

Bimodal Fusion Recognition of Cigarette Tobacco Mold Contamination Level Based on Odor and Spectral Information

Jinxia Liu,^{1,2} Qilong Yang,² Qiaoyu Zhang,² Zhe Jin,¹
Feng Li,¹ Yan Shi,² and Hong Men^{2*}

¹Technical Research and Development Center, Jilin Tobacco Industrial Co., Ltd., Changchun 130000, China

²School of Automation Engineering, Northeast Electric Power University, Jilin 132012, China

(Received September 5, 2025; accepted October 19, 2026)

Keywords: cigarette tobacco, mold detection, E-nose, hyperspectral information, deep learning

Mold detection in cigarette tobacco is crucial for tobacco industry quality control. Traditional detection methods often suffer from limited efficiency, accuracy, and objectivity, highlighting the need for intelligent and reliable detection approaches. As a typical chemical sensing platform, the electronic nose (e-nose) system can generate characteristic responses to volatile organic compounds released during the mold development process of tobacco through a sensor array, providing chemical sensing information associated with mold contamination. However, single modal information is susceptible to environmental factors and sample differences under complex sample conditions, and its detection selectivity and stability still have limitations. In this study, we aim to develop an accurate and robust multimodal detection framework by integrating complementary odor and spectral information from an e-nose and hyperspectral imaging. Samples at three levels of mold contamination, namely, uncontaminated, moderately contaminated, and severely contaminated, were collected, and their e-nose responses and hyperspectral information were acquired. We designed a Complementary Feature Enhancement Module (CFEM) to strengthen single-modal feature representation and constructed a Cross-Modal Interaction Fusion (CMIF) module for deep inter-modal fusion. These components form the Collaborative Cross-Modal Interaction Fusion Network (CCIF-Net) for mold detection. Experimental results demonstrated that CCIF-Net significantly outperforms classical network models across all evaluation metrics while also validating the superiority of multimodal information over single-modal approaches. The framework provides an effective solution for industrial mold detection and offers valuable insights for multimodal fusion research in other domains.

1. Introduction

As a key raw material for cigarettes, the quality of tobacco leaf determines the market competitiveness of cigarette products and directly affects consumers' smoking experience and health. The issue of mold has always been one of the main factors affecting the quality of

*Corresponding author: e-mail: menhong@neepu.edu.cn
<https://doi.org/10.18494/SAM5930>

cigarette tobacco.⁽¹⁾ More critically, mold growth can lead to the production of harmful substances, notably aflatoxins, known for their potent carcinogenic and mutagenic properties. Prolonged inhalation poses significant risks to human health.

The tobacco leaves that have developed mold exhibit different sensory and physicochemical characteristics owing to their inherent moisture absorption and storage conditions.⁽²⁾ Therefore, the current detection of mold in tobacco primarily relies on sensory evaluation by trained personnel and chemical composition analysis.⁽³⁾ The former approach is highly subjective, heavily dependent on the experience of the evaluator, and prone to inconsistency over time owing to susceptibility to environmental factors. In contrast, while the latter method offers relatively high precision and accuracy, its cumbersome, costly, and time-consuming nature hinders its suitability for practical industrial production and market regulation demands.

The electronic nose (e-nose), an instrument composed of a series of chemical sensors and a pattern recognition system, is a typical chemical sensing platform that mimics the olfactory system by capturing the characteristic responses of volatile compounds. It has been widely applied in quality inspection fields, such as meat, fruits, and rice. Shi *et al.* extracted time-domain features from e-nose information to construct a fast Pearson graph convolutional network for rice quality identification.⁽⁴⁾ Chotimah *et al.* developed a low-cost e-nose system that predicts bacterial growth in vacuum-packaged chicken under two different storage conditions (room temperature and refrigeration) by building a Support Vector Regression (SVR) model using e-nose information as input.⁽⁵⁾ Li *et al.* employed wavelet decomposition to reduce signal drift and combined the e-nose with a Random Forest (RF) classification model, increasing the classification accuracy for adulteration detection in camellia oil from 88.67 to 98.67%, significantly improving the detection precision for adulterated oils.⁽⁶⁾ Jia *et al.* used an e-nose combined with Support Vector Machine (SVM) to classify the freshness of beef, achieving an accuracy above 95% across all classification methods, with the SVM model reaching a classification accuracy of 100%, indicating that the e-nose combined with machine learning can efficiently detect beef spoilage.⁽⁷⁾ Cui *et al.* combined chemical analysis methods with the e-nose to analyze the quality and odor characteristics of nutmeg samples with varying degrees of mold.⁽⁸⁾ Tang and Wang proposed a Multi-Branch Self-Attention Module (MSAM) based on deep learning, which, in conjunction with e-nose technology, enables the rapid detection of rice quality under different storage conditions.⁽⁹⁾

Hyperspectral imaging technology, as a three-dimensional detection technique that combines image information with spectral information, is widely used in fields such as crop nutrient detection, pest and disease detection, and agricultural product quality inspection.⁽¹⁰⁾ Park *et al.* proposed a nondestructive detection method for the contamination of seaweed with a combination of hyperspectral imaging and a mixed feature selection algorithm.⁽¹¹⁾ By fusing visible–NIR and shortwave infrared dual-band spectra, they achieved the precise measurement of the contamination levels of fresh and dried seaweed. Calderini *et al.* developed a quantitative prediction model that integrates hyperspectral imaging in the visible light range with a one-dimensional convolutional neural network to meet the detection needs for astaxanthin content in

Haematococcus pluvialis suspensions.⁽¹²⁾ Under the same cultivation conditions, the average prediction error for astaxanthin content was as low as 5.9%, significantly outperforming the predictive performance of single-wavelength or dual-wavelength linear regression models.⁽¹²⁾ Li *et al.* proposed a method for detecting the vitality of legume seeds based on hyperspectral imaging and a multi-scale spectral attention residual network.⁽¹³⁾ They constructed samples of different vitality levels through artificial accelerated aging, combined with the Successive Projections Algorithm (SPA) for double feature wavelength selection to optimize model input.⁽¹³⁾ Rashvand *et al.* combined hyperspectral imaging with machine learning algorithms to develop a nondestructive method for detecting the carbohydrate content in processed common bean flour.⁽¹⁴⁾ They showed that although deep learning models have a higher accuracy than machine learning models, the latter not only exhibit reliable accuracy but also have a significant impact on time analysis.⁽¹⁴⁾ Ahmed *et al.* developed a nondestructive detection method for fertilized eggs based on hyperspectral imaging and integrated learning.⁽¹⁵⁾ By using synthetic data augmentation and multi-strategy optimization, they constructed multiple classification models. This method overcomes the efficiency bottleneck of traditional egg candling, providing a green technological pathway for the intelligent sorting of fertilized eggs on a large scale in hatcheries and for the sustainable development of the egg industry.⁽¹⁵⁾

The e-nose and hyperspectral imaging technology are two crucial detection methods, each with its unique advantages. The e-nose analyzes the components of odors, while hyperspectral imaging technology uses spectral information to reflect physicochemical characteristics. However, the information contained in a single modality still has some limitations. Therefore, in this study, we aimed to develop an accurate and reliable multimodal detection framework by integrating complementary odor and spectral information from the e-nose and hyperspectral imaging for cigarette tobacco mold identification. To achieve this goal, we designed a novel fusion network that integrates the detection data from the e-nose and hyperspectral instruments, adaptively selecting important features from gas and spectral information to identify the moldiness of cigarette tobacco. The main contributions are as follows.

- 1) A multimodal tobacco mold detection dataset integrating e-nose responses and hyperspectral information is established using cigarette samples with three mold contamination levels, providing complementary chemical sensing and spectral features for nondestructive quality assessment.
- 2) A Complementary Feature Enhancement Module (CFEM) is designed for data fusion to enhance and optimize intra-modal features in e-nose and hyperspectral information.
- 3) A Cross-Modal Interaction Fusion (CMIF) module based on Cross-Modal Attention (CMA) is designed to achieve deep interaction and fusion between e-nose and hyperspectral information.
- 4) A Collaborative Cross-Modal Interaction Fusion Network (CCIF-Net) is proposed for cross-modal feature interaction and fusion. By incorporating CFEM and the CMIF module, the nondestructive detection of tobacco shred quality at different mold contamination levels is achieved.

2. Materials and Methods

Sixteen mainstream cigarette brands in the market were selected as experimental samples, with detailed information provided in Table 1. The selected brands covered major tobacco-producing regions, including Yunnan, Hunan, and Guizhou, and involved different processing types, such as traditional flue-cured tobacco and modern blended cigarettes. In addition, the selected samples included high-, medium-, and low-price categories to ensure a comprehensive representation of mainstream commercial cigarettes available.

2.1 Samples and instruments

The tobacco shreds used in the experiment were those taken out directly after opening commercially available cigarettes. To obtain unmolded, moderately contaminated, and severely contaminated samples representing three distinct levels of mold contamination, we designed the mold experiment. Environmental temperature and relative humidity are important factors affecting tobacco mold development and the growth and reproduction of molds. A high-humidity storage environment promotes moisture absorption by tobacco shreds and provides suitable conditions for mold growth. A relative humidity of approximately 85% facilitates moisture absorption by tobacco shreds, while a temperature of approximately 25 °C meets the requirements for mold growth and reproduction. Therefore, we selected 25 ± 1 °C and $85 \pm 1\%$ RH as the cultivation conditions for severely contaminated samples to simulate the high-humidity environment in which cigarettes are prone to mold development during storage. In addition, preliminary experiments were conducted to observe changes in the mold status of tobacco shreds under different cultivation periods, and 22 days was finally selected as the cultivation period. Therefore, the severely contaminated experimental parameters were

Table 1
Details of cigarette brands in the experimental sample.

No.	Brands	Tar content (mg)	Smoke nicotine content (mg)	CO content (mg)	Origin
1	Furongwang (Hard Middle)	10	1	9	Hunan Province
2	Guiyan (Red Middle)	8	0.8	7	Guizhou Province
3	Diamond (Soft Hehua)	10	1	10	Hebei Province
4	Huangshan (Hongfangyin Middle)	10	1	10	Anhui Province
5	Double Happiness (Baishun)	11	1.1	11	Shanghai
6	Huanghelou (Soft Blue)	11	1.1	13	Hubei Province
7	Golden Leaf (Gold Middle)	10	0.9	8	Henan Province
8	PRIDE (Kuanzhai Haoyun Slim)	8	0.8	7	Sichuan Province
9	Liqun (Soft Old Version)	6	0.6	7	Zhejiang Province
10	Nanjing (Twelve Hairpins of Jinling Mint)	6	0.6	5	Jiangsu Province
11	Renmin Dahuitang (Middle)	10	1	9	Liaoning Province
12	Yunyan (Lesser Panda Homeland)	10	1	11	Yunnan Province
13	Yuxi (Soft)	11	1	11	Yunnan Province
14	Yunyan (Congrong Hard)	10	1	11	Yunnan Province
15	Changbaishan (Tianlan)	10	0.9	11	Changbai Mountain
16	Zhongnanhai (5 mg)	5	0.4	6	Beijing

determined as follows: 60 g of commercially available cigarette tobacco shreds were incubated in a constant temperature and humidity incubator at 25 ± 1 °C and $85 \pm 1\%$ RH for 22 days.

The sampling schedule is as follows:

- 1) Stage 1. The first sampling was conducted on Day 0, before placing the samples in the constant temperature and humidity incubator. The second sampling occurred on Day 1. The samples collected during this initial stage were defined as uncontaminated samples.
- 2) Stage 2. The third sampling was conducted the next day. After that, samples were taken every other day to complete the 4th to 11th samplings. The samples collected during this intermediate stage were defined as moderately contaminated samples.
- 3) Stage 3. From Days 19 to 22, daily sampling was conducted, representing the 12th to 15th samplings. The samples collected during this later stage were defined as severely contaminated samples.

For a clearer illustration of the sampling days, a schematic diagram is provided in Fig. 1. In this figure, the yellow, green, and purple dots denote the sampling days of the first, second, and third stages, respectively.

This sampling strategy comprehensively captured the spectrum of mold contamination states in the tobacco shred samples. Each day, 1 g of samples underwent five information acquisition cycles, with six replicate measurements performed daily. Ultimately, 450 distinct tobacco shred sample groups representing different mold contamination levels were obtained.

Figure 2 shows the framework for e-nose and hyperspectral information fusion. The PEN3 e-nose (Airsense Analytics Inc., Schwerin, Germany) employs sensor array technology comprising three components: sensor perception, signal transmission, and information processing. The sensor array integrates ten MOS sensors, with key parameters detailed in Table 2. The sensor codes listed in Table 2, such as W1C, W5S, and W3C, represent different MOS sensor channels in the PEN3 e-nose and are used to distinguish sensors with different gas-sensing characteristics. Different sensor codes correspond to different response characteristics toward volatile compounds. For example, W1C is mainly sensitive to aromatic compounds, W5S shows a relatively high response to nitrogen oxides, and W3C is mainly sensitive to ammonia and aromatic compounds. Owing to their cross-sensitivity, the ten MOS sensors collectively generate a characteristic response fingerprint of the volatile compounds rather than relying on a single sensor for specific compound identification. The response value is defined as the ratio of electrical conductivity (G) when the sensor array interacts with volatile gases from tobacco

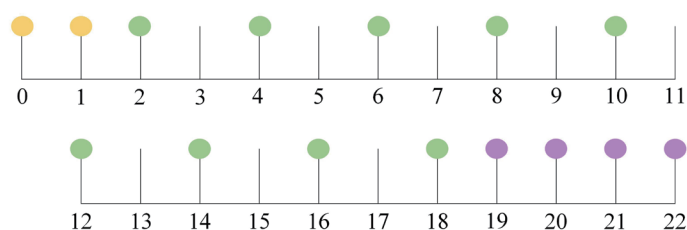


Fig. 1. (Color online) Schematic diagram of sampling days.

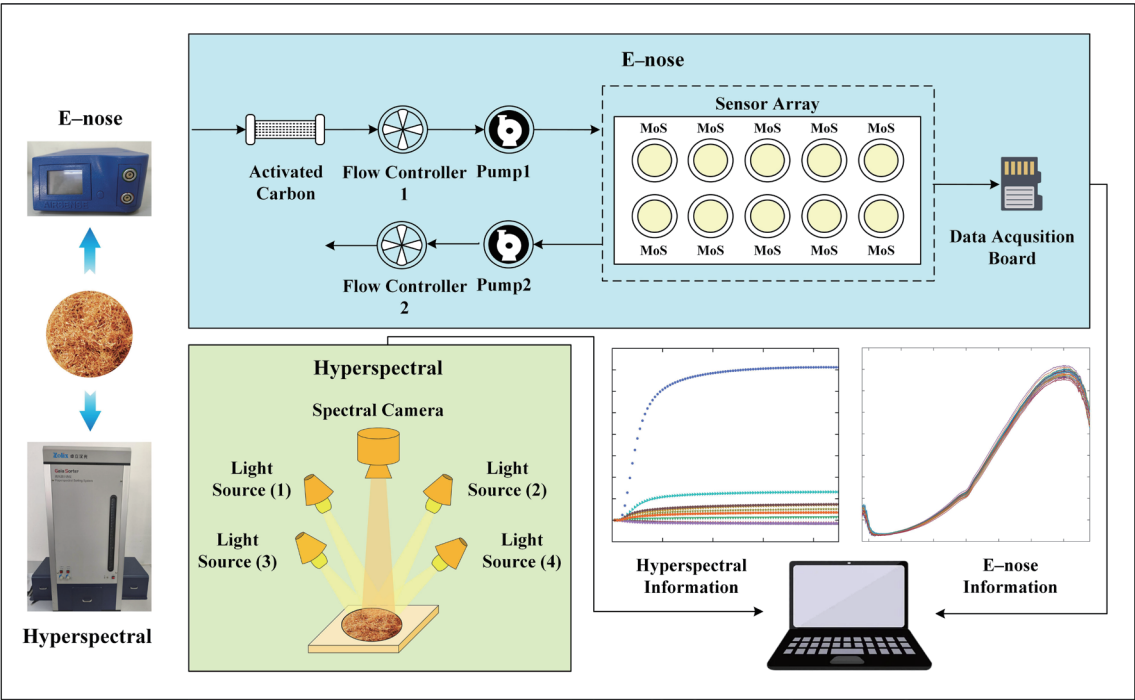


Fig. 2. (Color online) Framework for e-nose and hyperspectral information fusion.

Table 2
Key parameters and detection limits of e-nose sensors.⁽⁴⁾

No.	Sensor	Primary substances*	Detectability (ppm)
1	W1C	Aromatic compounds	10
2	W5S	Nitrogen oxides	1
3	W3C	Aromatic compounds and ammonia	10
4	W6S	Hydrogen	100
5	W5C	Less polar compounds, alkanes, and aromatic components	1
6	W1S	Methane, broad range sensitivity	100
7	W1W	Sulfur compounds, organic sulfur compounds, and many terpenes	1
8	W2S	Ethanol and some aromatic compounds	100
9	W2W	Sulfur organic compounds and aromatic components	1
10	W3S	High-concentration alkanes	10

*Substances to which the sensors are primarily sensitive.

shreds to the baseline conductivity (G_0) measure against activated carbon-filtered purified air, expressed as G / G_0 .

Hyperspectral information acquisition for tobacco shreds was performed using a GaiaSorter hyperspectral imaging system (Zollihan Optoelectronic Instrument Instruments Co., Ltd., Beijing, China). This system integrates the following components: a hyperspectral camera, uniform illumination source, motorized translation stage, control software (SpecView), and computer. Tobacco shred samples are positioned at the center of the motorized stage. The hyperspectral camera captures images by recording light reflected from samples illuminated by a tungsten-halogen lamp. The instrument operates across the spectral range of 380 to 1030 nm,

covering 520 spectral bands. The camera features a 1344×1024 pixel resolution, providing a high spectral resolution of 1.26 nm.

2.2 Detection experiments

The experimental procedure for detecting odor information in tobacco shreds at varying mold contamination levels using the e-nose is as follows.

- 1) The e-nose is preheated for 1 h to stabilize sensor output characteristics.
- 2) 1 g of tobacco shreds is weighed and sealed in a 300 ml sample vial, then maintained at 25 °C for 10 min to achieve headspace gas saturation.
- 3) The sensor chamber was purged and the baseline was calibrated. Prior to each measurement, the sensor chamber is purged with filtered clean air at a flow rate of 300 ml/min for 600 s to establish a clean baseline and ensure sensor cleanliness.
- 4) Following calibration, detection commences with a sampling period of 90s at 300 ml/min. Volatile compounds interact with sensors, generating response signals captured in real time via data acquisition cards using dynamic headspace sampling.
- 5) Replicates. Steps 2–4 are repeated to generate parallel samples. Consequently, each sampling time point produces 30 parallel samples (6×5), yielding a total of 450 samples.

In practical applications, sensor aging and damage should also be considered. The performance of MOS sensors is mainly affected by target gas type, humidity, and long-term baseline drift. Therefore, the sensor condition should be examined before each experiment and degraded sensors should be replaced promptly.

The experimental procedure for detecting spectral information in tobacco shreds at varying mold contamination levels using a hyperspectral imaging system is as follows.

- 1) System parameter initialization. The camera exposure time was set to 10 ms, the motorized stage speed to 0.28 cm/s, and the distance between tobacco shred samples and the imaging platform to 10 cm.
- 2) Dark and white correction was applied to mitigate the effects of ambient light, sensor sensitivity, and camera parameters on raw images. Specifically, dark reference images were acquired with the lens covered, while white reference images were obtained with the lens uncovered. The correction formula is as follows.

$$R_n = \frac{R_a - R_b}{R_c - R_b} \times 100\% \quad (1)$$

Here, R_a represents the raw hyperspectral image of tobacco shreds, R_b denotes the standard full-black reference image, R_c signifies the standard full-white reference image, and R_n indicates the corrected hyperspectral image.

- 3) Tobacco shred samples were positioned at the center of the motorized stage and transported through the spectral camera's acquisition area at a uniform speed to capture hyperspectral images.

- 4) Raw tobacco shred images were corrected using the dark/white reference images, completing one data acquisition cycle. To maintain generality, steps 2 and 3 were repeated. Each sampling time point yielded 30 parallel sample sets, resulting in 450 total sample sets.

2.3 Local feature extraction

The local feature extraction of e-nose information is as shown in Eq. (2). The inherent cross-sensitivity of e-nose and significant temporal variability in response values led to deeply correlated gas information across sensor and time dimensions. To address this, asymmetric convolutional kernels were employed for sequential convolution processing, effectively capturing and analyzing local feature dynamics in these two dimensions. Specifically, $S \times 1$ convolution first extracts time-varying features within individual sensors, while subsequent $1 \times S$ convolution analyzes response differences among different sensors at identical time points.

$$E = \text{Conv}_{1 \times 3}(\text{Conv}_{3 \times 1}(E)) \quad (2)$$

The local feature extraction of hyperspectral information is as shown in Eq. (3). A 3×3 convolution is utilized to preliminarily extract spatial features from input data, exploring and emphasizing local spatial and interband correlations within spectral information to better capture key spectral and spatial features.

$$H = \text{Conv}_{3 \times 3}(H) \quad (3)$$

2.4 Design of CFEM

We designed CFEM to enable complementary information interaction and enhancement between e-nose and hyperspectral information during feature extraction. Figures 3(a) and 3(b) illustrate the e-nose-guided and hyperspectral-guided feature enhancement structures, respectively. Both enhancement architectures incorporate Global Average Pooling (GAP) and 1D convolution to extract deep-level key features from their respective modalities. Subsequently, Sigmoid activation generates modality-specific weights that are applied to the complementary modality to amplify cross-modal information. This approach enables each modality to fully leverage its intrinsic global features while effectively integrating critical characteristics from the complementary modality, thereby enhancing the model's performance in identifying mold in tobacco shreds.

The e-nose-guided feature enhancement is given by

$$H_{input} = E_{input} \otimes \text{Sigmoid}\left(f\left(\text{GAP}\left(E_{input}\right)\right)\right) + H_{input}. \quad (4)$$

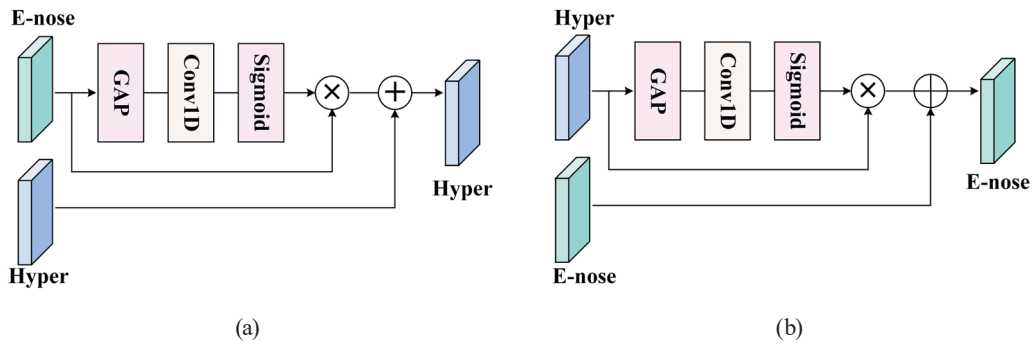


Fig. 3. (Color online) Structure of CFEM: (a) e-nose-guided feature enhancement and (b) hyperspectral-guided feature enhancement.

The hyperspectral-guided feature enhancement is given by

$$E_{input} = H_{input} \otimes \text{Sigmoid}\left(f\left(GAP\left(H_{input}\right)\right)\right) + E_{input}, \quad (5)$$

where $f(\cdot)$ denotes the Conv1D operation, H_{input} represents the hyperspectral input features, and E_{input} denotes the e-nose input features.

2.5 Design of CMIF

We designed CMIF to achieve the deep fusion of e-nose and hyperspectral information through fine-grained feature transformation and attention mechanisms. Its core operation, CMA, establishes correlations between different modal features. The module employs a bidirectional query-weighting mechanism, allowing each modality to not only represent itself but also interpret and utilize information from the complementary modality. Each branch utilizes its self-generated Query to interrogate the other modality's Key and employs its own Value to generate weighted features, ensuring the comprehensive utilization and integration of both modalities' information. The detailed implementation is as follows.

- 1) First, the CFEM outputs of the e-nose and hyperspectral features undergo 1×1 convolutions to generate the corresponding Query, Key, and Value vectors.
- 2) Next, e-nose-guided cross-modal attention computation is performed using Eq. (6). The e-nose modality utilizes its own Query vector and the hyperspectral modality's Key vector to calculate attention weights. These weights are then applied to its own Value features for weighted fusion, forming e-nose-dominated cross-modal fusion features.

$$CMA(E_{in}, H_{in}, E_{in}) = \text{SoftMax}\left(\left(f_q(E_{in}) \cdot f_k^T(H_{in})\right) / \alpha\right) \cdot f_v(E_{in}) \quad (6)$$

Here, $f_q(\cdot)$, $f_k(\cdot)$, and $f_v(\cdot)$ denote convolutional layers with 1×1 kernels.

- 3) Similarly, hyperspectral-guided CMA computation is performed using Eq. (7). The hyperspectral modality employs its own Query vector and the e-nose modality's Key vector

to compute attention weights. These weights are applied to its own Value features to obtain hyperspectral-dominated cross-modal fusion features.

$$CMA(H_{in}, E_{in}, H_{in}) = \text{SoftMax}\left(\left(f_q(H_{in}) \cdot f_k^T(E_{in}) / \alpha\right) \cdot f_v(H_{in})\right) \quad (7)$$

- 4) Following the establishment of global dependencies between multimodal features via CMA, CMIF is executed to obtain dynamically and adaptively fused multimodal features, and computed as

$$CMIF = [CMA(E_{in}, H_{in}, E_{in}) + E_{in}, CMA(H_{in}, E_{in}, H_{in}) + H_{in}] \quad (8)$$

Here, $[\cdot]$ denotes the concatenation operation.

2.6 Design of mold identification network for tobacco shreds

To improve the detection accuracy of mold contamination in tobacco shreds, we integrated gas and spectral information. The mold identification network (CCIF-Net) is illustrated in Fig. 4, with its primary workflow as follows.

- 1) E-nose information undergoes sequential asymmetric convolutions (1×3 and 3×1) for local feature extraction, while hyperspectral information undergoes 3×3 convolution for local feature extraction.

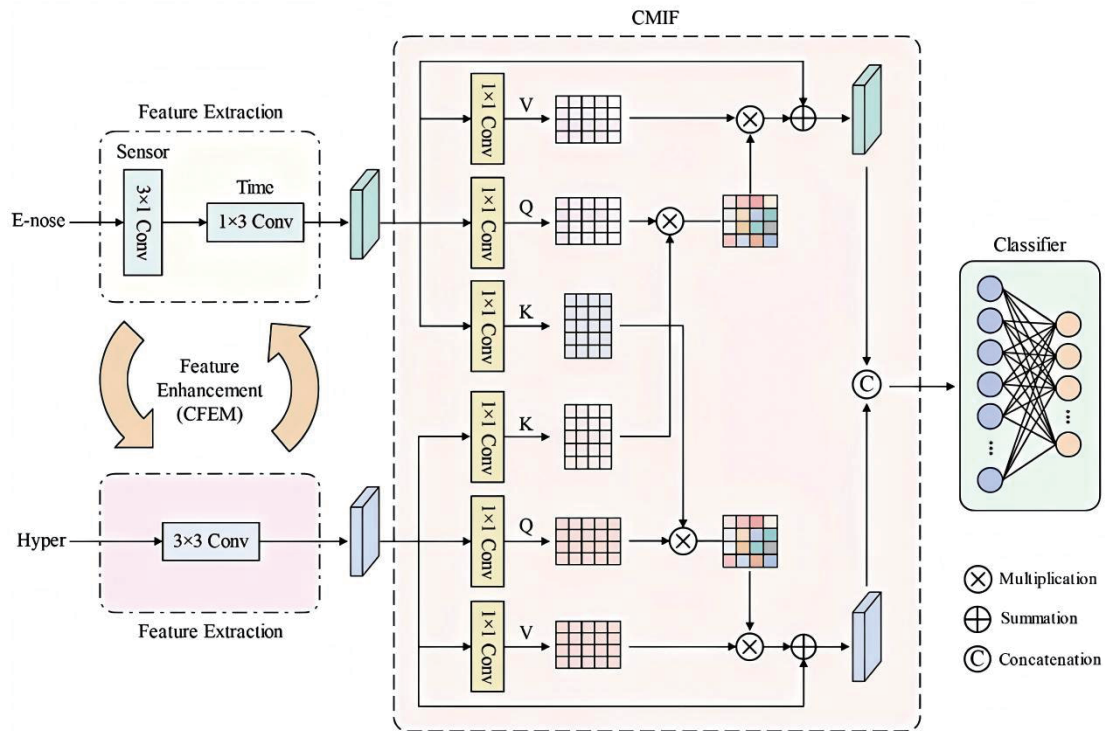


Fig. 4. (Color online) Tobacco shred mold identification network framework.

- 2) The local features from both modalities are then fed into CFEM for preliminary information interaction and complementary enhancement.
- 3) CFEM-processed features are subsequently input to the CMIF module for deep cross-modal interaction and fusion, outputting fused features.
- 4) Finally, the fused features are fed into the classification module for final recognition.

2.7 Model evaluation metrics

Accuracy, precision, recall, and F1-score were employed to evaluate the recognition performance of the model. The definitions and calculation methods of these evaluation metrics are as follows:

1) Accuracy

Accuracy represents the proportion of correctly classified samples among all the samples. It is calculated as the number of correctly classified samples divided by the total number of samples and directly reflects the overall classification correctness.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (9)$$

Here, TP represents the number of samples correctly predicted as positive. TN represents the number of samples correctly predicted as negative. FP represents the number of samples incorrectly predicted as positive. FN represents the number of samples incorrectly predicted as negative.

2) Precision

Precision refers to the proportion of samples that are actually positive among all the samples predicted as positive. It is used to evaluate the accuracy of positive predictions.

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

3) Recall

Recall refers to the proportion of correctly predicted positive samples among all the samples that are actually positive. It evaluates the ability of the model to identify positive samples based on the actual class labels.

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

(4) F1-score

F1-score is the harmonic mean of precision and recall, which is used to comprehensively evaluate the performance of classification models.

$$\frac{2}{F_1} = \frac{1}{Precision} + \frac{1}{Recall} \quad (12)$$

$$F_1 = \frac{2Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \quad (13)$$

3. Results and Discussion

All algorithms are implemented on a computer equipped with a 12th Gen Intel(R) Core(TM) i7-12700H CPU and an NVIDIA GeForce RTX 3060 GPU. All programs are executed in a Python 3.9 environment with CUDA 11.8 for computation.

3.1 Information preprocessing

Figure 5(a) shows the sensor response curve, exhibiting a characteristic rapid ascent followed by gradual stabilization. Figures 5(b)–5(d) present radar plots of gas information from tobacco

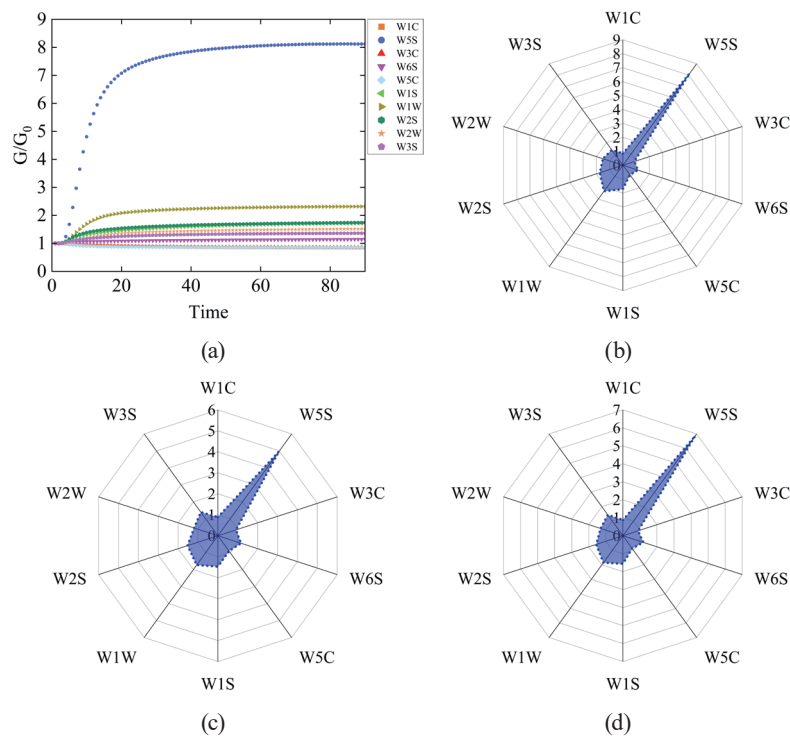


Fig. 5. (Color online) E-nose responses of cigarette tobacco shreds at varying mold contamination levels: (a) sensor response curve, (b) radar plot of unseverely contaminated tobacco shreds, (c) radar plot of moderately contaminated tobacco shreds, and (d) radar plot of severely contaminated tobacco shreds.

shreds at different mold contamination levels, where radial axes represent corresponding sensor outputs. Radar plots are employed here because they can simultaneously display the response intensities of the 10 sensors to each sample, intuitively presenting the multidimensional odor profile of each sample and facilitating the channel-by-channel comparison of the response differences among samples with different degrees of mold. In addition, using radar plots to display the multi-sensor response patterns of e-noses is a common practice in this field.^(16–19) As shown in Figs. 5(b)–5(d), among all the sensors, W5S demonstrates the highest response values, followed by W1W, suggesting their heightened sensitivity to volatile compounds from mold-contaminated tobacco. Conversely, W3C and W5C exhibit lower response magnitudes, indicating weaker detection capabilities for these compounds. Overall, the radar plot patterns exhibit considerable similarity across the different mold contamination levels.

Figures 6(a)–6(d) illustrate representative average spectral results of unmolded cigarette tobacco shred samples. Different colors in the figures represent different parallel experiments, reflecting the spectral variations among repeated measurements. Raw spectral information is processed using the Savitzky–Golay (S–G) method, effectively reducing spectral noise and enhancing curve smoothness. Standard Normal Variate (SNV) transformation primarily eliminates scattering effects while improving spectral contrast. Multiplicative Scatter Correction (MSC) corrects spectral distortions caused by particle size variations and light scattering. This enhances spectral aggregation within homogeneous samples, suppresses noise interference, and corrects baseline drifts and offsets.

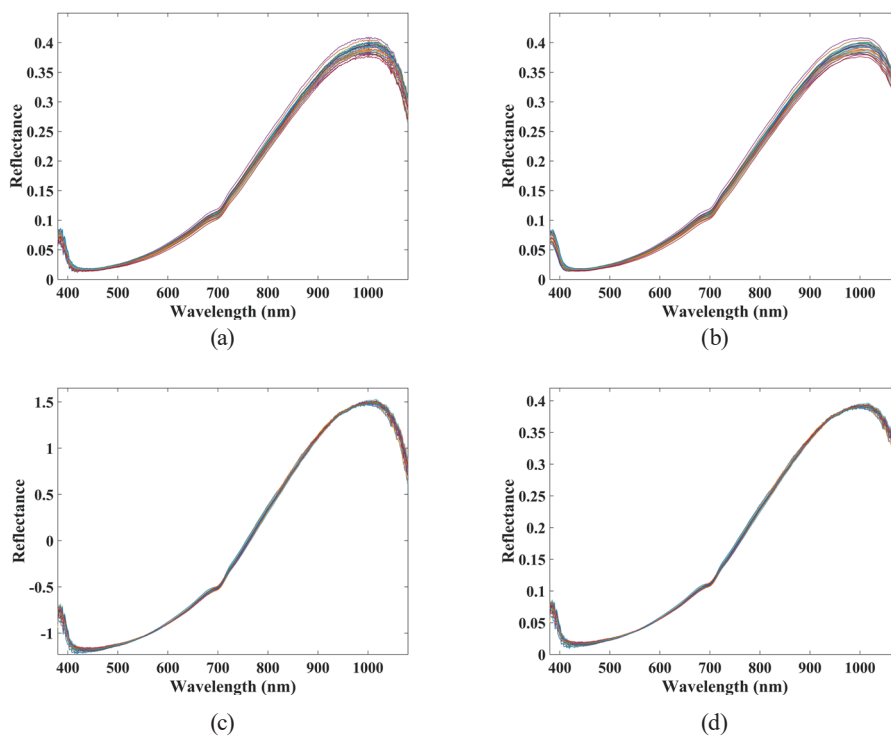


Fig. 6. (Color online) Average spectral results of unmolded cigarette tobacco shred samples: (a) raw data of unmolded tobacco shreds, (b) S–G results, (c) SNV results, and (d) MSC results.

3.2 Bimodal fusion

To maintain correspondence between gas information dimensions and hyperspectral band numbers for gas-spectral information fusion, the following procedures were implemented:

- 1) For odor information, all response data from 10 e-nose sensors over 52 s were selected to construct a 450×520 odor feature matrix.
- 2) For spectral information, regions of interest (ROIs) measuring 240×240 pixels were first selected. The average values of all pixels within each ROI were computed as feature values, forming a 450×520 spectral feature matrix.

This enables the data-level integration of the odor and spectral information, facilitating direct interaction and mutual influence between the two data modalities. Furthermore, data-level fusion facilitates the enhanced learning of the correlations between features derived from the odor and spectral information. The fused data serves as direct input to CCIF-Net without preprocessing.

3.3 Hyperparameters

During network training, the Adam optimizer was employed to update network parameters. Adam adaptively adjusts the learning rate of each parameter by combining the first-order and second-order moment estimations of gradients, thereby improving the stability and convergence efficiency of the training process. The cross-entropy loss function was used to compute the training loss, thereby enabling error backpropagation. The training iteration count was set to 200 epochs with a batch size of 40. From the total of 450 samples representing three mold contamination levels of tobacco shreds, 360 samples were randomly selected for the training set and 90 samples for the test set. The classification performance of CCIF-Net was evaluated on the basis of accuracy, precision, recall, and F1-score. To mitigate randomness, the mean and standard deviation of 20 computational results served as the final classification metrics.

3.4 Ablation results

To comprehensively evaluate the effectiveness of the proposed model in e-nose and hyperspectral fusion data analysis, two primary ablation experiments were designed herein. These experiments were aimed to systematically investigate the impact of CFEM and the CMIF modules on model performance. Table 3 shows the performance impact of the two modules.

First, when neither CFEM nor the CMIF module were incorporated, CCIF-Net exhibited relatively low performance metrics: accuracy (95.72%), precision (96.17%), recall (96.10%), and F1-score (96.50%). Without effective intermodal interaction and fusion mechanisms, the model struggled to fully leverage complementary features within the multimodal data.

Second, with the standalone CFEM incorporated, all model metrics showed significant improvement. This outcome demonstrated the CFEM effectiveness in individually enhancing the feature representation capabilities of both the e-nose and hyperspectral modalities.

Table 3
Ablation results of CCIF-Net with different structures.

CFEM	CMIF	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
√	√	95.72 ± 3.15	96.17 ± 0.02	96.10 ± 0.03	96.50 ± 0.03
		97.17 ± 1.08	97.70 ± 0.01	96.70 ± 0.007	97.12 ± 0.01
√	√	97.94 ± 0.53	98.45 ± 0.004	97.06 ± 0.007	97.67 ± 0.006
	√	99.56 ± 0.24	99.54 ± 0.006	99.41 ± 0.007	99.46 ± 0.009

Third, when only the CMIF module was added, its performance surpassed that of the CFEM-only configuration. This reflected how the CMIF module successfully achieves more efficient intermodal information fusion and interaction through its cross-modal attention mechanism.

Finally, when both CFEM and the CMIF module were integrated, all model performance metrics reached optimal levels: accuracy (99.56%), precision (99.54%), recall (99.41%), and F1-score (99.46%). This substantial improvement over the single-module configurations validated the effective synergistic interaction between the proposed CFEM and the CMIF module.

In conclusion, the ablation experiments effectively validated the indispensability of both CFEM and the CMIF module within CCIF-Net. The most outstanding performance achieved through their integrated use demonstrated the rationality and effectiveness of the proposed method and architectural design for fusing e-nose and hyperspectral information.

3.5 Performance comparison and analysis of CCIF-Net

In this study, we introduced and successfully implemented CCIF-Net for detecting mold in cigarette tobacco. This model integrated CFEM and the CMIF module to enhance the data interaction representation capability within modalities and promote the deep fusion of data between modalities. Compared with other established network architectures, CCIF-Net demonstrated outstanding performance in the cigarette tobacco mold detection task. In this study, we used CCIF-Net as the basic framework and compared and analyzed it with six classic network models to validate its effectiveness. The classic network models used for comparison included LeNet,⁽²⁰⁾ MobileNet V1⁽²¹⁾ and V2,⁽²²⁾ Transformer,⁽²³⁾ ViT,⁽²⁴⁾ and ResNet-50.⁽²⁵⁾ These models represent different categories of feature learning approaches, including conventional convolutional neural networks, lightweight feature extraction networks, residual networks, and attention-based architectures, which have been widely applied in sensor signal analysis, image recognition, and hyperspectral data classification tasks. Therefore, they were selected as representative baseline models for evaluating the effectiveness of the proposed CCIF-Net.

Table 4 presents the performance comparisons of different network models for tobacco shred mold detection. The results demonstrated that CCIF-Net outperforms all comparative models across evaluation metrics. Specifically, CCIF-Net achieved significantly higher performance than traditional CNN architectures such as LeNet, lightweight convolutional models (MobileNet V1/V2), and even advanced models including Transformer, ViT, and ResNet-50. Although these networks exhibit strong feature representation capabilities, they lack dedicated mechanisms for modeling the complex correlations and complementary information between heterogeneous

Table 4
Performance comparison between CCIF-Net and classical networks.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
LeNet	92.22 ± 1.91	93.26 ± 0.02	93.61 ± 0.01	93.20 ± 0.01
MobileNet V1	92.74 ± 3.36	92.60 ± 0.03	92.52 ± 0.03	92.56 ± 0.03
MobileNet V2	93.85 ± 2.30	93.72 ± 0.02	93.68 ± 0.02	93.70 ± 0.02
Transformer	94.32 ± 1.29	94.25 ± 0.02	94.16 ± 0.03	94.20 ± 0.02
ViT	94.87 ± 1.27	94.78 ± 0.02	94.70 ± 0.02	94.74 ± 0.02
ResNet-50	95.46 ± 2.26	95.38 ± 0.02	95.32 ± 0.02	95.35 ± 0.25
CCIF-Net	99.56 ± 0.24	99.54 ± 0.006	99.41 ± 0.007	99.46 ± 0.009

sensing modalities. In contrast, CCIF-Net effectively improves feature representation through CFEM and establishes deep cross-modal interactions through CMIF, thereby fully exploiting the complementary chemical sensing information from the e-nose and spectral information from hyperspectral imaging for more accurate tobacco mold detection.

3.6 Comparative analysis of recognition results

To evaluate the impact of dataset selection on model performance, we employed three distinct datasets: E-nose, Hyper, and Fusion. These datasets reflect the characteristics of both single-modal and multimodal data. As shown in Table 5, when solely using e-nose data, the model attained the following: accuracy (96.17%), precision (96.78%), recall (95.89%), and F1-score (96.04%). With hyperspectral data alone, performance improved to the following: accuracy (97.34%), precision (97.57%), recall (97.73%), and F1-score (97.23%).

When fusing e-nose and hyperspectral data for joint feature learning, the classification performance of CCIF-Net improved significantly: accuracy increased to 99.56%, precision reached 99.54%, recall rose to 99.41%, and F1-score attained 99.46%. These results demonstrated the complementary characteristics of e-nose and hyperspectral data. Integrating these two modalities enabled the model to comprehensively capture subtle feature differences across varying mold contamination levels in cigarette tobacco. This enhanced recognition capability while mitigating the information gaps and uncertainties inherent in single-modal approaches.

4. Model Limitations and Future Extensions

Our proposed CCIF-Net demonstrated reliable recognition performance in tobacco mold detection. However, the contaminated samples were mainly prepared by artificially controlling the degree of mold contamination. This process differs from natural mold development during actual storage, which is affected by multiple factors such as temperature, humidity, and microbial growth. In addition, we prepared the severely contaminated samples only under specific temperature and humidity conditions. The effects of different environmental conditions on mold development and model detection performance remain to be further investigated. Moreover, the e-nose and hyperspectral imaging equipment used in our experiments involved relatively high costs and limited portability. These factors restricted their application in rapid detection under industrial conditions. In future work, we will further evaluate the applicability of the model using naturally severely contaminated samples collected under different environmental

Table 5
CCIF-Net results on different datasets.

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
E-nose	96.17 ± 3.34	96.78 ± 0.03	95.89 ± 0.01	96.04 ± 0.03
Hyper	97.34 ± 1.29	97.57 ± 0.01	97.73 ± 0.01	97.23 ± 0.01
Fusion	99.56 ± 0.24	99.54 ± 0.006	99.41 ± 0.007	99.46 ± 0.009

conditions. We will also investigate the contribution of individual sensor channels to provide a reference for sensor selection, array optimization, and the development of new sensors targeting key volatile compounds while combining portable sensing devices with lightweight intelligent algorithms to support a faster and on-site detection of tobacco mold.

5. Conclusions

In this study, we proposed and implemented CCIF-Net, a network integrating e-nose and hyperspectral data, to overcome the limitations of single-modality detection—specifically inadequate stability and constrained feature representation—in identifying varying mold contamination levels in tobacco shreds. The principal contributions and conclusions are as follows.

- 1) CFEM was designed to effectively enhance the inherent representation capability of both e-nose and hyperspectral modalities through intra-modal feature enhancement.
- 2) The CMA mechanism was designed, enabling the construction of CMIF to achieve deep interaction and fusion between modalities, effectively compensating for the limitations of single-modality information.
- 3) On the basis of CFEM and CMIF, CCIF-Net was designed. Compared with classical network architectures, CCIF-Net achieved optimal classification performance for mold identification in tobacco shreds.
- 4) Comparative experiments between single-modality and fused data further confirmed that multimodal data fusion significantly outperforms single-modality approaches, effectively enhancing the accuracy and reliability of mold contamination detection in tobacco shreds.

In summary, the proposed CCIF-Net significantly enhances feature interaction and fusion capabilities between e-nose and hyperspectral modalities through the design of CFEM and CMIF, along with the incorporation of CMA. This approach has achieved remarkable classification performance in identifying varying mold contamination levels in tobacco. The research outcomes provide a solid theoretical foundation and technical support for developing intelligent multimodal detection methodologies. They also hold significant reference value for advancing quality inspection technologies both within the tobacco industry and in other fields.

Acknowledgments

This work was supported by the Science and Technology Program of Jilin Tobacco Industry Co., Ltd. (HTGS202540130).

References

- 1 J. Zhou, Y. Cheng, L. Yu, J. Zhang, and X. Zou: Appl. Microbiol. Biotechnol. **106** (2022) 131. <https://doi.org/10.1007/s00253-021-11703-2>
- 2 K. Fu, X. Song, Y. Cui, Q. Zhou, Y. Yin, J. Zhang, and Y. Su: Front. Microbiol. **15** (2024) 1399777. <https://doi.org/10.3389/fmicb.2024.1399777>
- 3 J. K. Bernat, K. J. Jackson, E. J. Krüsemann, S. Boesveldt, S. F. Rudy, and R. Talhout: Tobacco Control **32** (2023) e95. <https://doi.org/10.1136/tobaccocontrol-2021-056681>
- 4 Y. Shi, M. Liu, A. Sun, J. Liu, and H. Men: IEEE Sens. J. **21** (2021) 21175. <https://doi.org/10.1109/JSEN.2021.3079424>
- 5 C. Chotimah, K. Saifullah, F. N. Laily, M. Puspita, K. O. Kombo, S. N. Hidayat, and K. Triyana: J. Food Meas. Charact. **18** (2024) 8825. <https://doi.org/10.1007/s11694-024-02847-6>
- 6 D. Li, H. Jiang, G. Yang, Z. Gong, and T. Wen: LWT Food Sci. Technol. **210** (2024) 116822. <https://doi.org/10.1016/j.lwt.2024.116822>
- 7 W. Jia, H. Lv, Y. Liu, W. Zhou, Y. Qin, and J. Ma: Food Sci. Nutr. **12** (2024) 9856. <https://doi.org/10.1002/fsn3.3910>
- 8 Y. Cui, Y. Yao, R. Yang, Y. Wang, J. Liang, S. Ouyang, and Y. Yan: Molecules. **18** (2023) 6051. <https://doi.org/10.3390/molecules28166051>
- 9 X. Tang and N. Wang: Sens. Actuators, A. **379** (2024) 115930. <https://doi.org/10.1016/j.sna.2024.115930>
- 10 L. Singh, O. Mutanga, P. Mafongoya, K. Peerbhay, and J. Crous: Remote Sens. Appl.: Soc. Environ. **25** (2022) 100676. <https://doi.org/10.1016/j.rsase.2021.100676>
- 11 J. J. Park, S. K. Park, D. Y. Yun, G. Lee, S. S. Kim, K. J. Park, and J. S. Cho: Food Biosci. **66** (2025) 106272. <https://doi.org/10.1016/j.fbio.2025.106272>
- 12 M. L. Calderini, S. Pääkkönen, A. Yli-Tuomola, H. Timilsina, K. Pulkkinen, I. Pölönen, and P. Salmi: Algal Res. **87** (2025) 103979. <https://doi.org/10.1016/j.algal.2025.103979>
- 13 S. Li, L. Sun, X. Jin, G. Feng, L. Zhang, and H. Bai: Microchem. J. **211** (2025) 113133. <https://doi.org/10.1016/j.microc.2025.113133>
- 14 M. Rashvand, G. Paterna, S. Laveglia, H. Zhang, A. Shenfield, T. Gioia, and F. Genovese: J. Food Compos. Anal. **142** (2025) 107424. <https://doi.org/10.1016/j.jfca.2025.107424>
- 15 M. W. Ahmed, A. Sprigler, J. L. Emmert, R. N. Dilger, G. Chowdhary, and M. Kamruzzaman: Smart Agric. Technol. **10** (2025) 100857. <https://doi.org/10.1016/j.atech.2025.100857>
- 16 X. Sun, T. Li, Z. Fang, X. Yuan, Q. Li, W. Fang, S. Li, S. Li, L. Zhu, Y. Du, K. Yan, H. Luo, and Q. Ma: LWT Food Sci. Technol. **249** (2026) 119408. <https://doi.org/10.1016/j.lwt.2026.119408>
- 17 Z. Ren, K. Liang, Y. Liu, X. Wu, C. Zhang, X. Mei, and Y. Zhang: Agriculture. **15** (2025) 415. <https://doi.org/10.3390/agriculture15040415>
- 18 W. Zhang, Y. Xu, J. Liu, X. Li, H. Zheng, W. Jin, and A. M. Abd El-Aty: Front. Nutr. **13** (2026) 1879933. <https://doi.org/10.3389/fnut.2026.1879933>
- 19 F. Hu, A. Zhang, Z. L. Ji, K. Thakur, J. G. Zhang, and Z. J. Wei: Food Chem.: X **29** (2025) 102820. <https://doi.org/10.1016/j.fochx.2025.102820>
- 20 Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner: Proc. IEEE. **86** (2002) 2278. <https://doi.org/10.1109/5.726791>
- 21 A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam: Proc. 2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR, 2017) 2261–2269.
- 22 M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen: Proc. 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR, 2018) 4510–4520.
- 23 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin: Proc. 2017 Neural Information Processing Systems (NeurIPS, 2017) 1–15.
- 24 A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby: Proc. 2021 Int. Conf. Learning Representations (ICLR, 2021) 1–22.
- 25 K. He, X. Zhang, S. Ren, and J. Sun: Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition (CVPR, 2016) 770–778.