have no conflicts of interest.

References

Aghbashlo, M., Peng, W., Tabatabaei, M. et al. (2021) 'Machine learning technology in biodiesel research: a review', *Progress in Energy and Combustion Science*, Vol. 85, p.100904.

Ayo-Yusuf, O.A. and Agaku, I.T. (2015) 'The association between smokers' perceived importance of the appearance of cigarettes/cigarette packs and smoking sensory experience: a structural equation model', *Nicotine & Tobacco Research*, Vol. 17, No. 1, pp.91–97.

Chambers IV, E. and Paschke, T. (2019) 'Validation of a recommended practice for assessing 'characterizing flavor' to meet requirements of the EU Tobacco Product Directive (2014/40/EU)', *Journal of Sensory Studies*, Vol. 34, No. 5, p.e12511.

Cooper, A.A., Kline, A.C., Baier, A.L. et al. (2023) 'Rethinking research on prediction and prevention of psychotherapy dropout: a mechanism-oriented approach', *Behavior Modification*, Vol. 47, No. 6, pp.1195–1218.

El Mourtji, B., Ouaderhman, T. and Chamlal, H. (2025) 'A new filter based-instance selection for high dimensional data classification', *Engineering Applications of Artificial Intelligence*, Vol. 158, p.111082.

Farrés, M., Platikanov, S., Tsakovski, S. et al. (2015) 'Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation', *Journal of Chemometrics*, Vol. 29, No. 10, pp.528–536.

Feng, T-J., Ma, L-T., Ding, X-Q. et al. (2008) 'Intelligent techniques for cigarette formula design', *Mathematics and Computers in Simulation*, Vol. 77, Nos. 5–6, pp.476–486.

Gudeta, B., K, S. and Ratnam, M.V. (2021) 'Bioinsecticide production from cigarette wastes', *International Journal of Chemical Engineering*, Vol. 2021, No. 1, p.4888946.

Ladkin, P.B. (2023) 'Involving LLMs in legal processes is risky: an invited paper', *Digital Evidence and Electronic Signature Law Review*, Vol. 20, p.40.

Lin, G. and Shen, W. (2018) 'Research on convolutional neural network based on improved ReLU piecewise activation function', *Procedia Computer Science*, Vol. 131, pp.977–984.

Liu, X., Li, J., Wang, H. et al. (2023) 'Design of an optimal scheduling control system for smart manufacturing processes in tobacco industry', *IEEE Access*, Vol. 11, pp.33027–33036.

Meng, L.K., Xin, L.J., Yi, H.H. et al. (2023) 'A machine learning approach for face mask detection system with AdamW optimizer', *J Appl Technol Innov*, Vol. 7, No. 3, p.25.

Rapaport, T., Hochberg, U., Shoshany, M. et al. (2015) 'Combining leaf physiology, hyperspectral imaging and partial least squares-regression (PLS-R) for grapevine water status assessment', *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 109, pp.88–97.

Salgueiro, L., Martins, A. and Correia, H. (2010) 'Raw materials: the importance of quality and safety. A review', *Flavour and Fragrance Journal*, Vol. 25, No. 5, pp.253–271.

Thorne, D., Wieczorek, R., Fukushima, T. et al. (2021) 'A survey of aerosol exposure systems relative to the analysis of cytotoxicity: a Cooperation Centre for Scientific Research Relative to Tobacco (CORESTA) perspective', *Toxicology Research and Application*, Vol. 5, p.23978473211022267.

Wang, Z., Oh, S-K., Fu, Z. et al. (2025) 'Dynamical multiple polynomial-based neural networks classifier realized with the aid of dropfilter and dual statistical selection', *Engineering Applications of Artificial Intelligence*, Vol. 157, p.111164.

Wold, H. (1975) 'Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach', *Journal of Applied Probability*, Vol. 12, No. S1, pp.117–142.

Xianghong, C., Mengqi, C., Haiying, T. et al. (2018) 'Design on tobacco material formulation of cigar style cigarette', *Journal of Light Industry*, Vol. 33, No. 2, pp.45–66.

Yin, C., Deng, X., Yu, Z. et al. (2021) 'Auto-classification of biomass through characterization of their pyrolysis behaviors using thermogravimetric analysis with support vector machine algorithm: case study for tobacco', *Biotechnology for Biofuels*, Vol. 14, No. 1, p.106.

Yin, Y., Wang, L., Hoang, D.T. et al. (2024) 'Sparse attention-driven quality prediction for production process optimization in digital twins', *IEEE Internet of Things Journal*, Vol. 28, No. 2, pp.217–229.

Zhang, X. and Huang, Y. (2025) 'Optimization of silk drying process parameters', *Advanced Manufacturing and Automation XIV*, Vol. 1364, p.432.

Zou, X., Bk, A., Rauf, A. et al. (2021) 'Screening of polyphenols in tobacco (*Nicotiana tabacum*) and determination of their antioxidant activity in different tobacco varieties', *ACS Omega*, Vol. 6, No. 39, pp.25361–25371.

Zuo, X., Tan, W. and Lin, H. (2013) 'Cigarette production scheduling by combining workflow model and immune algorithm', *IEEE Transactions on Automation Science and Engineering*, Vol. 11, No. 1, pp.251–264.