International Journal of Food Microbiology 423 (2024) 110831

Contents lists available at ScienceDirect

International Journal of Food Microbiology

journal homepage: www.elsevier.com/locate/ijfoodmicro

Enhanced detection of *Aspergillus flavus* in peanut kernels using a multi-scale attention transformer (MSAT): Advancements in food safety and contamination analysis

Zhen Guo [a,b,c], Jing Zhang [a], Haifang Wang [d], Haowei Dong [a,b,c], Shiling Li [a,b,c], Xijun Shao [a,b,c], Jingcheng Huang [a,b,c], Xiang Yin [a], Qi Zhang [e], Yemin Guo [a,b,c,*], Xia Sun [a,b,c,*], Ibrahim Darwish [f]

[a] School of Agricultural Engineering and Food Science, Shandong University of Technology, No. 266 Xincun Xilu, Zibo, Shandong 255049, China
[b] Shandong Provincial Engineering Research Center of Vegetable Safety and Quality Traceability, No. 266 Xincun Xilu, Zibo, Shandong 255049, China
[c] Zibo City Key Laboratory of Agricultural Product Safety Traceability, No. 266 Xincun Xilu, Zibo, Shandong 255049, China
[d] Dongzhimen Hospital, Beijing University of Chinese Medicine, Beijing 100700, China
[e] Oil Crops Research Institute, Chinese Academy of Agricultural Sciences, Wuhan 430062, China
[f] Department of Pharmaceutical Chemistry, College of Pharmacy, King Saud University, P.O. Box 2457, Riyadh 11451, Saudi Arabia

ARTICLE INFO

Keywords:
Multi-scale attention
Transformer
Peanut kernels
Hyperspectral imaging
Aspergillus flavus

ABSTRACT

In this study, a multi-scale attention transformer (MSAT) was coupled with hyperspectral imaging for classifying peanut kernels contaminated with diverse *Aspergillus flavus* fungi. The results underscored that the MSAT significantly outperformed classic deep learning models, due to its sophisticated multi-scale attention mechanism which enhanced its classification capabilities. The multi-scale attention mechanism was utilized by employing several multi-head attention layers to focus on both fine-scale and broad-scale features. It also integrated a series of scale processing layers to capture features at different resolutions and incorporated a self-attention mechanism to integrate information across different levels. The MSAT model achieved outstanding performance in different classification tasks, particularly in distinguishing healthy peanut kernels from those contaminated with aflatoxigenic fungi, with test accuracy achieving 98.42±0.22%. However, it faced challenges in differentiating peanut kernels contaminated with aflatoxigenic fungi from those with non-aflatoxigenic contamination. Visualization of attention weights explicitly revealed that the MSAT model's multi-scale attention mechanism progressively refined its focus from broad spatial-spectral features to more specialized signatures. Overall, the MSAT model's advanced processing capabilities marked a notable advancement in the field of food quality safety, offering a robust and reliable tool for the rapid and accurate detection of *Aspergillus flavus* contaminations in food.

1. Introduction

Aflatoxins are secondary metabolites and highly toxic substances primarily produced by *Aspergillus flavus* and *Aspergillus parasiticus* (Bertani et al., 2024). Among them, aflatoxin B_1 (AFB_1) is the most toxic aflatoxin, which has been classified as a group I carcinogenic compound by the International Agency Research on Cancer (Romero-Sánchez et al., 2024). *Aspergillus flavus* is widely present in soil, plants, and nuts with peanut kernels being particularly susceptible to contamination (Salano et al., 2024; Yao et al., 2021). The infection of peanut kernels by *Aspergillus flavus* is a continuous process. In this process, nutrients like sugars, lipids, and proteins are metabolized by the fungus, resulting in persistent changes in the peanut kernel's tissue structure and composition due to the production of secondary metabolites like aflatoxins (Guo et al., 2023a; Achar et al., 2009). Many countries and organizations have established maximum levels for AFB_1 in peanut kernels to minimize human exposure to aflatoxins, with the European Union allowing a maximum of 2 μg/kg and China allowing a maximum of 20 μg/kg (Thati

* Corresponding authors at: School of Agricultural Engineering and Food Science, Shandong University of Technology, No. 266 Xincun Xilu, Zibo, Shandong 255049, China.
E-mail addresses: gym@sdut.edu.cn (Y. Guo), sunxia2151@sina.com (X. Sun).

https://doi.org/10.1016/j.ijfoodmicro.2024.110831
Received 1 April 2024; Received in revised form 18 June 2024; Accepted 18 July 2024
Available online 20 July 2024

et al., 2024; Ren et al., 2023). Accurate identification of contaminated peanut kernels is crucial to prevent aflatoxins from entering the food chain.

Traditional techniques for detecting fungal infections in food and agricultural products typically rely on microbiological methods, including culturing and plating methods, enzyme-linked immunosorbent assay, and serologically specific electron microscopy (Makarichian et al., 2022). These methods are known for their specificity and reliability in identifying fungal species. However, they are costly, time-consuming, and often require trained personnel for sample preparation making them unsuitable for large-scale, non-destructive screening or integration into online sorting systems. Hyperspectral imaging combined with chemometric and deep learning techniques is increasingly applied to detect fungal contamination and mycotoxins in food and agricultural products (Liu et al., 2020; Wu et al., 2020; Mishra et al., 2022; Lu et al., 2022; Weng et al., 2021). Short-wave infrared (SWIR) (1000–2500 nm) hyperspectral imaging provides vibrational information about overtones and combination bands in the infrared region related to chemical bonds like C—H, N—H, and O—H (Kimuli et al., 2018). Changes in the chemical composition and organizational structure of peanut kernels infected by *Aspergillus flavus* are accompanied by alterations in compound and functional group information, offering feasibility for SWIR hyperspectral imaging to detect this fungal contamination.

In most reported studies, samples infected with aflatoxigenic fungi are detected, which produce aflatoxins, overlooking the detection of non-aflatoxigenic fungal infections. Few studies have reported on the use of hyperspectral imaging to identify non-aflatoxigenic fungal infections. A study reports on using SWIR hyperspectral imaging to distinguish *Aspergillus flavus* contamination in corn kernels. The corn kernels are inoculated with aflatoxigenic fungi (AF13), non-aflatoxigenic fungi (AF36), and sterile distilled water. The competitive adaptive reweighted sampling (CARS)-partial least squares-discriminant analysis (PLS-DA) model shows overall accuracy of 89.8% and 89.3% using a 20-ppb concentration threshold and a 100-ppb concentration threshold, respectively (Tao et al., 2022). Another study reports on applying SWIR hyperspectral imaging (900–2500 nm) to distinguish corn kernels inoculated with aflatoxigenic fungi (AF13) and non-aflatoxigenic fungi (AF36) from healthy kernels. Based on complete average spectra extracted from the same kernel side, the PLS-DA shows the best overall prediction accuracy of 96.3% for the three categories and 97.8% for the aflatoxin-negative and aflatoxin-positive classes (Tao et al., 2020). In fact, aflatoxigenic fungi and non-aflatoxigenic fungi coexist in crops, seeds, or soil (Alaniz Zanon et al., 2018; Yin et al., 2009; Abbas et al., 2005). Moreover, non-aflatoxigenic strains of *Aspergillus flavus* are widely used to reduce aflatoxin contamination in peanut kernels, corn, and nuts (Hulikunte Mallikarjunaiah et al., 2017; Weaver and Abbas, 2019; Alaniz Zanon et al., 2016). Hence, detecting both aflatoxigenic fungal infection and non-aflatoxigenic fungal infection in peanut kernels is necessary.

In recent years, deep learning algorithms have offered new methods for fully exploiting hyperspectral image information. A 1-dimensional convolutional neural network (CNN) is used for pixel-level classification of aflatoxin contamination (Gao et al., 2021), and CNN-based pixel-spectral reshaping methods assess aflatoxin contamination in peanut kernels (Han and Gao, 2019). Since the introduction of Vision-Transformers, transformers have been effective alternatives to CNNs in tasks like image recognition and object detection (Azad et al., 2024). A recent study employs a 3-dimensional-convolutional neural network (3D-CNN) with a 2-dimensional-convolutional neural network (2D-CNN) to capture spectral and spatial features of hyperspectral images, using Gaussian-weighted feature tokenizers to convert extracted features, achieving an overall accuracy of 97% in identifying the growth years of Kudzu root (Xu et al., 2023). Another study introduces that hyperspectral imaging combined with a spectrum transformer network (STNet) is employed to identify tomato bacterial wilt severity (Wang et al., 2023). The STNet achieves an average F1-score of 0.9309, an overall accuracy of 91.93%, and a kappa coefficient of 0.8903. However, no study has yet reported on the application of transformers in detecting peanut kernels contamination by aflatoxigenic fungi or non-aflatoxigenic fungi.

Therefore, the objectives of this study are as follows. (1) Propose a multi-scale attention transformer (MSAT) to be coupled with hyperspectral imaging for classifying peanut kernels contaminated with diverse *Aspergillus flavus* fungi. (2) Innovatively design a multi-scale attention mechanism employing several multi-head attention layers to focus on both fine-scale and broad-scale features, integrating a series of scale processing layers to capture features at different resolutions and incorporating a self-attention mechanism to integrate information across different levels. (3) Design a series of experiments to find the optimal hyperparameter combination of MSAT and evaluate the performance of MSAT in different classification tasks. (4) Utilize attention weights visualization of attention heads across different layers to provide insights into how the model focused on various spatial-spectral features at different scales.

2. Materials

2.1. Samples preparation

Peanut kernels of the Weihua and Baisha varieties were sourced from a Zibo supermarket in China. A total of 2160 kernels were distributed into 2 contaminated groups and a control group, respectively. Sterilization involved immersing all samples in a 75% ethanol solution for 1 min, followed by triple rinsing in sterile water and subsequent placement in a sterile environment. Aflatoxigenic *Aspergillus flavus* (ATCC#28539) was acquired from China's National Strain Center, which produced aflatoxin B_1, B_2, G_1, and G_2. Non-aflatoxigenic *Aspergillus flavus* (CTCC#2020519) was obtained from the Oil Crops Research Institute, Chinese Academy of Agricultural Sciences, which did not produce aflatoxin. *Aspergillus flavus* was grown on PDA medium at 28°C for 5 days. Spores from this culture were then collected and adjusted to a concentration of 1×10^6 CFU/mL using sterile water, and used for the contamination of the peanut kernels. A total of 720 and 360 peanut kernels were used as the contaminated group 1 and contaminated group 2, which were incubated with aflatoxigenic fungi (ATCC#28539) or non-aflatoxigenic fungi (CTCC#2020519), respectively. The remaining 1080 peanut kernels inoculated with sterile water were used as the control group. These kernels were incubated until the 6th day under conditions set to 30°C and 85% relative humidity (Long et al., 2022). 180, 120 and 60 kernels were obtained daily to acquire hyperspectral images from the 1st day to the 6th days, which were selected from the control group, contaminated group 1 and contaminated group 2, respectively.

2.2. Aflatoxin detection

Methanol and acetonitrile (chromatography grade) were purchased from Tedia Company, Inc. (Fairfield, USA). NaCl, Na_2HPO_4, KH_2PO_4, KCl, HCl, Triton X-100 [$C_{14}H_{22}O$ $(C_2H_4O)_n$] and I_2 were obtained from Sinopharm Chemical Reagent Co., Ltd. (Shanghai, China). AFB_1, Aflatoxin B_2 (AFB_2), Aflatoxin G_1 (AFG_1) and Aflatoxin G_2 (AFG_2) were purchased from Pribolab (Qingdao, China). Seven concentrations of standard solutions were prepared for the establishment of calibration curves. The concentrations of AFB_1 and AFG_1 were set at 0.1, 0.5, 2.0, 5.0, 10.0, 20.0, and 40.0 μg/kg, respectively. For AFB_2 and AFG_2, the concentrations were established at 0.03, 0.15, 0.6, 1.5, 3.0, 6.0, and 12.0 μg/kg, respectively. Recovery experiments were conducted by spiking peanut kernels samples with 4 distinct concentrations of each aflatoxin. Concentrations for AFB_1 and AFG_1 were established at 5.0, 10.0, 20.0, and 40.0 μg/kg, while for AFB_2 and AFG_2, they were set at 1.5, 3.0, 6.0, and 12.0 μg/kg, respectively. Six independent replicates

were prepared for each concentration level.

Peanut kernel samples were crushed and sieved. Exactly 5 g of the samples was weighed and placed into a 50 mL centrifuge tube. Then, 20 mL of an acetonitrile-water solution was also added to the centrifuge tube. The mixture was homogenized using a vortex and then subjected to ultrasonic oscillation for 20 min. After oscillation, the suspension was centrifuged at 6000 r/min for 10 min. A volume of 4 mL of the supernatant was then transferred and mixed with 46 mL of a 1% Triton X-100 PBS solution for further analysis. Aflatoxins were extracted using immunoaffinity columns and analyzed using high-performance liquid chromatography (HPLC) system.

The HPLC system (1260 Infinity II LC System, Agilent Technologies, Santa Clara, CA) was used to measure the content of aflatoxin in peanut kernels after acquiring hyperspectral images. Detailed information on the detection method can be found in the publication by Campos et al. (2017). An analytical column (4.6 mm × 150 mm with particle size of 4 μm) (Poroshell 120, Agilent Technologies, Santa Clara, CA) was used to achieve HPLC separation and quantitative analysis. The mobile phase consisted of acetonitrile: methanol: water (16:16:68, v/v/v) at a flow rate of 1.0 mL/min and a temperature of 40°C. Aflatoxins were detected and quantified with fluorescence detection at 360 nm (excitation) and 440 nm (emission) wavelengths. 4 peanut kernels were randomly selected from each group daily to detect the content of aflatoxins by HPLC from the 1st day to the 6th days. There were no aflatoxins existing in the control group and contaminated group 2. This indicated that the non-aflatoxigenic fungi did not produce aflatoxins. Table 1 showed the aflatoxins content of peanut kernels in contaminated group 1. The content of AFB_1 in peanuts was higher than that of AFB_2, AFG_1 and AFG_2, and almost no AFG_1 was produced by the aflatoxigenic fungi.

As shown in Table 2, the recoveries of AFB_1 content in peanut kernels ranged from 95.14% to 101.78%. Seven concentrations of AFB_1 standard solutions (0.1, 0.5, 2.0, 5.0, 10.0, 20.0, and 40.0 μg/kg) were used to establish a calibration curve. The linear regression equation for AFB_1 was $y = 0.2226x + 0.1182$, where y was the concentration (μg/kg) and x was the peak area, with correlation coefficient of 0.9998, and a limit of detection of 0.06 μg/kg (Table 3).

2.3. Hyperspectral image acquisition and processing

The hyperspectral images were acquired using a SWIR hyperspectral imaging system (Isuzu Optics Corp., Taiwan, China). Each hyperspectral image contained 288 spectral bands ranging from 1000 to 2500 nm. The imaging process involved setting the exposure time to 2.7 ms and the mobile platform speed to 14.7 mm/s. Both a black reference image and a white reference image were utilized to calibrate the raw hyperspectral images, reducing the ambient noise impact (Guo et al., 2023b). Each hyperspectral image contained 60 peanut kernel images, yielding a total of 36 hyperspectral images.

A band ratio image (1133.1 nm gray image/1936.0 nm gray image) was used to generate a mask image using the OTSU method. The mask image was multiplied by each band in the calibrated hyperspectral image, resulting in a mask hyperspectral image only containing the peanut kernels. Peanut kernels were labeled one by one and extracted individually as the region of interest (ROI). Table 4 showed that the average height and the average with of ROI for peanut kernels were 56.43 pixel and 31.34 pixel. Therefore, the ROI size of peanut kernel was set to 60 × 30. Principal component analysis (PCA) was employed to reduce the dimension of the individual peanut kernel hyperspectral images. The number of components was fixed at 15 in PCA. Therefore, the individual peanut kernel hyperspectral images (60 × 30 × 15) were applied in the extraction of spatial-spectral features.

Table 1
Aflatoxins content in the contaminated group 1.

Aflatoxins variety	Range (μg/kg)	Mean (μg/kg)	SD (μg/kg)
AFB_1	0.00–232.35	24.41	39.48
AFB_2	0.00–61.27	6.26	9.66
AFG_1	0.00–3.62	0.72	0.76

Table 2
Recoveries of AFB_1, AFB_2, AFG_1 and AFG_2 from peanut kernel samples.

Aflatoxins	Added (μg/kg)	Determine ± SD (μg/kg)	Recovery (%)
AFB_1	5.0	4.757±0.090	95.14
	10.0	9.755±0.150	97.55
	20.0	20.355±0.210	101.78
	40.0	40.030±0.210	100.08
AFB_2	1.5	1.445±0.110	96.33
	3.0	2.955±0.120	98.50
	6.0	6.135±0.180	102.25
	12.0	12.040±0.175	100.33
AFG_1	5.0	4.535±0.135	90.70
	10.0	9.295±0.190	92.95
	20.0	19.705±0.210	98.53
	40.0	40.505±0.230	101.26
AFG_2	1.5	1.390±0.110	92.67
	3.0	2.835±0.100	94.50
	6.0	5.960±0.150	99.33
	12.0	12.175±0.155	101.46

Table 3
Linearity parameters and detection limits of HPLC methods.

Aflatoxins	Linear regression equation	Correlation coefficient	Limit of detection (μg/kg)
AFB_1	$y = 0.2226x + 0.1182$	0.9998	0.06
AFB_2	$y = 0.1203x + 0.0183$	0.9997	0.05
AFG_1	$y = 0.4814x + 0.3487$	0.9996	0.07
AFG_2	$y = 0.3562x + 0.0794$	0.9997	0.05

Table 4
Peanut sample ROI size.

Sample	Index	Mean	Max.	Min.	SD
Weihua	Pixel size	1581.23	2374.87	526.54	253.81
	Height	64.67	88.82	32.57	8.37
	Width	32.14	56.45	18.92	4.03
Baisha	Pixel size	1126.14	2047.23	423.61	282.13
	Height	48.19	79.29	24.67	8.58
	Width	30.54	50.46	16.86	4.48
Total	Pixel size	1353.69	2374.87	423.61	327.96
	Height	56.43	88.82	24.67	11.52
	Width	31.34	56.45	16.86	4.28

3. Methods

3.1. Spatial-spectral feature extraction

A complex CNN model was employed to obtain the spatial-spectral features of the individual peanut kernel hyperspectral images as illustrated in Fig. 1. The CNN model consisted of a succession of layers including two 3D convolutional layers, a 3D max pooling layer, two 2D convolutional layers, a 2D max pooling layer and two reshape operation layers. The CNN model utilized an Adam optimizer and a cross-entropy loss function, and included an accuracy metric for performance evaluation.

The model's inception involved the processing of an input tensor $X_0 \in \mathbb{R}^{m \times n \times p}$ through its first 3D convolutional layer. The numbers of the m, n and p were 60, 30 and 15, respectively. This layer was equipped with 16 filters of dimensions 5 × 5 × 3 and applied "same" padding. It preserved the spatial dimensions of X_0, producing an output tensor $X_1 \in \mathbb{R}^{m \times n \times p \times 16}$. Following this, a second 3D convolutional layer with 32 filters, each of size 3 × 3 × 3, further processed X_1 and yielded $X_2 \in \mathbb{R}^{m \times n \times p \times 32}$. The model then incorporated a 3D max pooling layer,

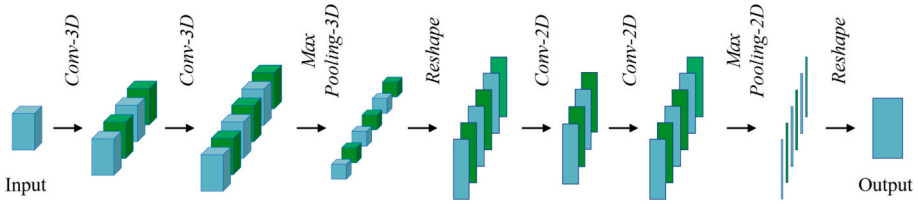

Fig. 1. CNN model architecture.

with a pool size of $6 \times 3 \times 3$, which reduced the spatial dimensions of X_2, resulting in $X_3 \in \mathbb{R}^{\frac{m}{6} \times \frac{n}{3} \times \frac{p}{3} \times 32}$. Then, the data was reshaped to be compatible with 2D convolutional layers by the first reshape layer, producing an output tensor $X_4 \in \mathbb{R}^{\frac{mn}{18} \times \frac{p}{3} \times 32}$. Utilizing 16 filters of size 5×5 and maintaining "same" padding, the first 2D convolutional layer transformed $X_5 \in \mathbb{R}^{\frac{mn}{18} \times \frac{p}{3} \times 16}$. This was succeeded by the second 2D layer with 32 filters of 3×3, leading to $X_6 \in \mathbb{R}^{\frac{mn}{18} \times \frac{p}{3} \times 32}$. Following this, a 2D max pooling layer with a pool size of 1×5 transformed X_6 into $X_7 \in \mathbb{R}^{\frac{mn}{18} \times \frac{p}{15} \times 32}$. Finally, the second reshape operation layer reshaped the X_7 into $X_8 \in \mathbb{R}^{\frac{mn}{18} \times 32}$, comprising 100 features, each with 32 channels. This sequential and structured approach emphasized spatial-spectral features extraction through a cascade of convolutional layers and pooling layers. It illustrated the CNN model's capacity to transform the input data into a nuanced feature representation.

3.2. Multi-head attention mechanism

Fig. 2 showed the structure of multi-head attention layer (Vaswani et al., 2017; Zheng et al., 2019). For each attention head h, the input tensor X was projected to a set of queries (Q_h), keys (K_h) and values (V_h) using trainable weight matrices $W_h^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_h^K \in \mathbb{R}^{d_{model} \times d_k}$ and $W_h^V \in \mathbb{R}^{d_{model} \times d_k}$, where d_{model} was the dimension of the X, d_k was the dimension of the key vector. This yielded $Q_h = XW_h^Q$, $K_h = XW_h^K$ and $V_h = XW_h^V$. The attention for each head was then computed as:

$$Attention_h(Q_h, K_h, V_h) = softmax\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right) V_h \quad (1)$$

where, $Attention_h(Q_h, K_h, V_h)$ was the attention for each head, *softmax* was applied over the keys for each query, and the d_k was used as a scaling factor to prevent the *softmax* function from entering regions.

The outputs of each head were concatenated and once again linearly transformed with a final weight matrix $W^O \in \mathbb{R}^{d_{model} \times d_{model}}$:

$$MHA(X) = Concat(Attention_1, Attention_2, \ldots, Attention_h) W^O \quad (2)$$

where, $MHA(X)$ was multi-head attention applied to the input X.

3.3. Multi-scale attention module

In this study, a multi-scale attention module was ingeniously architected, as shown in Fig. 3, integrating multiple multi-head attention

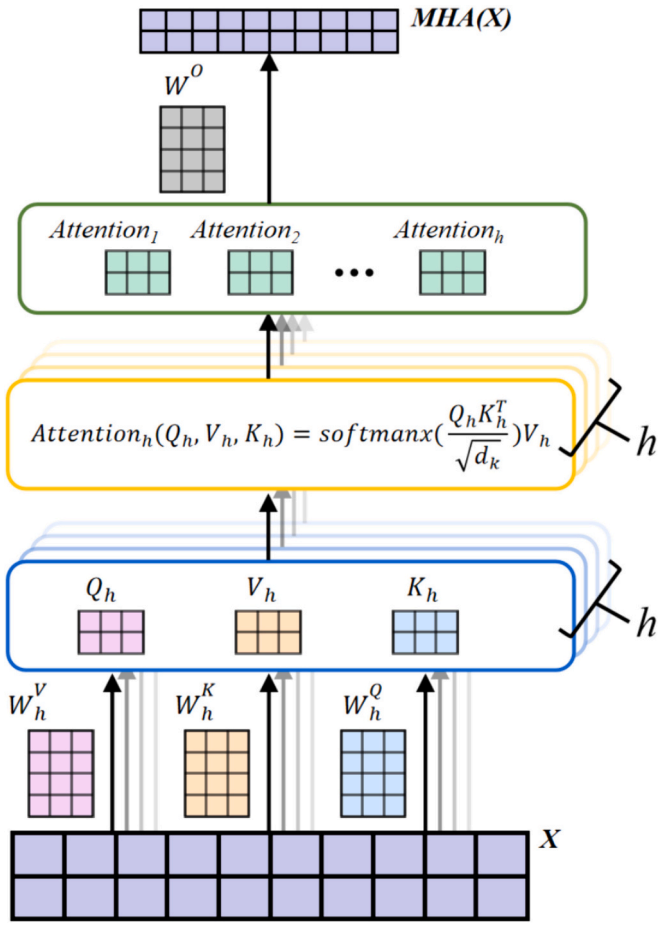

Fig. 2. Structure of multi-head attention layer.

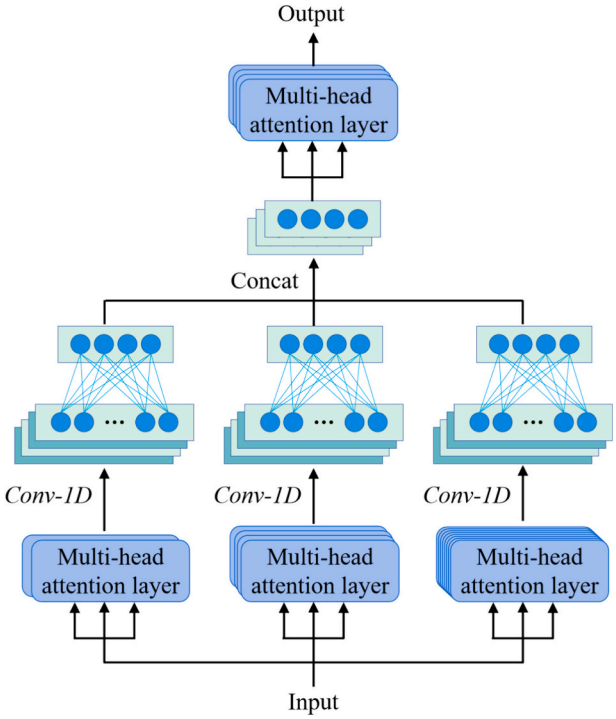

Fig. 3. Multi-scale attention module.

layers to distill features at varied scales from the input data tensor $X_1 \in \mathbb{R}^{s \times d_{model}}$. Where, s denoted the sequence length and d_{model} denoted the feature dimensionality.

Initially, each attention layer was parameterized with a distinct configuration of heads and key dimensions to cater to different information granularities within X_1. The first multi-head attention layer targeted fine-grained features, designed with $\frac{num\_heads}{2}$ attention heads and a key dimension of $\frac{d_{model}}{2}$. In contrast, the second multi-head attention layer and the third one addressed coarser scales of the input features, with *num_heads* and $2 \times$ *num_heads*, and key dimension of d_{model} and $2 \times d_{model}$, respectively. Subsequently, 3 processing layers operated on the outputs of the corresponding multi-head attention layers, and each scale processing layer consisted of a 1D convolution layer and a fully connected layer. These 1D convolution layers facilitated a diverse capture of feature information, with kernel sizes spanning 1, 3 and 5, respectively. Each scale processing layer produced outputs that preserved the original feature dimension d_{model} post-activation. The outputs of these scale processing layers were concatenated along the feature dimension, forming an enriched composite representation:

$$X_2 = \bigoplus_{i=1}^{3} Dense_i(Conv1D_i(MHA_i(X))) \quad (3)$$

where, $\bigoplus$ denoted the concatenation operation along the feature dimension, MHA_i represented the output of the i-th multi-head attention layer applied to X, $Conv1D_i$ represented the i-th 1D convolution layer, $Dense_i$ denoted the i-th fully connected layer applied to the output of $Conv1D_i$.

The final representation $X_2 \in \mathbb{R}^{s \times 3d_{model}}$ was achieved by concatenating the outputs from the three distinct processing paths. This enriched representation was subjected to a self-attention mechanism through another multi-head attention layer, designed with attention heads of *num_heads* and a key dimension of d_{model}, producing output X_3. The output was mathematically described by the equation:

$$X_3 = MHA_{num\_heads}(X_2) \quad (4)$$

where, the $X_3 \in \mathbb{R}^{s \times 3d_{model}}$, MHA_{num_heads} denoted multi-head attention with a specific number of heads.

The self-attention mechanism enabled the model to integrate information from different representation levels, considering the entire sequence information to generate a refined output. Finally, a final projection was applied to X_3 via a fully connected layer, mapping it back to the original dimensionality d_{model} and yielding X_4. This projection was mathematically described by the equation:

$$X_4 = X_3 W^O + b^O \quad (5)$$

where, $W^O \in \mathbb{R}^{3d_{model} \times d_{model}}$ and $b^O \in \mathbb{R}^{d_{model}}$ were the trainable parameters of the final projection layer, and $X_4 \in \mathbb{R}^{s \times d_{model}}$ was the resulting tensor that retained the original feature dimensionality post-transformation.

The multi-scale attention module was engineered to harness multi-head attention mechanisms across varied scales, effectively capturing dependencies in the input data X. It aggregated features through convolutional processing at multiple resolutions, integrated via self-attention mechanism, and refined the composite representation for subsequent processing.

3.4. Multi-scale attention transformer

Fig. 4 showed the architectures of the MSAT, including an encode layer, a 1D global average pooling layer and a fully connected layer. The encode layer consisted of a multi-scale attention module and a feed-forward neural network (FNN). The multi-scale attention module and FNN incorporated layer dropout and normalization for regularization. The FNN consisted of two linear transformations with a rectified linear unit (ReLU) activation in between. The output from the encoder layer

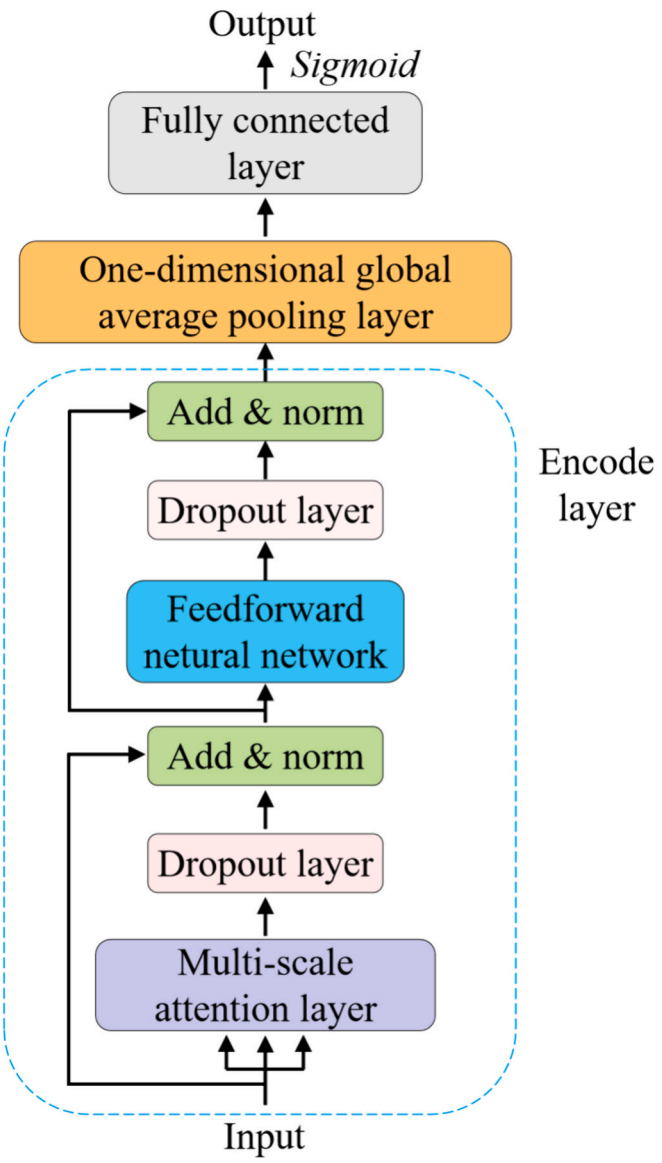

Fig. 4. Multi-scale attention transformer architecture.

underwent global average pooling, reducing the sequence dimension and preparing the feature representation for the classification layer. A fully connected layer with an activation function was employed to produce the final output. This entire process encapsulated the forward pass of the proposed MSAT model, which harnessed the multi-scale attention mechanism to effectively capture and process spectral-spatial features within hyperspectral images. The MSAT architectural design and operational flow were conducive to the extraction of intricate patterns and dependencies, facilitating the precise classification of the peanut kernels contaminated with diverse *Aspergillus flavus* fungi.

3.5. Experiment settings

The processing of hyperspectral images and the execution of models were conducted in a Python 3.7.10 environment, utilizing TensorFlow 2.10.0. For the purpose of training, experiments utilized the Adam optimizer, with the initial learning rate set at 0.001. The cross-entropy loss function was employed in the classification experiments. The batch size and training epochs were determined to be 128 and 50, respectively. Samples were randomly allocated into a training set and a testing set in a quantity ratio of 3:1, with 5-fold cross-validation being

implemented during the training phase. Five performance indices were employed to assess and compare performance, including test accuracy, test precision, test recall, test F1-score, and model run time (runtime). To provide a more objective evaluation and validation of the models, each experiment was repeated 10 times, and the mean and standard deviation of each metric were reported.

4. Results and discussion

4.1. Spectral analysis

Fig. 5(a) depicted that distinct spectral curves were observed across healthy peanut kernels (control group), aflatoxigenic fungus contamination peanut kernels (contaminated group 1) and non-aflatoxigenic fungus contamination peanut kernels (contaminated group 2). Healthy peanut kernels displayed a spectral profile with lower reflectance in the range of 1000–2500 nm. In contrast, peanut kernels infected with aflatoxigenic *Aspergillus flavus* showed increased reflectance which was in agreement with the previous works (Yuan et al., 2020; Jiang et al., 2016). It suggested alterations in moisture content, fat structures, and protein degradation due to the metabolic changes induced by the fungal infection and aflatoxin production. Non-aflatoxigenic infections also differed from the healthy peanut kernel spectrum, with distinct changes at 1114, 1302, 1861 and 2019 nm. Further analysis correlated specific wavelength bands with biochemical and structural changes. Variations in fat were revealed around 1126 nm to 1180 nm (Bilal et al., 2020). Variations in water content were observed around 1450 nm and 1940 nm corresponding to O—H stretching and bending vibrations in water molecules (Sundaram et al., 2012). Alterations in oils were identified by changes around 1751 nm, which were attributed to C—H stretching vibrations (Wang and Cheng, 2018). Protein and amino acid changes were indicated at 2136 nm which was associated with N—H stretching and C═O stretching (Williams et al., 2012). Moreover, significant reflectance differences were revealed between the contaminated group 1 and contaminated group 2, with the former generally showing higher reflectance. This was indicative of the unique biochemical alterations caused by the respective fungal strains. Finally, Fig. 5(b) showed the spectral curves of the 2160 peanut kernel samples.

4.2. Bayesian optimization

In the MSAT model architecture, the number of encoder layers used in the model was specified by num_layers, and the count of attention heads was determined by num_heads. The dimension of the FNN within the encoder layers, denoted as Dff, set the dimension of the FNN. The dropout rate defined the dropout ratio applied to various layers during training to prevent overfitting. These 4 hyperparameters were crucial for optimizing model performance, and their correct configuration required experimental determination. Consequently, the MSAT was optimized using Bayesian optimization via the Hyperopt's Tree-structured Parzen Estimator algorithm. The optimization focused on 4 key hyperparameters including num_layers, num_heads, Dff, and dropout rate. The objective function assessed model efficacy based on maximal validation accuracy, facilitating 200 experiments to identify the optimal configuration. Num_layers offered choices among 1 layer, 2 layers, and 3 layers, and num_heads included options for 2, 4, and 8 attention heads. The Dff parameter was set to select from 64, 128, or 256, and the dropout rate was configured to choose from 0.1, 0.3, or 0.5. This structured approach enabled a comprehensive exploration of the model's architecture within the defined bounds, ensuring a thorough investigation of the potential configurations for optimal performance.

Fig. 6(a) showed that when Dff was set to 128, 256, and 512, the median values of the test accuracy metric were 94.63%, 94.91%, and 95.00% in box plot, respectively. But the upper quartile of test accuracy was highest at 95.51% when Dff was 256. Additionally, with Dff at 128, 256, and 512, the median values of the runtime metric were 79.56 s, 39.28 s, and 46.68 s, respectively, indicating a faster runtime when Dff was 256 as shown in Fig. 6(e). Larger Dff values suggested that a more complex network could capture finer details in the data. Fig. 6(f–i) showed that when num_heads was 4, the median values of test accuracy, test precision, test recall, and test F1-score were the highest, and all of them reached at 95.19%. Furthermore, the median runtime was 39.25 s when num_heads was 4, which was smaller than the median runtimes when num_heads were 2 and 8 as displayed in Fig. 6(j). Models with a greater number of attention heads generally achieved better performance, signifying their ability to focus on more relevant features in the data. Fig. 6(k–o) demonstrated that when the number of encoder layer was set to 1, 2, and 3, the median values of the test accuracy were recorded at 94.63%, 94.81%, and 94.81%, respectively. Furthermore, the median values of the test F1-scores were also reported to be 94.63%, 94.81%, and 94.81%, respectively. These results confirmed that an increase in the number of encoder layers enhanced the testing precision and other metrics. However, the change in the runtime metric was more pronounced, with median values of 28.40 s, 39.26 s, and 79.58 s, respectively. A higher number of encoder layers brought a greater operational burden, indicating that the number of encoder layers

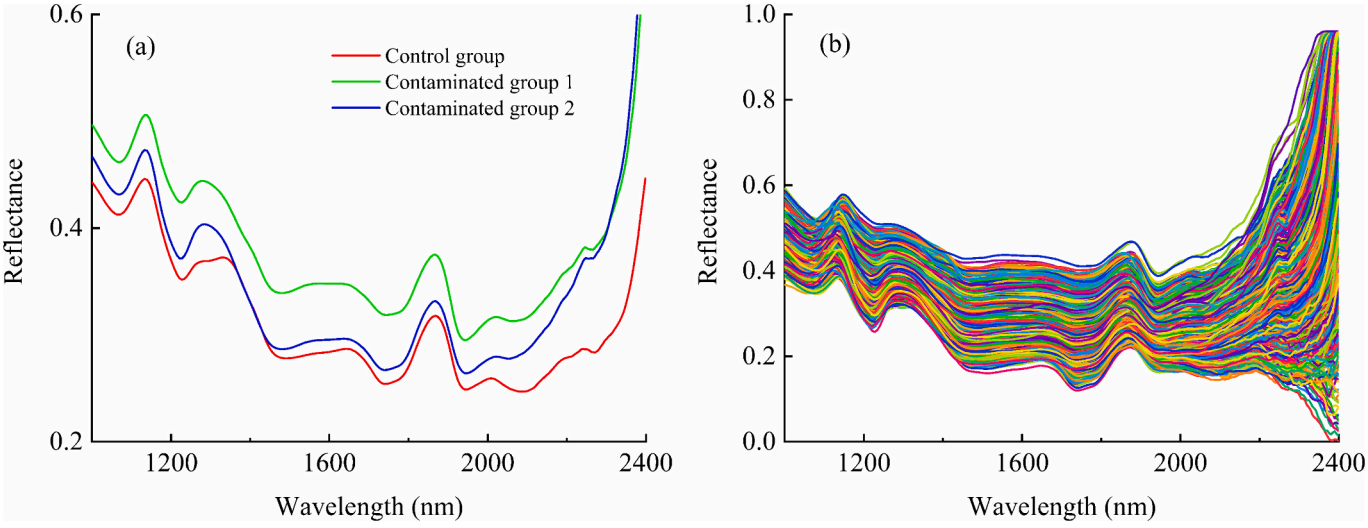

Fig. 5. Spectral curves of peanut kernels, (a) spectral curves of healthy peanut kernels (control group), aflatoxigenic fungus contamination peanut kernels (contaminated group 1) and non-aflatoxigenic fungus contamination peanut kernels (contaminated group 2), (b) spectral curves of the 2160 peanut kernels.

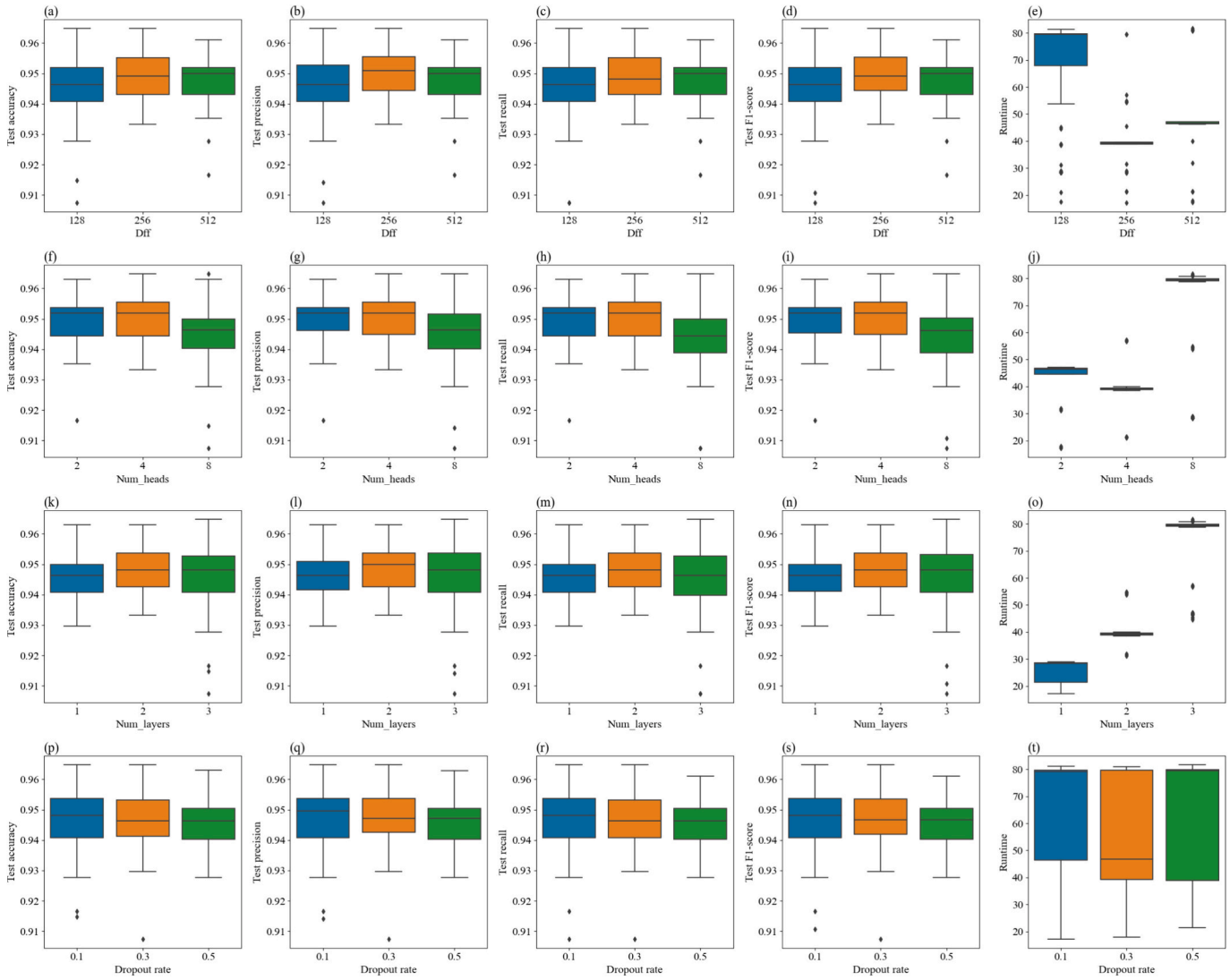

Fig. 6. Box plot of hyperparameters and classification metrics, (a) test accuracy of Dff, (b) test precision of Dff, (c) test recall of Dff, (d) test F1-score of Dff, (e) runtime of Dff, (f) test accuracy of num_heads, (g) test precision of num_heads, (h) test recall of num_heads, (i) test F1-score of num_heads, (j) runtime of num_heads, (k) test accuracy of num_layers, (l) test precision of num_layers, (m) test recall of num_layers, (n) test F1-score of num_layers, (o) runtime of num_layers, (p) test accuracy of dropout rate (q) test precision of dropout rate, (r) test recall of dropout rate, (s) test F1-score of dropout rate, (t) runtime of dropout rate.

significantly affected the model's runtime speed and computational efficiency. Finally, Fig. 6(p–t) showed that the dropout rate had no significant impact on the model's classifying ability, as the median values of all metrics were close when the dropout rate was 0.1, 0.3, and 0.5. However, the median runtime was the lowest at 46.74 s when the dropout rate was 0.3, suggesting a balance between overfitting and underfitting. In summary, after analyzing the MSAT's classification accuracy and operational efficiency, the optimal hyperparameter combination was identified as a Dff of 256, a num_heads of 4, a num_layers of 1, and a dropout rate of 0.3. These findings demonstrated the efficacy of this specific combination in enhancing model performance while ensuring computational efficiency. This study underlined the effectiveness of Bayesian optimization in fine-tuning hyperparameters in complex machine learning models, illustrating the intricate interplay between multiple hyperparameters and their impact on overall performance.

4.3. Ablation experiments

In this study, a detailed analysis was conducted on the effects of various components in the MSAT, particularly focusing on the multi-scale attention module. The analysis was structured around a series of ablation experiments in Table 5. The intact model was equipped with all its components including multi-scale attention layers, scale processing layers, multi-scale CNN layers, fully connected layers and a self-attention layer. It achieved the highest classification performance: 95.00±0.78% in test accuracy, 95.02±0.77% in test precision, 95.00 ±0.78% in test recall, and 95.01±0.78% in test F1-score. This established the collective importance of these components. After replacing the multi-scale attention layers with a single-scale attention layer, a marginal decrease in performance was observed, with test accuracy slightly dropping to 94.70±0.74%. This highlighted the value of multi-scale layers in capturing comprehensive features within the data. The absence of scale processing layers led to the most pronounced impact, with test accuracy reducing to 93.93±0.47% and test F1-score to 93.92 ±0.47%. In contrast, missing only multi-scale CNN layers or fully connected layers had less impact than that of the scale processing layers on the test accuracy, dropping to 94.04±0.60% and 94.80±1.27%, respectively. This indicated that the scale processing layers played a significant role in enhancing the model's sensitivity to diverse features in the data. Furthermore, the omission of the self-attention layer resulted in a significant decline in performance, with test accuracy reducing to

Table 5
Model performance in ablation experiments.

No.	Multi-scale attention layers	Scale processing layers	Multi-scale CNN layers	Fully connected layers	Self-attention layer	Test accuracy (%)	Test precision (%)	Test recall (%)	Test F1-score (%)	Runtime (s)
1	√	√	√	√	√	95.00±0.78	95.02±0.77	95.00±0.78	95.01±0.78	21.38±0.93
2	×	√	√	√	√	94.70±0.74	94.70±0.74	94.69±0.73	94.69±0.73	13.05±0.62
3	√	×	×	×	√	93.93±0.47	93.94±0.46	93.91±0.48	93.92±0.47	18.99±0.34
4	√	√	×	√	√	94.04±0.60	94.04±0.60	94.02±0.61	94.03±0.60	20.17±0.27
5	√	√	√	×	√	94.80±1.27	94.81±1.26	94.80±1.27	94.81±1.27	20.32±0.74
6	√	√	√	√	×	94.57±0.92	94.57±0.92	94.56±0.93	94.56±0.92	18.91±0.74

Note: × indicated the ablation of the component, and √ indicated the existence of the component.

94.57±0.92%. These findings underscored each component's contribution to the model's efficacy in classifying the peanut kernels contaminated with diverse *Aspergillus flavus* fungi. The multi-scale attention mechanism was shown to be vital for effective data processing.

4.4. Models performance in different classification tasks

4.4.1. Models performance in classifying the peanut kernels contaminated with diverse Aspergillus flavus fungi

Representative classical deep learning models often exhibited increased architectural complexity, incorporating advanced neural network structures. Consequently, a comprehensive evaluation of various representative classical deep learning algorithms was conducted to identify the peanut kernels contaminated with diverse *Aspergillus flavus* fungi. A total of 6 algorithms were analyzed in this 3-class classification task, including MSAT, DenseNet-121 (Huang et al., 2017), ResNet-50 (He et al., 2016), SqueezeNet (Iandola et al., 2016), Xception (Chollet, 2017) and CapsuleNet (Pv et al., 2019).

Table 6 revealed that MSAT was the most impressive model, consistently achieving high scores across all evaluated metrics, including a test accuracy of 95.00±0.78%, test precision of 95.02 ±0.77%, test recall of 95.01±0.78%, and test F1-score of 95.01±0.78%. Its runtime, though moderate at 21.38±0.93 s, was acceptable given its performance. Xception also exhibited commendable performance, closely mirroring MSAT in terms of test accuracy (94.72±0.81%) and test F1-score (94.76±0.76%). However, Xception had a significantly longer runtime of 100.00±0.00 s. SqueezeNet, while not as precise as MSAT or Xception, offered a favorable balance between test accuracy (91.61±4.45%) and processing speed (10.71±0.88 s), although it was not as precise as MSAT or Xception. Among the other models, DenseNet-121, ResNet-50, and CapsuleNet demonstrated significantly inferior performance compared to MSAT and Xception, especially in terms of test accuracy and test F1-score. For example, DenseNet-121 recorded a test accuracy of 86.80±10.80% and a relatively low test F1-score of 86.49 ±11.29%. CapsuleNet exhibited the weakest performance with a test accuracy of only 76.24±3.01% and a test F1-score of 73.62±3.06%. The results clearly indicated the superiority of MSAT in classifying the peanut kernels contaminated with diverse *Aspergillus flavus* fungi. Overall, MSAT and Xception stood out as the most effective models for this classification task, with MSAT providing an optimal combination of high performance and reasonable runtime.

Fig. 7(a) showed that the MSAT model demonstrated a rapid and efficient learning curve. In the initial training phase, MSAT achieved a training accuracy of 68.81%, significantly outperforming other models. As the training progressed, MSAT continued to exhibit superior performance, with a notable jump to 90.52% in the second epoch, finally achieving 94.95%. This consistency contrasted sharply with the gradual and less pronounced improvements seen in other models. Although Xception showed outstanding results, achieving the highest training accuracy of 96.28% in the 50th epoch, it started to converge around the 20th epoch. In terms of validation accuracy, the MSAT model again outshone its counterparts as shown in Fig. 7(b). Starting with an initial validation accuracy of 81.08%, it quickly escalated to around 94.41%, finally reaching a validation accuracy of 95.15%. It demonstrated not only high learning efficiency but also effective generalization capabilities. In comparison, other models like DenseNet-121 and ResNet-50 struggled to achieve similar levels of validation accuracy, indicating potential overfitting or less effective learning strategies. Coupled with high final accuracies in both training and validation phases, the MSAT's rapid convergence speed set it apart from the other models.

4.4.2. Models performance in distinguishing healthy peanut kernels and those contaminated with aflatoxigenic fungi or non-aflatoxigenic fungi

In a further examination of the MSAT, a 2-class classification task was conducted to assess its capability in distinguishing healthy peanut kernels and those contaminated with aflatoxigenic fungi. The task was also analyzed for identifying healthy peanut kernels and those contaminated with non-aflatoxigenic fungi as shown in Table 7. In the first classification task (No. 1–6), MSAT exhibited exceptional performance with the highest test accuracy of 98.42±0.22%, test precision of 99.16±1.41%, and a test F1-score of 97.97±0.27%. Although its runtime was longer than those of some other models, it was reasonable at 26.93±1.03 s. Xception and SqueezeNet also showed excellent results, with SqueezeNet achieving the highest test recall of 97.06±1.53%. It also achieved a slightly faster runtime of 10.89±0.97 s. In the second classification task (No. 7–12), MSAT again demonstrated superior capability, achieving a test accuracy of 97.22±0.51%, a test precision of 99.76±0.50%, a test recall of 89.46±2.01%, and a test F1-score of 94.32 ±1.10%, with a slower runtime of 17.24±1.07 s compared with the first task. SqueezeNet and Xception remained strong competitors, presenting high test accuracy and test F1-scores. Notably, DenseNet-121, ResNet-50, and CapsuleNet were less effective in both tasks, and their performances significantly dropped in the second classification. DenseNet-121's test accuracy plunged to 70.28±14.27% with a very low test F1-score of 8.76±13.57%. Similarly, ResNet-50 and CapsuleNet showed substantial declines in test recall and test F1-score. In summary, the

Table 6
Models performance in classifying peanut kernels contaminated with diverse *Aspergillus flavus* fungi.

Algorithms	Test accuracy (%)	Test precision (%)	Test recall (%)	Test F1-score (%)	Runtime (s)
MSAT	95.00±0.78	95.02±0.77	95.00±0.78	95.01±0.78	21.38±0.93
DenseNet-121	86.80±10.80	87.39±10.50	85.69±12.12	86.49±11.29	74.98±2.57
ResNet-50	89.48±8.59	89.70±8.50	89.33±8.64	89.52±8.57	59.66±1.30
SqueezeNet	91.61±4.45	92.06±4.62	90.93±4.36	91.48±4.40	10.71±0.88
Xception	94.72±0.81	94.83±0.73	94.70±0.80	94.76±0.76	100.00±0.00
CapsuleNet	76.24±3.01	80.92±3.50	67.59±3.63	73.62±3.06	14.60±0.54

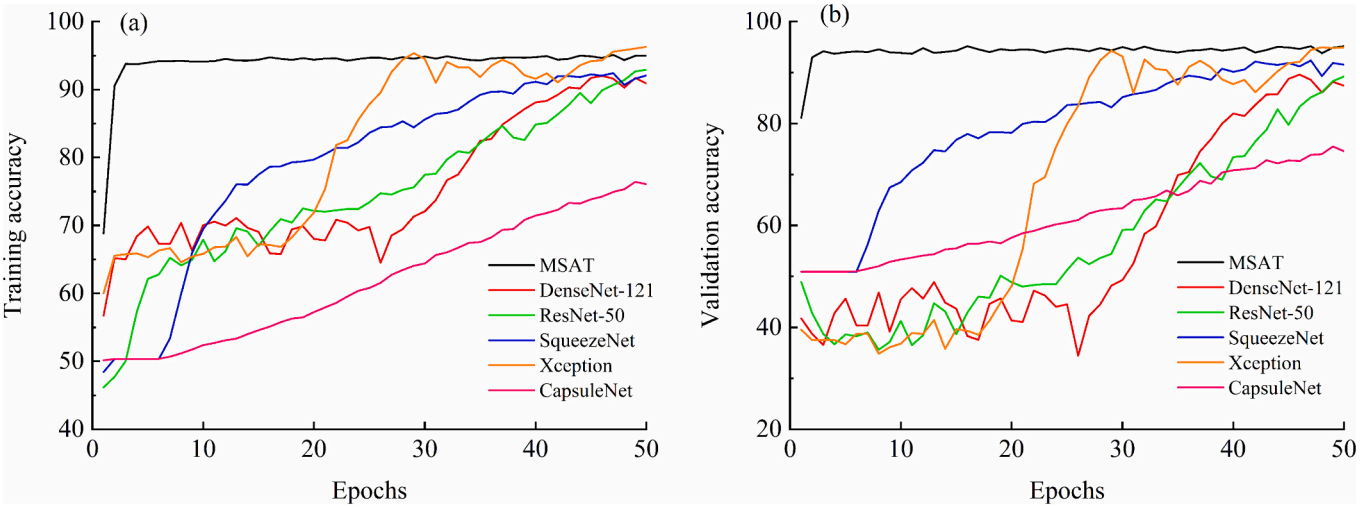

Fig. 7. Learning curves of MSAT and deep learning models, (a) training accuracy curves, (b) validation accuracy curves.

Table 7
Models performance in distinguishing healthy peanut kernels and those contaminated with aflatoxigenic fungi or non-aflatoxigenic fungi.

No.	Algorithms	Test accuracy (%)	Test precision (%)	Test recall (%)	Test F1-score (%)	Runtime (s)
1	MSAT	98.42±0.22	99.16±1.41	96.84±1.07	97.97±0.27	26.93±1.03
2	DenseNet-121	86.02±6.56	98.58±2.57	65.65±17.67	77.42±12.93	87.90±10.75
3	ResNet-50	79.04±8.57	82.48±13.16	66.55±24.74	69.31±17.02	59.61±1.47
4	SqueezeNet	98.04±0.44	98.00±1.77	97.06±1.53	97.50±0.56	10.89±0.97
5	Xception	98.13±0.46	97.88±2.40	97.46±1.58	97.63±0.53	100.00±0.00
6	CapsuleNet	91.07±1.44	89.76±3.57	87.46±2.67	88.52±1.72	14.91±0.65
7	MSAT	97.22±0.51	99.76±0.50	89.46±2.01	94.32±1.10	17.24±1.07
8	DenseNet-121	70.28±14.27	50.55±48.70	12.47±30.99	8.76±13.57	56.07±3.0.
9	ResNet-50	71.81±11.26	42.41±49.35	13.01±27.95	11.23±17.45	43.00±1.46
10	SqueezeNet	97.19±0.80	98.96±1.49	90.11±2.81	94.30±1.70	9.24±0.94
11	Xception	96.83±0.33	99.88±0.37	87.85±1.44	93.47±0.73	77.28±1.15
12	CapsuleNet	77.22±1.50	88.53±5.82	13.33±6.19	22.73±9.54	11.68±0.59

analysis revealed that MSAT outperformed the classical deep learning models in distinguishing healthy peanut kernels and those contaminated with aflatoxigenic fungi or non-aflatoxigenic fungi.

4.4.3. Models performance in identifying peanut kernels contaminated with aflatoxigenic fungi and those contaminated with non-aflatoxigenic fungi

Compared with the 3-class classification task, MSAT performed better in the 2-class classification tasks. This suggested that peanut kernels contaminated with aflatoxigenic fungi and those contaminated with non-aflatoxigenic fungi might be difficult to differentiate. Therefore, further analysis using MSAT and deep learning algorithms was recommended. Table 8 demonstrated that the performance of MSAT in distinguishing peanut kernels contaminated with aflatoxigenic fungi from those contaminated with non-aflatoxigenic fungi indeed declined, with a test accuracy of only 91.11±0.35% and a test recall further dropping to 82.44±1.93%. This suggested that it was more difficult for MSAT to identify peanut kernels contaminated by two different types of *Aspergillus flavus* fungi than it was to identify fungal-contaminated kernels from healthy peanut kernels. Among the other models, SqueezeNet and Xception still performed better, with their test accuracies being 90.28±0.98% and 87.33±4.97%, respectively. However, the performance of DenseNet-121, ResNet-50, and CapsuleNet remained relatively unsatisfactory.

4.4.4. Confusion matrix in different classification tasks

Fig. 8 showed confusion matrixes in differentiating healthy samples, aflatoxigenic fungal contaminated samples, and non-aflatoxigenic fungal contaminated samples, denoted as the control group, contaminated group 1 and contaminated group 2, respectively. Fig. 8(a) displayed that the MSAT exhibited minor errors in differentiating these 3-class samples. Specifically, the model occasionally misclassified the control group as the contaminated group 1 (12 samples). Additionally, samples in the contaminated group 1 were misclassified as the control group (1 sample) or as the contaminated group 2 (3 samples). The least misclassification occurred in the contaminated group 2, with 1 sample and 4 samples being mistaken for the control group and contaminated group 1, respectively. These errors could be attributed to the model's difficulty in discerning spatial-spectral feature differences between the

Table 8
Models performance in identifying peanut kernels contaminated with aflatoxigenic fungi and those contaminated with non-aflatoxigenic fungi.

Algorithms	Test accuracy (%)	Test precision (%)	Test recall (%)	Test F1-score (%)	Runtime (s)
MSAT	91.11±0.35	88.90±1.68	82.44±1.93	85.52±0.66	13.35±0.80
DenseNet-121	65.67±14.18	65.31±34.00	39.30±37.65	35.07±18.42	48.51±2.33
ResNet-50	66.76±8.16	61.51±21.11	56.16±33.20	48.41±18.23	38.24±1.37
SqueezeNet	90.28±0.98	89.16±3.57	81.37±3.72	84.97±1.55	7.78±0.89
Xception	87.33±4.97	85.56±7.74	73.95±17.09	77.75±12.72	57.60±1.22
CapsuleNet	79.81±5.47	89.95± 8.35	43.84±22.87	54.18±23.14	9.19±0.55

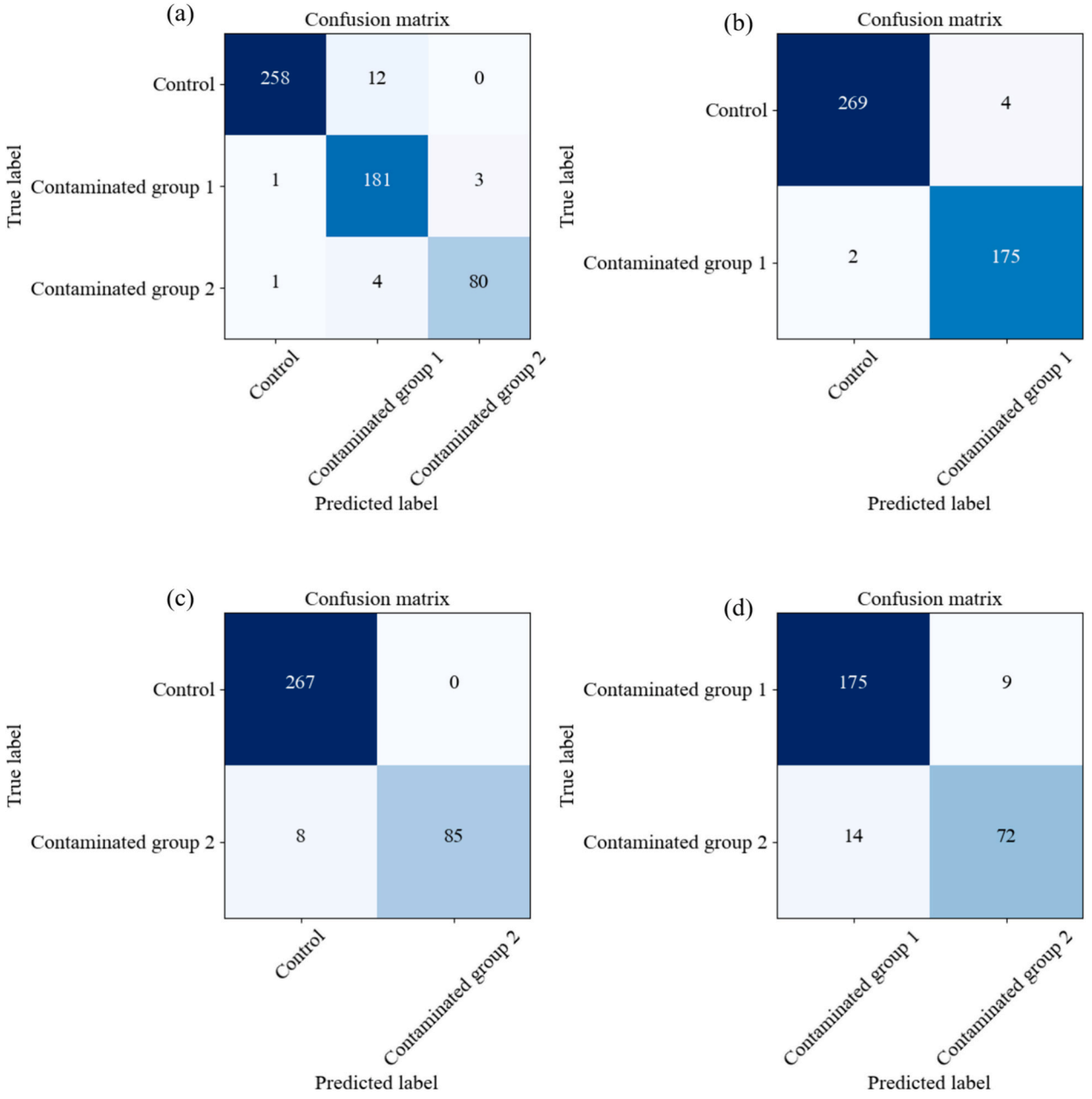

Fig. 8. confusion matrixes in differentiating healthy samples, aflatoxigenic fungal contaminated samples, and non-aflatoxigenic fungal contaminated samples, denoted as control group, contaminated group 1 and contaminated group 2, respectively, (a) result of classifying healthy samples, aflatoxigenic fungal contaminated samples, and non-aflatoxigenic fungal contaminated samples, (b) result of classifying healthy samples and aflatoxigenic fungal contaminated samples, (c) result of classifying healthy samples and non-aflatoxigenic fungal contaminated samples, (d) result of classifying aflatoxigenic fungal contaminated samples and non-aflatoxigenic fungal contaminated samples.

peanut kernel types. In the 2-class classifications of Fig. 8(b) and (c), the MSAT model showed an even higher level of test accuracy, with fewer misclassifications. In Fig. 8(b), only 4 samples of the control group were misclassified as the contaminated group 1, and 2 samples of the contaminated group 1 were misclassified as the control group. The model achieved perfect precision in identifying aflatoxigenic contaminated samples. Fig. 8(c) presented that the model misclassified 8 samples of the contaminated group 2 as the control group. In Fig. 8(d), the model differentiated aflatoxigenic fungal contaminations and non-aflatoxigenic fungal contaminations. 14 samples of the contaminated group 2 were misclassified as the contaminated group 1, and 9 samples of the contaminated group 1 were misclassified as the contaminated group 2. The results revealed that the MSAT model showed high accuracy and test precision in all tasks, although the samples of misclassification were existing. It highlighted the challenges in distinguishing subtly different types of contaminations in peanut kernels and the importance of further refining the model to enhance its discriminative capabilities.

4.5. *Attention weights visualization*

To analyze the role of the multi-scale attention mechanism in the 3-class classification task, a detailed study was conducted on each head in each layer of the multi-scale attention layers. The visualization of attention weights across different layers provided insights into how the model focused on various spatial-spectral features at different scales as

shown in Fig. 9. The first multi-head attention layer with 2 attention heads focused on capturing fundamental spatial-spectral features as illustrated in Fig. 9(a) and (b). These attention heads concentrated on broad spatial-spectral patterns, laying the groundwork for more nuanced differentiation in subsequent layers. Fig. 9(c–f) depicted the 4 attention heads of the second multi-head attention layer. A noticeable shift towards more specific spatial-spectral features was observed. These attention heads appeared to focus on more distinct spatial-spectral features, which was indicative of the presence or absence of aflatoxin. This layer's attention patterns were more refined than the first layer, honing in on particular spatial-spectral regions that differentiated healthy peanut kernels from those contaminated with diverse *Aspergillus flavus* fungi with greater specificity. The third layer's attention heads demonstrated the most specialized attention patterns as pictured in Fig. 9(g–n). Each attention head in this layer exhibited distinct focus areas, corresponding to high-level spatial-spectral features critical for accurate classification. This layer's nuanced attention distribution was indicative of its role in final decision-making, integrating insights from previous layers to arrive at a precise classification. Overall, the visualization and analysis of each attention head across all layers revealed a progressive refinement in the MSAT's focus. Starting from broad spatial-spectral features to more specialized signatures, each layer contributed uniquely to the MSAT's overall ability to accurately classify the peanut kernels contaminated with diverse *Aspergillus flavus* fungi. This multi-scale attention approach was instrumental in handling the complexity and high dimensionality of the data, demonstrating the potential in advanced classification tasks.

5. Conclusion

The MSAT model was proposed to accurately classify peanut kernels contaminated with diverse *Aspergillus flavus* fungi, which employed a sophisticated multi-scale attention mechanism to enhance its classification capabilities. This mechanism operated on the principle of processing data across various scales, enabling the model to capture a wide range of feature granularities. The MSAT model utilized several multi-head attention layers, which had varying heads and key dimensions, allowing the model to focus on both fine-scale and broad-scale features. In addition, the model also integrated a series of scale processing layers to capture features at different resolutions and incorporated a self-attention mechanism through an additional multi-head attention layer integrating information across different levels. The spectral analysis highlighted distinct spectral differences among healthy peanut kernels, aflatoxigenic fungi contaminated peanut kernels, and non-aflatoxigenic fungi-contaminated peanut kernels. These differences were linked to biochemical changes caused by the fungal infection. In optimizing the MSAT model through Bayesian optimization, the optimal hyperparameters were found to be a Dff of 256, a num_heads of 4, a num_layer of 1, and a dropout rate of 0.3. This combination achieved a test accuracy of 95.00% and high operational efficiency. Ablation experiments

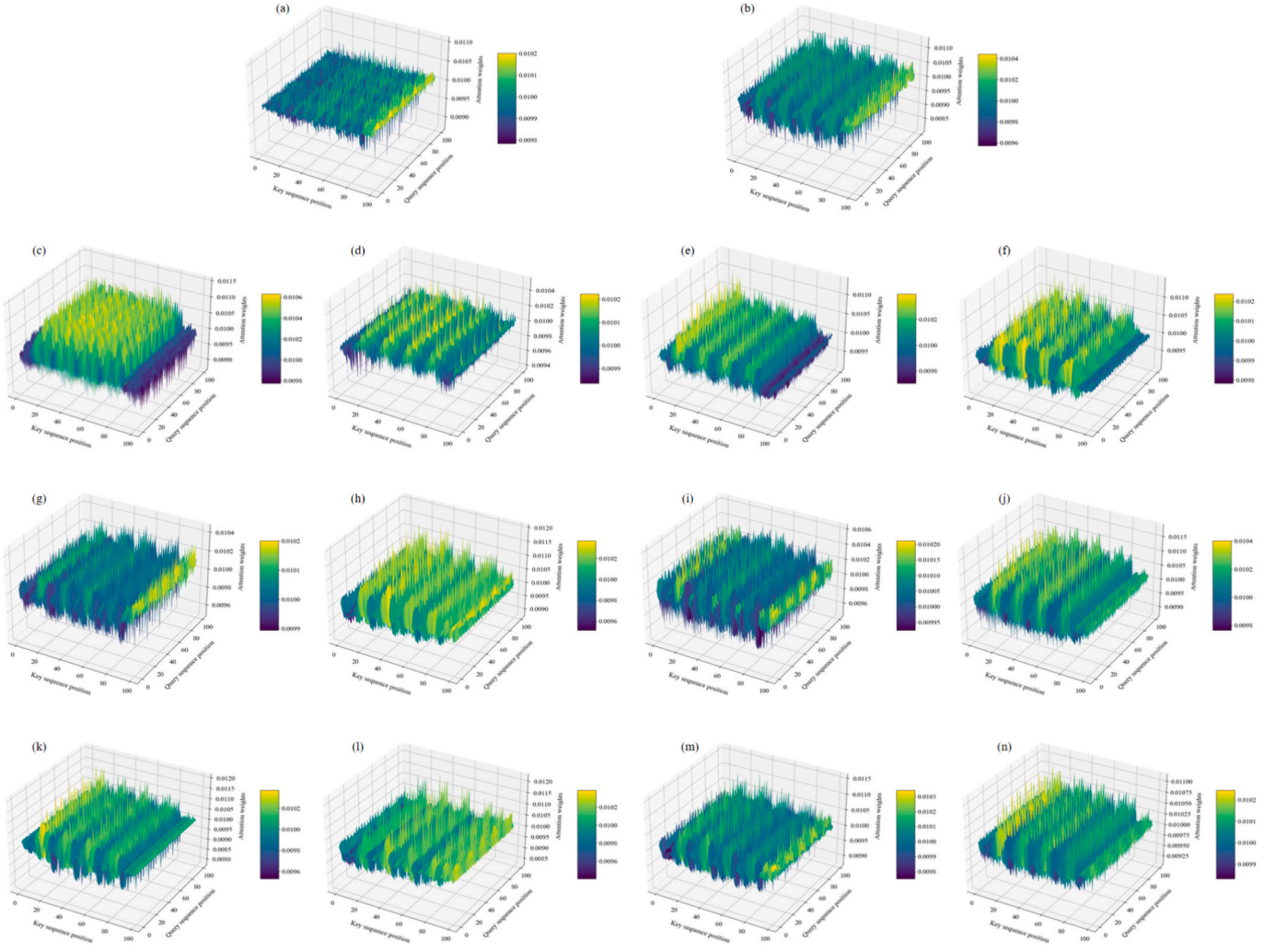

Fig. 9. Attention weights visualization, (a) layer 1 head 1, (b) layer 1 head 2, (c) layer 2 head 1, (d) layer 2 head 2, (e) layer 2 head 3, (f) layer 2 head 4, (g) layer 3 head 1, (h) layer 3 head 2, (i) layer 3 head 3, (j) layer 3 head 4, (k) layer 3 head 5, (l) layer 3 head 6, (m) layer 3 head 7, (n) layer 3 head 8.

demonstrated the crucial role of the model's components. Replacing or omitting certain layers led to varied decreases in performance. The intact model with all components achieved the highest performance metrics (test accuracy, test precision, test recall, and test F1-score all around 95.00%). The model's performance was also compared against other deep learning algorithms in various classification tasks. MSAT consistently outperformed classical models like DenseNet-121, ResNet-50, SqueezeNet, Xception, and CapsuleNet, particularly in distinguishing healthy kernels from those contaminated with aflatoxigenic fungi or non-aflatoxigenic fungi. In 2-class classification tasks, MSAT demonstrated even higher accuracy, though it faced challenges in differentiating aflatoxigenic contaminated peanut kernels and non-aflatoxigenic contaminated peanut kernels. Finally, the visualization of attention weights revealed that the MSAT model's multi-scale attention mechanism progressively refined its focus from broad spatial-spectral features to more specialized signatures. In summary, MSAT's multi-scale attention mechanism was key to its success, effectively combining multi-head attention layers at various scales with convolutional processing and self-attention. This sophisticated approach enabled the accurate classification of peanut kernels contaminated with diverse *Aspergillus flavus* fungi, showcasing its ability to detect fungal contamination in food and agriculture products.

Ethical approval

This article has no any study with human participants or animals by any of the authors.

CRediT authorship contribution statement

Zhen Guo: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Jing Zhang:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Haifang Wang:** Writing – original draft, Methodology, Conceptualization. **Haowei Dong:** Methodology, Investigation, Conceptualization. **Shiling Li:** Writing – original draft, Methodology, Conceptualization. **Xijun Shao:** Writing – original draft, Methodology, Investigation. **Jingcheng Huang:** Methodology, Investigation, Conceptualization. **Xiang Yin:** Resources, Methodology. **Qi Zhang:** Investigation, Conceptualization. **Yemin Guo:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Funding acquisition, Conceptualization. **Xia Sun:** Methodology, Investigation, Conceptualization. **Ibrahim Darwish:** Investigation, Funding acquisition.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 32372438, 31772068, 31872909), Funding Project for the Central Government to Guide the Development of Local Science and Technology (YDZX2022163), Natural Science Foundation of Shandong Province (ZR2023MC088), Shandong Province Major Applied Technology Innovation Project (SD2019NJ007), Technological Innovation Guidance Project of Science and Technology Department of Gansu Province (22CX8NA023) and Weifang Science and Technology Development Project (2021ZJ1103). The authors also extended their appreciation to the Researchers Supporting Project number (RSPD2024R944), King Saud University, Riyadh, Saudi Arabia, for funding this work.

References

Abbas, H.K., Weaver, M.A., Zablotowicz, R.M., Horn, B.W., Shier, W.T., 2005. Relationships between aflatoxin production and sclerotia formation among isolates of *Aspergillus* section *Flavi* from the Mississippi Delta. Eur. J. Plant Pathol. 112 (3), 283–287.

Achar, P.N., Hermetz, K., Rao, S., Apkarian, R., Taylor, J., 2009. Microscopic studies on the *Aspergillus flavus* infected kernels of commercial peanuts in Georgia. Ecotoxicol. Environ. Saf. 72 (8), 2115–2120.

Alaniz Zanon, M.S., Barros, G.G., Chulze, S.N., 2016. Non-aflatoxigenic *Aspergillus flavus* as potential biocontrol agents to reduce aflatoxin contamination in peanuts harvested in Northern Argentina. Int. J. Food Microbiol. 231, 63–68.

Alaniz Zanon, M.S., Clemente, M.P., Chulze, S.N., 2018. Characterization and competitive ability of non-aflatoxigenic *Aspergillus flavus* isolated from the maize agro-ecosystem in Argentina as potential aflatoxin biocontrol agents. Int. J. Food Microbiol. 277, 58–63.

Azad, R., Kazerouni, A., Heidari, M., Aghdam, E.K., Molaei, A., Jia, Y., Jose, A., Roy, R., Merhof, D., 2024. Advances in medical image analysis with vision Transformers: a comprehensive review. Med. Image Anal. 91, 103000.

Bertani, F.R., Mencattini, A., Gambacorta, L., de Ninno, A., Businaro, L., Solfrizzo, M., Gerardino, A., Martinelli, E., 2024. Aflatoxins detection in almonds via fluorescence imaging and deep neural network approach. J. Food Compos. Anal. 125, 105850.

Bilal, M., Xiaobo, Z., Arslan, M., Tahir, H.E., Azam, M., Junjun, Z., Basheer, S., Abdullah, 2020. Rapid determination of the chemical compositions of peanut seed (*Arachis hypogaea*) using portable near-infrared spectroscopy. Vib. Spectrosc. 110, 103138.

Campos, W.E.O., Rosas, L.B., Neto, A.P., Mello, R.A., Vasconcelos, A.A., 2017. Extended validation of a sensitive and robust method for simultaneous quantification of aflatoxins B_1, B_2, G_1 and G_2 in Brazil nuts by HPLC-FLD. J. Food Compos. Anal. 60, 90–96.

Chollet, F., 2017. Xception: deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807.

Gao, J., Zhao, L., Li, J., Deng, L., Ni, J., Han, Z., 2021. Aflatoxin rapid detection based on hyperspectral with 1D-convolution neural network in the pixel level. Food Chem. 360, 129968.

Guo, Z., Zhang, J., Dong, H., Sun, J., Huang, J., Li, S., Ma, C., Guo, Y., Sun, X., 2023a. Spatio-temporal distribution patterns and quantitative detection of aflatoxin B_1 and total aflatoxin in peanut kernels explored by short-wave infrared hyperspectral imaging. Food Chem. 424, 136441.

Guo, Z., Zhang, J., Ma, C., Yin, X., Guo, Y., Sun, X., Jin, C., 2023b. Application of visible-near-infrared hyperspectral imaging technology coupled with wavelength selection algorithm for rapid determination of moisture content of soybean seeds. J. Food Compos. Anal. 116, 105048.

Han, Z., Gao, J., 2019. Pixel-level aflatoxin detecting based on deep learning and hyperspectral imaging. Comput. Electron. Agric. 164, 104888.

He, K., Zhang, X., Ren, S., et al., 2016. Deep residual learning for image recognition. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016, pp. 770–778.

Huang, G., Liu, Z., van der Maaten, L., et al., 2017. Densely connected convolutional networks. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. CVPR 2017, pp. 4700–4708.

Hulikunte Mallikarjunaiah, N., Jayapala, N., Puttaswamy, H., Siddapura Ramachandrappa, N., 2017. Characterization of non-aflatoxigenic strains of *Aspergillus flavus* as potential biocontrol agent for the management of aflatoxin contamination in groundnut. Microb. Pathog. 102, 21–28.

Iandola, F.N., Moskewicz, M.W., Ashraf, K., et al., 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size. CoRR (abs/1602.07360).

Jiang, J., Qiao, X., He, R., 2016. Use of near-infrared hyperspectral images to identify moldy peanuts. J. Food Eng. 169, 284–290.

Kimuli, D., Wang, W., Wang, W., Jiang, H., Zhao, X., Chu, X., 2018. Application of SWIR hyperspectral imaging and chemometrics for identification of aflatoxin B_1 contaminated maize kernels. Infrared Phys. Technol. 89, 351–362.

Liu, Z., Jiang, J., Qiao, X., Qi, X., Pan, Y., Pan, X., 2020. Using convolution neural network and hyperspectral image to identify moldy peanut kernels. LWT 132, 109815.

Long, Y., Huang, W., Wang, Q., Fan, S., Tian, X., 2022. Integration of textural and spectral features of Raman hyperspectral imaging for quantitative determination of a single maize kernel mildew coupled with chemometrics. Food Chem. 372, 131246.

Lu, Y., Jia, B., Yoon, S.C., Zhuang, H., Ni, X., Guo, B., Gold, S.E., Fountain, J.C., Glenn, A. E., Lawrence, K.C., Zhang, H., Guo, X., Zhang, F., Wang, W., 2022. Spatio-temporal patterns of *Aspergillus flavus* infection and aflatoxin B_1 biosynthesis on maize kernels probed by SWIR hyperspectral imaging and synchrotron FTIR microspectroscopy. Food Chem. 382, 132340.

Makarichian, A., Chayjan, R.A., Ahmadi, E., Zafari, D., 2022. Early detection and classification of fungal infection in garlic (*A. sativum*) using electronic nose. Comput. Electron. Agric. 192, 106575.

Mishra, G., Panda, B.K., Ramirez, W.A., Jung, H., Singh, C.B., Lee, S.H., Lee, I., 2022. Application of SWIR hyperspectral imaging coupled with chemometrics for rapid and non-destructive prediction of Aflatoxin B_1 in single kernel almonds. LWT 155, 112954.

Pv, A., Buddhiraju, K.M., Porwal, A., 2019. Capsulenet-based spatial-spectral classifier for hyperspectral images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12 (6), 1849–1865.
Ren, M., Dong, Y., Wang, J., Lin, J., Qu, L., Zhou, Y., Chen, Y., 2023. Computer vision-assisted smartphone microscope imaging digital immunosensor based on click chemistry-mediated microsphere counting technology for the detection of aflatoxin B_1 in peanuts. Anal. Chim. Acta 1278, 341687.
Romero-Sánchez, I., Gracia-Lor, E., Madrid-Albarrán, Y., 2024. Aflatoxin detoxification by thermal cooking treatment and evaluation of in vitro bioaccessibility from white and brown rice. Food Chem. 436, 137738.
Salano, E.N., Mulwa, R.M., Obonyo, M.A., 2024. Peanut (*Arachis hypogea*) accessions differentially accumulate aflatoxins upon challenge by *Aspergillus flavus*: implications for aflatoxin mitigation. J. Agric. Food Res. 15, 100923.
Sundaram, J., Kandala, C.V., Govindarajan, K.N., Subbiah, J., 2012. Sensing of moisture content of in-shell peanuts by NIR reflectance spectroscopy. J. Sensor Technol. 02 (01), 17910.
Tao, F., Yao, H., Hruska, Z., Kincaid, R., Rajasekaran, K., Bhatnagar, D., 2020. A novel hyperspectral-based approach for identification of maize kernels infected with diverse *Aspergillus flavus* fungi. Biosyst. Eng. 200, 415–430.
Tao, F., Yao, H., Hruska, Z., Kincaid, R., Rajasekaran, K., 2022. Near-infrared hyperspectral imaging for evaluation of aflatoxin contamination in corn kernels. Biosyst. Eng. 221, 181–194.
Thati, R., Seetha, B.S., Alegete, P., Mudiam, M.K.R., 2024. Molecularly imprinted dispersive micro solid-phase extraction coupled with high-performance liquid chromatography for the determination of four aflatoxins in various foods. Food Chem. 433, 137342.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Proces. Syst. (2017-Dec., no pagination).
Wang, Y., Cheng, J., 2018. Rapid and non-destructive prediction of protein content in peanut varieties using near-infrared hyperspectral imaging method. Grain Oil Sci. Technol. 1 (1), 40–43.
Wang, X., Yang, W., Yang, Y., Huang, M., Zhu, Q., 2023. Identification of tomato bacterial wilt severity based on hyperspectral imaging technology and spectrum Transformer network. Ecol. Inform. 78, 102353.
Weaver, M.A., Abbas, H.K., 2019. Field displacement of Aflatoxigenic *Aspergillus flavus* strains through repeated biological control applications. Front. Microbiol. 10, 1.
Weng, S., Han, K., Chu, Z., Zhu, G., Liu, C., Zhu, Z., Zhang, Z., Zheng, L., Huang, L., 2021. Reflectance images of effective wavelengths from hyperspectral imaging for identification of *Fusarium* head blight-infected wheat kernels combined with a residual attention convolution neural network. Comput. Electron. Agric. 190, 106483.
Williams, P.J., Geladi, P., Britz, T.J., Manley, M., 2012. Investigation of fungal development in maize kernels using NIR hyperspectral imaging and multivariate data analysis. J. Cereal Sci. 55 (3), 272–278.
Wu, N., Jiang, H., Bao, Y., Zhang, C., Zhang, J., Song, W., Zhao, Y., Mi, C., He, Y., Liu, F., 2020. Practicability investigation of using near-infrared hyperspectral imaging to detect rice kernels infected with rice false smut in different conditions. Sensors Actuators B Chem. 308, 127696.
Xu, Z., Hu, H., Wang, T., Zhao, Y., Zhou, C., Xu, H., Mao, X., 2023. Identification of growth years of Kudzu root by hyperspectral imaging combined with spectral-spatial feature tokenization transformer. Comput. Electron. Agric. 214, 108332.
Yao, Y., Gao, S., Ding, X., Zhang, Q., Li, P., 2021. Topography effect on *Aspergillus flavus* occurrence and aflatoxin B_1 contamination associated with peanut. Curr. Res. Microb. Sci. 2, 100021.
Yin, Y., Lou, T., Yan, L., Michailides, T.J., Ma, Z., 2009. Molecular characterization of toxigenic and atoxigenic Aspergillus flavus isolates, collected from peanut fields in China. J. Appl. Microbiol. 107 (6), 1857–1865.
Yuan, D., Jiang, J., Qi, X., Xie, Z., Zhang, G., 2020. Selecting key wavelengths of hyperspectral imagine for nondestructive classification of moldy peanuts using ensemble classifier. Infrared Phys. Technol. 111, 103518.
Zheng, J., Xia, A., Shao, L., Wan, T., Qin, Z., 2019. Stock volatility prediction based on self-attention networks with social information. In: CIFEr 2019-IEEE Conf. Comput. Intell. Financ. Eng. Econ (no pagination).