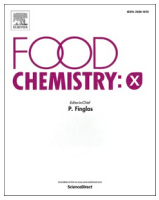
Food Chemistry: X 23 (2024) 101759

Contents lists available at ScienceDirect

Food Chemistry: X

journal homepage: www.sciencedirect.com/journal/food-chemistry-x



Accurate and visualiable discrimination of Chenpi age using 2D-CNN and Grad-CAM++ based on infrared spectral images

Li Jun Tang [1], Xin Kang Li [1], Yue Huang, Xiang-Zhi Zhang, Bao Qiong Li [*]

School of Pharmacy and Food Engineering, Wuyi University, Jiangmen, 529020, PR China

ARTICLE INFO

Keywords:
FTIR spectral image
Chenpi
2D-CNN
Grad-CAM++
Feature visualization

ABSTRACT

Dried tangerine peel ("Chenpi"), has numerous clinical and nutritional benefits, with its quality being significantly influenced by its storage age, referred to as "Chen Jiu Zhe Liang" in Chinese. Concequently, the rapid and accurate identification of Chenpi's age is important for consumers. In this study, Fourier transform infrared spectroscopy (FTIR) was employed to capture spectral images of Chenpi. These FTIR images were then analyzed using a two-dimensional convolutional neural networks (2D-CNN) model, achieving a discrimination accuracy of 97.92%. To address the "black box" nature of the 2D-CNN, Gradient-weighted Class Activation Mapping Plus Plus (Grad-CAM++) was utilized to highlight the important regions contributing to the model's performance. Additionally, six other machine learning models were developped using features identified by the 2D-CNN to validate their effectiveness. The results demonstrated that the combination of FTIR spectral images and 2D-CNN provides a highly effective method for accurately determining the age of Chenpi.

1. Introduction

Chenpi, also known as dried tangerine peel, is created by sun-drying the peel of tangerines, which is known for its health benefits, including regulating qi, aiding digestion and reducing phlegm (Wang et al., 2023). Chenpi is highly valued for its medicinal properties, with the age of Chenpi being a key factor in determining its quality, which is known as "Chen Jiu Zhe Liang" in Chinese (Sun et al., 2023). The traditional methods for identifying the age of Chenpi involve sensory and physical examinations, such as surface observation, aroma identification, and taste testing. However, these methods require a high level of skill and experience and lack quantitative data, making it difficult to standardize evaluations and compare results across different batches. In contrast, chemical methods provide a more precise and objective approach to identifying the age of Chenpi by analyzing its chemical composition and the changes that occur over time. The commonly used techniques including high-performance liquid chromatography (HPLC) (Li et al., 2019), gas chromatography–mass spectrometry (GC–MS) (Shi et al., 2024), and liquid chromatography-mass spectrometry (LC-MS) (Yang et al., 2022). These methods demonstrate high sensitivity and resolution in detecting chemical components such as flavonoids, phenolic acids, polysaccharides, and volatile oils. When combined with data processing methods like artificial neural network (ANN), principal component analysis (PCA) and partial least squares-discriminant analysis (PLS-DA), the accuracy and efficiency of these detection techniques are enhanced. For instance, GC–MS combined with ANN allows for precise analysis of volatile compounds (Qu et al., 2015). Similarly, PCA and PLS-DA can be employed alongside HPLC or LC-MS for the differentiation of Chenpi from different years based on the analysis of flavonoids, and phenolic acids (Liang et al., 2022; Wang et al., 2016). However, despite the improved capabilities offered by these combined approaches, there are certain limitations. High-precision analytical techniques typically require complex and expensive equipment, as well as specialized skills, which limits their widespread application. Additionally, these methods

Abbreviations: Chenpi, dried tangerine peel; FTIR, Fourier transform infrared spectroscopy; HPLC, high-performance liquid chromatography; GC–MS, gas chromatography–mass spectrometry; LC-MS, liquid chromatography-mass spectroscopy; ANN, artificial neural network; PCA, principal component analysis; 2D-CNN, two-dimensional convolutional neural networks; Grad-CAM++, Gradient-weighted Class Activation Mapping Plus Plus; GNN, graph neural networks; LSTM, long short-term memory networks; GAN, generative adversarial networks; RNN, recurrent neural networks; DRL, transformer models and deep reinforcement learning; DA, data augmentation; AdaBoost, adaptive boosting; GBDT, gradient boosting decision tree; PLS-DA, partial least squares-discriminant analysis; LR, logistic regression; KNN, K-nearest neighbors; DT, decision tree; XGBoost, extreme gradient boosting.

* Corresponding author at: Dongcheng Village, Jiangmen, Guangdong 529020, PR China.
E-mail address: libq201406@163.com (B.Q. Li).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.fochx.2024.101759
Received 30 July 2024; Received in revised form 19 August 2024; Accepted 20 August 2024
Available online 22 August 2024

still rely on traditional machine learning algorithms, which may fall short when handling complex, non-linear data. These liminations underscore the need for simpler, innovative, and accurate methods for identifying the age of Chenpi.

Fourier transform infrared spectroscopy (FTIR) is a powerful spectroscopy technique that offers the advantages of non-destructive analysis, comprehensive chemical information, high sensitivity, and rapid data acquisition. FTIR covers a broad spectral range, enabling the detection of a wide variety of functional groups and molecular vibrations, thus providing detailed chemical information. For example, Pan et al. (Pan et al., 2022) and Zhang et al. (Zhang et al., 2022) successfully identified the geographical origin and storage age of Chenpi using near-infrared spectroscopy combined with machine learning methods. With the increasing computational power, deep learning techniques such as convolutional neural networks (CNN), graph neural networks (GNN), long short-term memory network (LSTM), autoencoder, generative adversarial networks (GAN), recurrent neural networks (RNN), transformer models, and deep reinforcement learning (DRL), have been developed and made remarkable advancements in a wide range of research fields (Dou et al., 2023). Li et al. (Qin et al., 2024) applied FTIR combined with CNN-LSTM model to classify Chenpi samples of different ages, achieving a discriminant accuracy higher than 96.5%. However, while these methods have enabled rapid, non-destructive and accurate discrimination of Chenpi with different ages, they face challenges with model interpretability, making it difficult to understand how specific input features influence the model's predictions. Therefore, developing rapid, efficient, accurate, and interpretable methods for identifying the age of Chenpi is crucial. Such advancements would not only enhance the reliability of the models but also provide deeper insights and support for scientific research and practical applications.

In recent years, the need for reliable methods to authenticate and evaluate the quality of Chenpi has driven the exploration of advanced analytical techniques. Li et al. (Li et al., 2024) highlighted in their review that infrared spectroscopy has emerged as an important technique for identifying Chenpi authenticity, offering more distinctive identification characteristics compared to other physical and chemical methods. Among the above mentioned deep learning methods, CNN demonstrated excellent performance in image classification (Liu et al., 2022), due to their ability to capture local and spatial hierarchies of features within images. By integrating the detailed chemical information provided by FTIR with the powerful image classification capabilities of CNN, this study aims to develop an accurate and novel model for determining the age of Chenpi based on FTIR spectral images. To enhance model interpretability, we employ Gradient-weighted Class Activation Mapping Plus Plus (Grad-CAM++) (Chattopadhay et al., 2018), an advanced visualization technique that can highlight critical regions in the spectral images, providing insights into the CNN's decision-making process. Although the combination of CNN and Grad-CAM++ has been used to identify the geographical origins of traditional Chinese medicine samples based on hyperspectral imaging (Cai et al., 2023), our study is the first to apply this approach using FTIR spectral images to determine the age of Chenpi. Additionally, we have conducted a more in-depth analysis to emphasize both the interpretability and accuracy of the model.

The objectives of this study are: (1) to develop a two-dimensional convolutional neural networks (2D-CNN) model capable of accurately classifying the age of Chenpi using FTIR spectral images, (2) to utilize Grad-CAM++ for visualizing the focus areas of the 2D-CNN model during age determination, thereby avoiding a “black box” phenomenon and enhancing the transparency of the model's predictions, and (3) to validate the effectiveness of features identified by the 2D-CNN by establishing six different machine learning models using these highlighted features. By addressing these objectives, this study aims to provide a reliable and interpretable solution for Chenpi age identification, benefiting the food industry in quality control and increasing consumer confidence in Chenpi products.

2. Sample and data description

In this study, 39 Chenpi samples were purchased from ten different companies located in Jiangmen, Guangdong Province, China, specifically sourced from the Xinhui region. These samples were harvested in the years 2014, 2016, 2018 and 2020. Detailed information of the Chenpi samples is summarized in Table S1 (Supporting information). Notably, the production of Xinhui Chenpi follows the traditional process of sun-drying fresh citrus peels and then storing them in a ventilated and dry environment to mature. In order to minimise the influence of unknown external factors on the analysis results, all samples were subjected to experimental analysis as soon as possible after purchase.

FTIR data were collected from the 39 Chenpi samples in the mid-infrared spectral range of 400–4000 cm^{-1} (FTIR, Nicolet iS5, Thermo Fisher Scientific Inc., USA). The samples were pre-treated, milled and removed by a 60 mesh sieve. The FTIR spectrum of the Chenpi powder was measured by potassium bromide (KBr, purchased from Shanghai Macklin Biochemical Co., Ltd., Shanghai, China) pellet method. The samples were weighted 1 mg and mixed with 100 mg of KBr uniformly, pressed it into a transparent flake for measurement. The accompanying OMNIC software was used for infrared spectroscopy analysis.

The objective of this study was to accurate classify and analyze these samples based on their harvest year. Each sample was measured six times, generating a total of 234 spectral data from the original 39 samples. The data were organized into a data matrix of size 234×7468.

3. Methodology

3.1. Deep learning

3.1.1. Two-dimensional convolutional neural networks

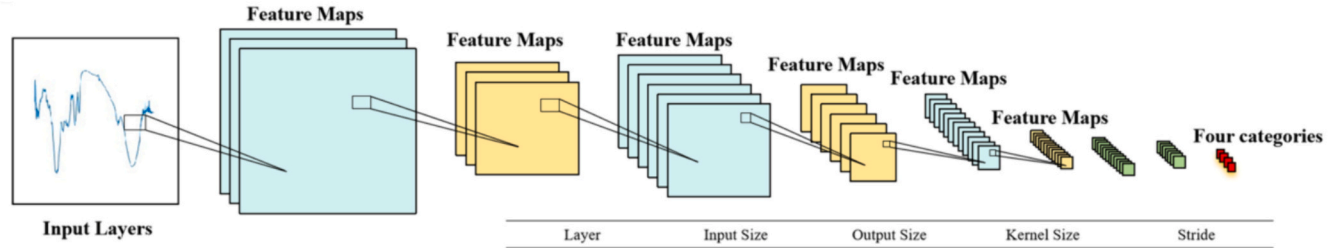
FTIR spectral classification using 2D-CNN involves treating spectral data as two-dimensional images, extracting spatial and spectral features through convolution and pooling, and progressively learning abstract features to achieve accurate classification. This approach has high potential in infrared spectral tasks due to its ability to handle complex features, improve classification accuracy, and reduce overfitting. By carefully designing the network structure and appropriately adjusting the parameters, these advantages can be fully utilized to enhance spectral analysis.

This paper introduces a 2D-CNN model specially designed for FTIR spectral image processing tasks. The model employs three successive convolutional layers to extract hierarchical features from the input images. Each convolutional layer consists of convolution operations followed by batch normalization, ReLU activation, and maximum pooling. This design not only enhances feature extraction capabilities but also reduces the model's parameter count, thereby improving computational efficiency. Starting with a three-channel input image, the initial convolutional layer progressively transforms it into feature maps with 16, 32, and 64 channels, respectively. These feature maps are further compressed through fully connected layers, enabling the mapping from high-dimensional data to the final classification predictions. The output layer utilizes the log softmax function to generate probabilities for each category, making the model suitable for multi-category classification tasks.

The structured 2D-CNN architecture exemplifies the efficacy of deep learning in image recognition and classification. By stacking convolutional layers, the model extracts intricate details from input images, enabling precise classification. The architectural diagram of the 2D-CNN structure was illustrated in Fig. 1.

3.1.2. Visualization method

The 2D-CNN is often considered as a “black box” model in deep learning. To gain deeper insights into the decision-making process of the 2D-CNN model and validate its interpretability, we employed the Grad-CAM++ method for model visualization in this study. Grad-CAM++ is



Layer	Input Size	Output Size	Kernel Size	Stride
Input	256×256×3	-	-	-
Conv1	3×32×32	16×15×15	3×3	2
MaxPool1	16×15×15	16×7×7	2×2	2
Conv2	16×7×7	32×3×3	3×3	2
MaxPool2	32×3×3	32×1×1	2×2	2
Conv3	32×1×1	64×1×1	3×3	2
MaxPool3	64×1×1	64×1×1	2×2	2
FC1	64×3×3	64	-	-
FC2	64	10	-	-
Output	10	4	-	-

Fig. 1. The architectural diagram of the 2D-CNN structure.

an advanced version of the Grad-CAM technique, which generates class activation maps by leveraging gradient information to highlight the image regions the model focuses on when recognizing target categories (Moujahid et al., 2022). In the present study, we first propagated the input images through the trained 2D-CNN model to obtain its predictions. Subsequently, we used the Grad-CAM++ method to compute the gradients of the output layer with respect to the convolutional feature maps, generating class activation maps based on these gradients. Finally, these maps were overlaid onto the original images to create intuitive heatmaps, revealing the critical regions the model relies on during classification decisions.

By utilizing Grad-CAM++ for visualization, we not only validated the model's performance on specific tasks but also uncovered its internal mechanisms, providing a clearer understanding of the model's behavior. This visualization method provides strong support for interpreting deep learning models, thereby enhancing their transparency and credibility.

3.1.3. Data augmentation

Deep learning models relys on large datasets for effective training, and data augmentation (DA) techniques can be used to enhance model performance by increasing the diversity of instances through synthesized data (Shorten & Khoshgoftaar, 2019). In addition, DA helps to prevent overfitting problems during the training phase of the model and maintain the model's generalization ability, making it an effective method for improving the overall performance of deep learning models (Hao et al., 2023). When performed in image classification assignment, DA methods commonly include flipping (flipping the image horizontally or vertically), rotation, scaling and cropping, panning, contrast and brightness adjustments, noise addition, Gaussian blurring, etc. (Khalifa et al., 2022).

3.2. Machine learning methods

3.2.1. Variable selection methods

When optimizing model performance through variable selection, choosing the appropriate algorithms is crucial. In this study, we employ AdaBoost (adaptive boosting) and Boruta as the primary methods for variable selection. These methods address the challenges of high-dimensional data and complex relationships, enhancing prediction accuracy and ensuring that the selected variables significantly contribute to the interpretability and practicality of the research.

AdaBoost is a powerful ensemble learning algorithm that works by iteratively training a multiple weak classifiers, typically simple models like decision trees and then combining them into a strong predictive model, resulting in a highly accurate ensemble model (Tang et al., 2021). In variable selection, AdaBoost effectively identifies features that have significantly impact on predicting the target variable. Unlike traditional methods such as stepwise regression or simple filtering methods, AdaBoost excels at capturing complex interactions between features, which are often missed by simpler methods. This capability leads to enhanced prediction accuracy and better generalization performance of the model. In this study, AdaBoost was chosen for variable selection because of its ability to handle high-dimensional data and complex feature structures, ensuring that the selected variables provide maximal information and predictive power.

Boruta algorithm is a robust feature selection method based on random forests classifer, specifically designed to assess the importance of variables in predicting the target variable (Kursa & Rudnicki, 2010). Unlike standard random forests, which directly rank features based on their importance, Boruta introduces a more rigorous comparison by adding randomly generated “shadow features” to the dataset (Yu et al., 2024). These shadow features are duplicates of the original features with their values randomly shuffled, serving as a baseline to test the importance of each real feature. This method helps avoid overfitting and bias in feature selection, making Boruta particularly suitable for complex datasets and high-dimensional scenes, such as infrared spectral data. In this study, Boruta was chosen for its ability to rigorously evaluate the significance of each variable, ensuring that only the most relevant features are selected, thereby improving the model's interpretability and reducing the risk of including irrelevant or redundant information.

3.3. Classification methods

Classification in FTIR spectroscopy involves categorizing spectra into distinct groups based on their unique patterns or features. Typically, machine learning models are trained on a dataset of known FTIR spectra, enabling them to classify new, previously unseen spectra accurately. In this study, we employed six widely-used machine learning algorithms for FTIR data analysis: gradient boosting decision tree (GBDT) (Wu et al., 2021), partial least squares-discriminant analysis (PLS-DA) (Naeim Mohamad Asri et al., 2022), logistic regression (LR) (Akturk et al., 2024), K-nearest neighbors (KNN) (Cunningham & Delany, 2021), decision tree (DT) (Chen et al., 2020), and extreme gradient boosting (XGBoost) (Sheridan et al., 2016).

GBDT is an ensemble machine learning algorithm that combines multiple decision trees as base learners to improve model performance. The algorithm iteratively builds decision trees by focusing on the residuals (errors) of the previous trees, allowing the model to gradually correct its mistakes. In each iteration, a new decision tree is trained using the residuals from the previous iteration, optimizing the model by

minimizing the loss function along the negative gradient direction (Liang et al., 2020). This process continues until the model achieves the desired accuracy or a pre-set number of iterations is eached. In this study, the parameters for the bosted tree were fine-tuned using a grid search method, and the model's performance was cross-validated with 4 folds. The optimal parameters were set as follows: "*n_estimators*" to 300, "*learning_rate*" to 0.01, and "*max_depth*" to 3, ensuring a balance between model complexity and prediction accuracy.

PLS-DA is a statistical analysis method that combines PLS regression and linear discriminant analysis to classify data sets into different categories. The method is particularly useful for high-dimensional data, where it reduces the number of latent variables by finding the most relevant features that contribute to the prediction of the target variable (Pokhrel et al., 2023). PLS-DA simultaneously models the relationships between predictor and response variables, allowing for effective classification even in the presence of multicollinearity among the predictors. In this study, five-fold cross-validation was employed to determine the optimal hyperparameters, with the best performance achieved when the number of components ("*n_components*") was set to 10.

LR is a classic machine learning algorithm widely used for binary classification problems. The algorithm fits a decision boundary, which can be linear or polynomial, to separate the data into different classes (Jin et al., 2022). It then calculates the probability of each data point belong to a particular class based on this boundary, using a logistic function as the predictor function. One of the key advantages of LR is its ability to avoid making inaccurate assumptions about the data distributions, making it robust in various situations. In this study, L2 regularization was applied to LR model to prevent overfitting, with the regularization parameter "*C*" set to 100, which controls the trade-off between maximizing the likelihood and minimizing the magnitude of the coefficients.

KNN is a popular parameter-free supervised learning algorithm that is widely used in classification problems. The algorithm classifies a data point by analyzing the "K" closest data points (neighbors) in training set and assigning the most common class among them to the new data point (Ali et al., 2019). The distances between data points is typically measured using Euclidean distance, although other metrics can also be used. In this study, we utilized 5-fold cross-validation to determine the optimal number of neighbors ("*k*"). The accuracy was evaluated for "*k*" value ranging from 1 to 20, and the optimal "*k*" value was identified as 5, providing the best balance between bias and variance.

DT is a popular supervised learning method used for both classification and regression problems. The algorithm works by recursively splitting the dataset into smaller subsets based on the value of the most informative feature at each step. It selects a feature as the current node and calculates its information gain or entropy. The feature with the highest information gain or entropy is chosen for the split. This process repeats until the stopping condition is met (Kim, 2016). In this study, grid search was employed to tune the hyperparameters of the decision tree model, and each parameter configuration was evaluated using 4-fold cross-validation. The optimal settings were determined to be a maximum depth of 3, a minimum number of samples required to split an internal node of 2, and a minimum number of samples required to be at a leaf node of 4.

XGBoost is an advanced implementation of gradient boosting that has gained widespread recognition for its high performance in both classification and regression tasks. It enhances model accuracy through the integration of multiple weak classifiers (typically decision trees) into a single strong learner (Qiu et al., 2022). XGBoost introduces additional regularization terms and penalty functions in its objective function to control model complexity and assess overfitting risks, making it more robust than standard gradient boosting. In this study, the model's parameters were carefully tuned to optimize performance. The final settings were "*n_estimators*" at 400, "*learning_rate*" at 0.2, and "*max_depth*" at 2, allowing the model to achieve high predictive accuracy while maintaining a controlled complexity.

3.4. Model evaluation

In all models, the entire dataset was randomly divided into training, validation and test sets in a 6:2:2 ratio. Model hyperparameters were initially adjusted using the training set. The accuracy or loss on the validation set was then used to further tune the hyperparameters, helping to accelerate the model's convergence. Finally, the test set was used to evaluate the model's performance.

Evaluating the performance of a classifier model is crucial. Common evaluation metrics include classification accuracy, precision, recall, and F1_score (Adegun et al., 2023). The metrics were definedas follows:

Accuracy measures the proportion of correct predictions out of the total number of cases examined, providing an overall evaluation of the model's classification performance. It is calculated as:

$$Accracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where, *TP* (True Positives) is the number of samples correctly predicted as positive, *FP* (False Positives) is the number of negative samples incorrectly predicted as positive, *TN* (True Negatives) is the number of samples correctly predicted as negative, and *FN* (False Negatives) is the number of positive samples incorrectly predicted as negative.

Precision measures the proportion of true positive cases out of all predicted positive cases, helping us understand the reliability of the model in correctly identifying samples belonging to a particular class. It is calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall, also known as sensitivity, measures the proportion of actual positive cases that are correctly identified by the model. It is calculated as:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1_score is the weighted average of Precision and Recall, providing a balanced assessment of the model's performance, particularly when dealing with inbalanced datasets. It is calculated as:

$$F1\_score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

Confusion matrix can provide a detailed analysis of the model's performance by reporting the counts of *FP*, *FN*, *TP* and *TN*, allowing for straightforward comparsion between different models on the same test.

Spectral/image data processing and classification model development were carried out using Python (version 3.7.0, 64-bit) with PyCharm Professional (version 2021.1.4) on a Windows 10 platform. The machine learning algorithms were implemented using scikit-learn (version 1.0.1). The 2D-CNN model was developed using PyTorch (version 1.10.0). All data analysis procedures were executed on a computer equipped with an Intel(R) Core(TM) i9-9900K processor clocked at 3.6 GHz and an NVIDIA GeForce GTX 1660 graphics card.

4. Result and discussion

4.1. Spectral profile of samples

The mid-infrared spectra of four Chenpi samples, each with different storage year, was illustrated in Fig. S1 (see supporting information). The results reveal a significant overlap in the spectral features among Chenpi samples from different years, which propose a challenge for differentiating the sample classes using conventional methods such as cosine similarity or Euclidean distance. To overcome this limitation and improve the classification accuracy, we constructed a 2D-CNN specifically designed for spectral image classification. This model's performance was then compared with traditional machine learning methods

that utilize spectral data. The analysis and validation process for the 2D-CNN model is illustrated in Fig. 2.

4.2. Deep learning classification results

4.2.1. Classification results of 2D-CNN

Plotting the loss value and accuracy curves of datasets is an essential method for evaluating the learning progress and detecting overfitting or underfitting in a deep learning model (Zhao et al., 2021). The loss curves indicate a trend of convergence and reduction, suggesting that the learning process is been optimized. Meanwhile, the accuracy curves show the model's performance improvements on both the training and validation sets, highlighting its ability to learn and generalise. As shown in Fig. 3a, the training and validation loss and accuracy for the 2D-CNN model demonstrate a clear reduction in training loss and a corresponding increase in accuracy over the epochs, reflecting the model's learning efficiency (Dong, 2024), which is a positive outcome. After evaluating the relationship between the number of iterations and the model's runtime, we identified the optimal parameter settings for the 2D-CNN model as follows: "*epoch*" at 50, "*learning rate*" at 0.008, and "*batchsize*" at 12.

Then, the 2D-CNN model was established using the optimized parameters (as depicted in Step 1 in Fig. 2). The classification results for the test set are illustrated in Fig. 4a and Fig. 4b. The model achieved impressive performance on the test set, misclassifying only one out of 47 samples. The final evaluation metrics on the test set obtain an accuracy of 97.92%, precision of 98.12%, recall of 97.91%, and F1_score of 97.92% (Table 1). These results highlight the robustness and precision of the 2D-CNN architecture in making accurate predictions, validating its potential for application in classification tasks. Moreover, the high accuracy and consistency across multiple evaluation metrics suggest that the 2D-CNN model is well-suited for applications that require precise and reliable classification.

In order to enhance the diversity and complexity of the data, reduce model overfitting, and improve model robustness, ten data augmentation methods were employed, as illstarted in Fig. 5. While the performance of a model often depends on the richness of the provided information (Luo et al., 2022), excessive augmentation can lead to information loss, making it challenging for the model to learn effective features. Additionally, it increases computational demands, prolongs training time and escalates resource costs. Therefore, it is crucial to carefully selected appropriate augmentation methods. To explore the impact of different augmentation strategies, we randomly performed combinations of 3, 7, and all 10 augmentation methods, resulting in datasets of different sizes (Table 1). These datasets were also split into training, validation and test sets, and 2D-CNN models were established accordingly. Among them, the model utilizing the three augmentation (3-DA) methods achieved the highest classification accuracy. The corresponding training and validation loss and accuracy for this model are presented in Fig. 3b. In constrast, the models based on 7-DA and 10-DA combinations, although benefiting from larger data sizes, exhibited a significant drop in model performance. When comparing the accuracy before (Fig. 4a and Fig. 4b) and after augmentation (Fig. 4c and Fig. 4d), it was found that the model's accuracy remained consistent, suggesting that the original data was sufficiently diverse and representative, providing the important features needed for the model to learn effectively. As a result, further augmentation did not lead to significant improvements in performance. This enabled the 2D-CNN to achieve optimal discrimination accuracy without the need for additional augmentation. Consequently, we opted to use the original data for subsequent analyses.

4.2.2. Visualization of 2D-CNN identified features

The Grad-CAM++ algorithm is employed to generate a heatmap of the gradients of the weights in the final convolutional layer, thus providing a visual representation of the network model's decision-making process (as illustrated in Step 2 of Fig. 2). On the Grad-CAM++ heatmap, the darker red areas indicate the features that are most critical for category localisation. In this study, these features were visualized, and the results for Chenpi samples from four different years are displayed in Fig. 6a. It is evident that the highlighted areas vary across different years, reflecting how the model's focus shifts based on the imput spectral images. This visualization enhances the transparency and interpretability of the model's decisions, allowing us to intuitively grasp the key features the model focuses on when distinguishing between categories.

Following this, the effectiveness of the features selected by 2D-CNN were validated, through successful validation of the 2D-CNN selected

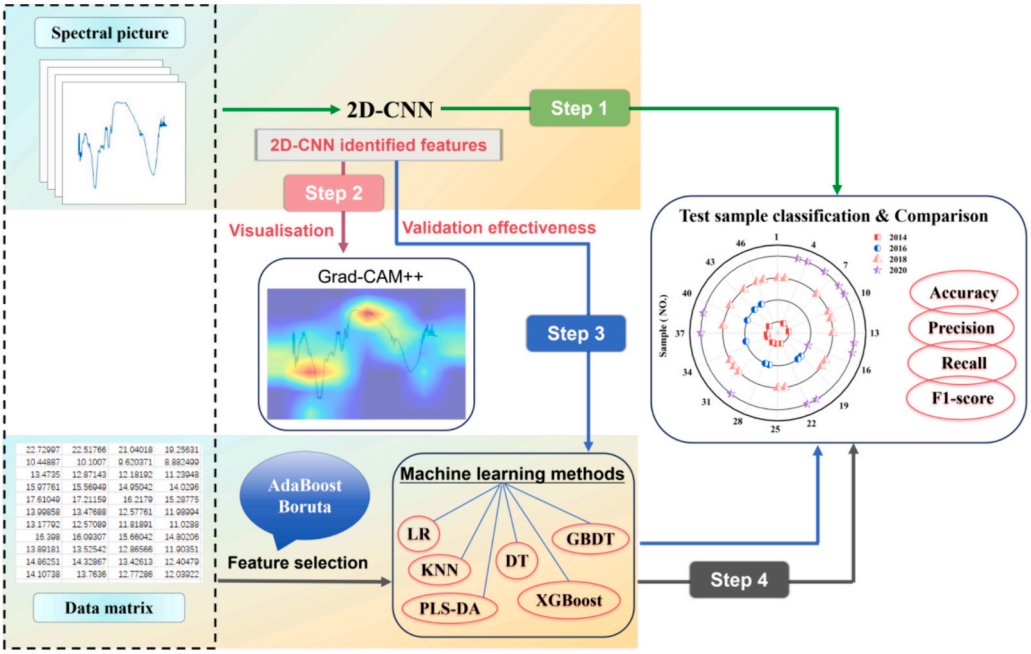


Fig. 2. The process of 2D-CNN for classification and model performance validation.

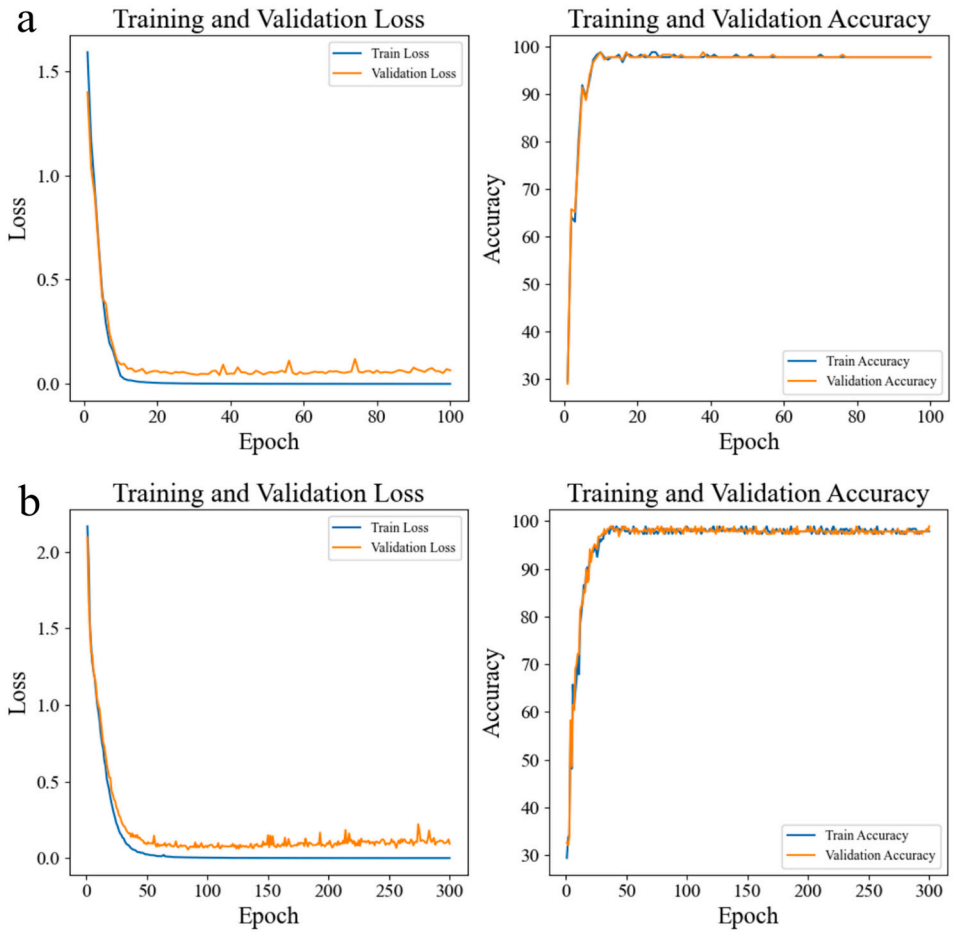


Fig. 3. Loss and accuracy curves: (a) 2D-CNN model before data augmentation, (b) 2D-CNN model after data enhancement.

features can further confirm their relevance and importance in accurate classification.

4.2.3. *Validation effectiveness of 2D-CNN identified features*

To validate the effectiveness of the features identified by the 2D-CNN, we combined the most critical features pinpointed by the Grad-CAM++ method, as depicted in Fig. 6b. These features were subsequently input into six machine learning methods-GBDT, LR, DT, XGBoost, KNN and PLS-DA (as illustrated in Step 3 of Fig. 2). The performance of six machine learning models established using 2D-CNN identified features were presented in Table 2. In addition, the comparison of reference year and predicted year of Chenpi samples in test set was illustrated in Fig. S2. Among the models, the GBDT and LR models established using 2D-CNN identified features show similar high performance with 2D-CNN model established on original images. This suggests that the feature selection process was successful in isolating the most informative aspects of the spectral images, thereby maintaining high classification accuracy. Although the DT, XGBoost and KNN moldes on 2D-CNN identified features show a drop in performace compared to previous methods, they still perform reasonable well, underscoring the robustness and generalizability of the features. These findings are statistically supported by cross-validation, further confirming the effectiveness of the features extracted by the 2D-CNN. Finally, the PLS-DA model on the 2D-CNN identified features shows the lowest performace among all the models. This deviation from the expected results suggests the need for further investigation. To address this, we extend our research by appliying the six aforementioned machine learning methods in conjunction with different variable selection methods to establish discriminant models.

4.3. *Machine learning modelling results*

At first, six machine learning models were established without applying any feature selection techniques. As can be seen from Table 2, models such as XGBoost and DT showed high accuracy values of 93.62% and 91.49%, respectively, demonstrating their strong performance in the classification task. Similarly, the LR, KNN, and PLS-DA models also performed relatively well, with accuracy values of 87.13%, 87.23% and 85.21%, respectively, indicating their ability to handle the dataset effectively despite the presence of potentially irrelevant features. However, GBDT model had a lower accuracy of 70.21%, indicating that the model might be negatively affected by irrelevant or redundant variables. This observation underscores the importance of feature selection in enhancing model performance by eliminating noise and focusing on the most informative variables.

Then, we applied Adaboost and Boruta algorithms to achieve feature selection, resulting in the extraction of 37 and 358 features from the original 7468 spectral data points (Fig. 6c and Fig. 6d), respectively. The results presented in Table 2 indicated that the introduction of feature selection had a significant impact on model performance. Specifically, AdaBoost feature selection markedly improved the classification accuracy of GBDT, DT and XGBoost models, maintained the accuracy of KNN, while slightly decreased the accuracy for LR and significantly for PLS-DA. On the other hand, Boruta feature selection improved the accuracy of GBDT, LR and DT models, while reduced the accuracy of KNN, PLS-DA and XGBoost models. These results highlight the importance of integrating machine learning algorithms with appropriate feature selection methods to improve classification accuracy.

Finally, the results of six models based on three feature selection

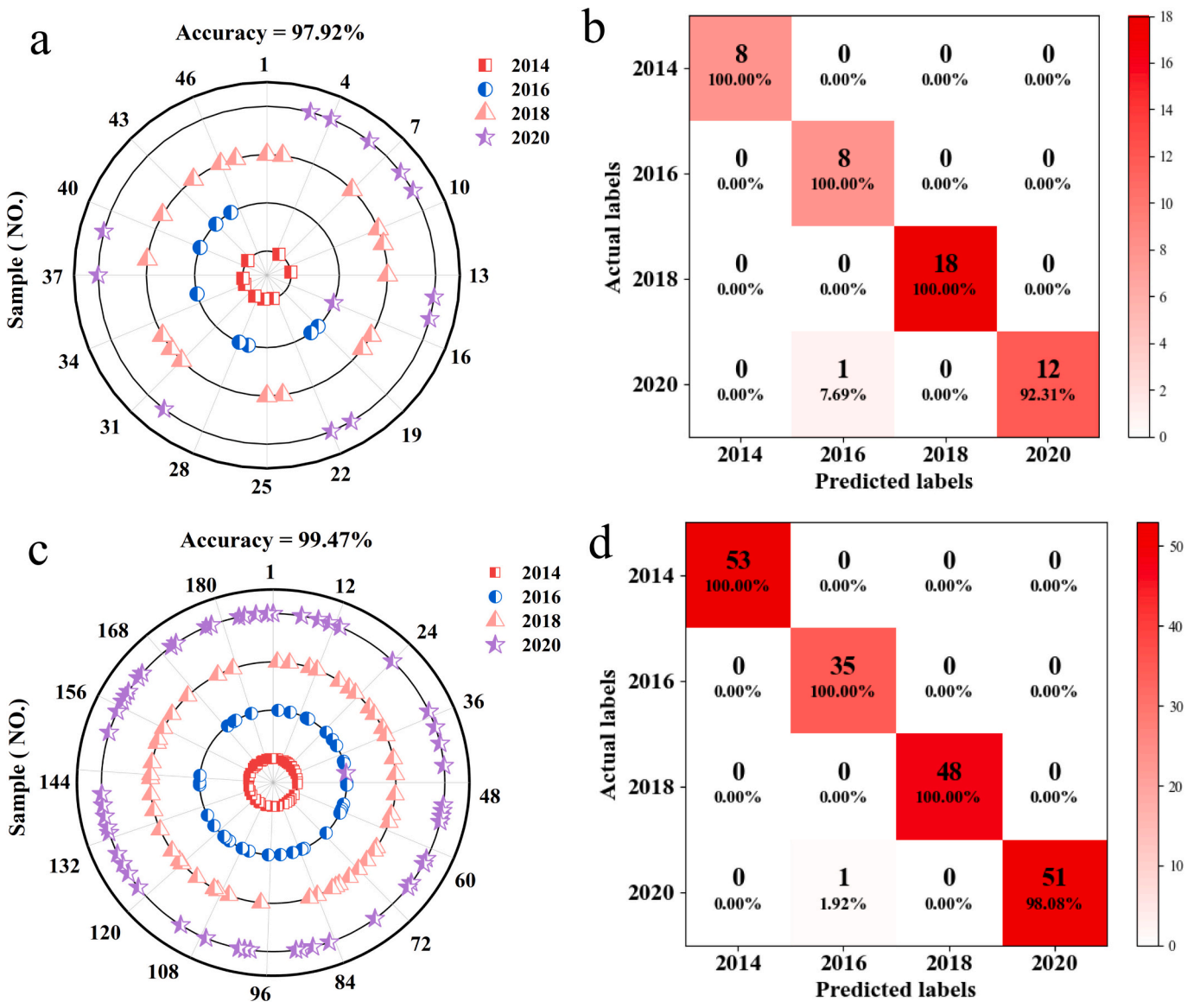


Fig. 4. Radar chart of classification results: (a) 2D-CNN model before data augmentation, (c) 2D-CNN model after data enhancement. Confusion matrix of classification results: (b) 2D-CNN model before data augmentation, (d) 2D-CNN model after data enhancement.

Table 1
The discrimination performance of 2D-CNN model before and after data augmentation.

Inputs	Data size	Accuracy (%)	Precision (%)	Recall (%)	F1_score (%)
Original images	234	97.92	98.12	97.91	97.92
3-DA	936	99.47	99.48	99.47	99.47
7-DA	1872	75.73	64.45	75.73	67.23
10-DA	2574	87.02	89.13	87.02	85.57

methods were compared (Fig. S3). The comparsion reveals that models such as GBDT, LR, KNN, and PLS-DA built on 2D-CNN identified features exhibit higher accuracy than those established using other two feature selection methods (Fig. S3a). This demonstrates the superior ability of 2D-CNN to extract relevant and informative features, thereby enhancing the performance of these models compared to Adaboost and Boruta. In addition, the accuracy of the DT model remains unchanged regardless of the feature selection method, while the XGBoost model established on 2D-CNN identified features achieves better accauacy with Boruta and lower with AdaBoost. The other three parameters (Fig. S3b-d) followed a similar trend as accuracy. Overall, these analysis results highlight the effectiveness of the features extracted by 2D-CNN. Moreover, the 2D-CNN model established on the FTIR spectral images demonstrated superior performance, and the features selected by 2D-CNN exhibited good classification accuracy accross machine learning models.

4.4. Key advantages of the proposed method

Compared to some reported methods used for Chenpi age discrimination (Liang et al., 2022; Qu et al., 2015; Wang et al., 2016), the proposed method's novelty lies in its unique integration of FTIR spectral images with 2D-CNN, a combination that has not been widely explored in Chenpi analysis. While traditional techniques like PCA and PLS-DA focus on data processing, this approach takes advantage of 2D-CNN's powerful image classification abilities. Additionally, the use of Grad-CAM++ for model interpretability introduces a new level of transparency to the analysis, addressing the common issue of interpretability in traditional models and allowing researchers to understand how the model arrives at its decisions.

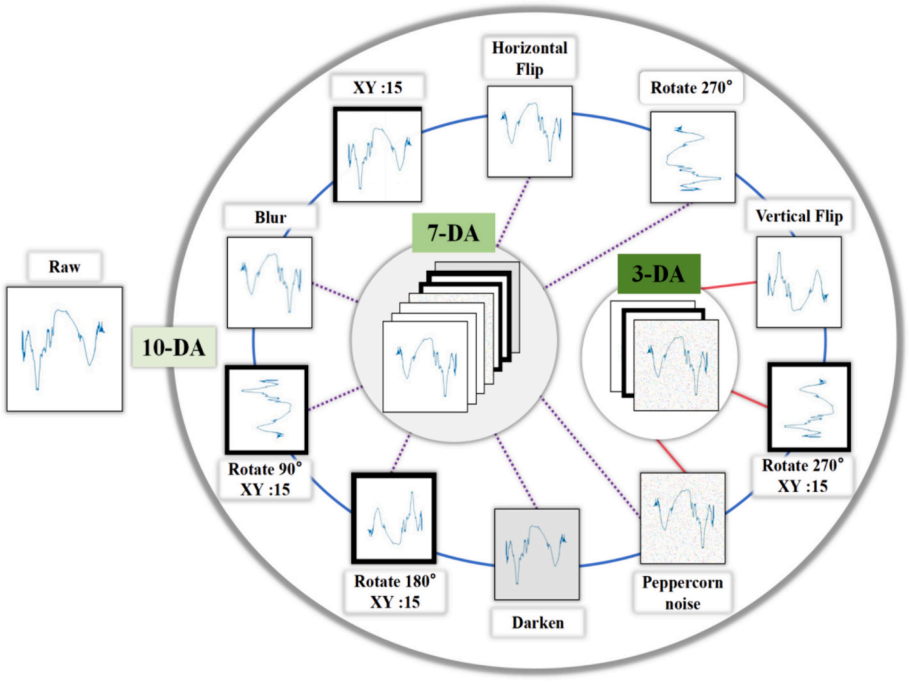


Fig. 5. The different kinds of employed data enhancement methods.

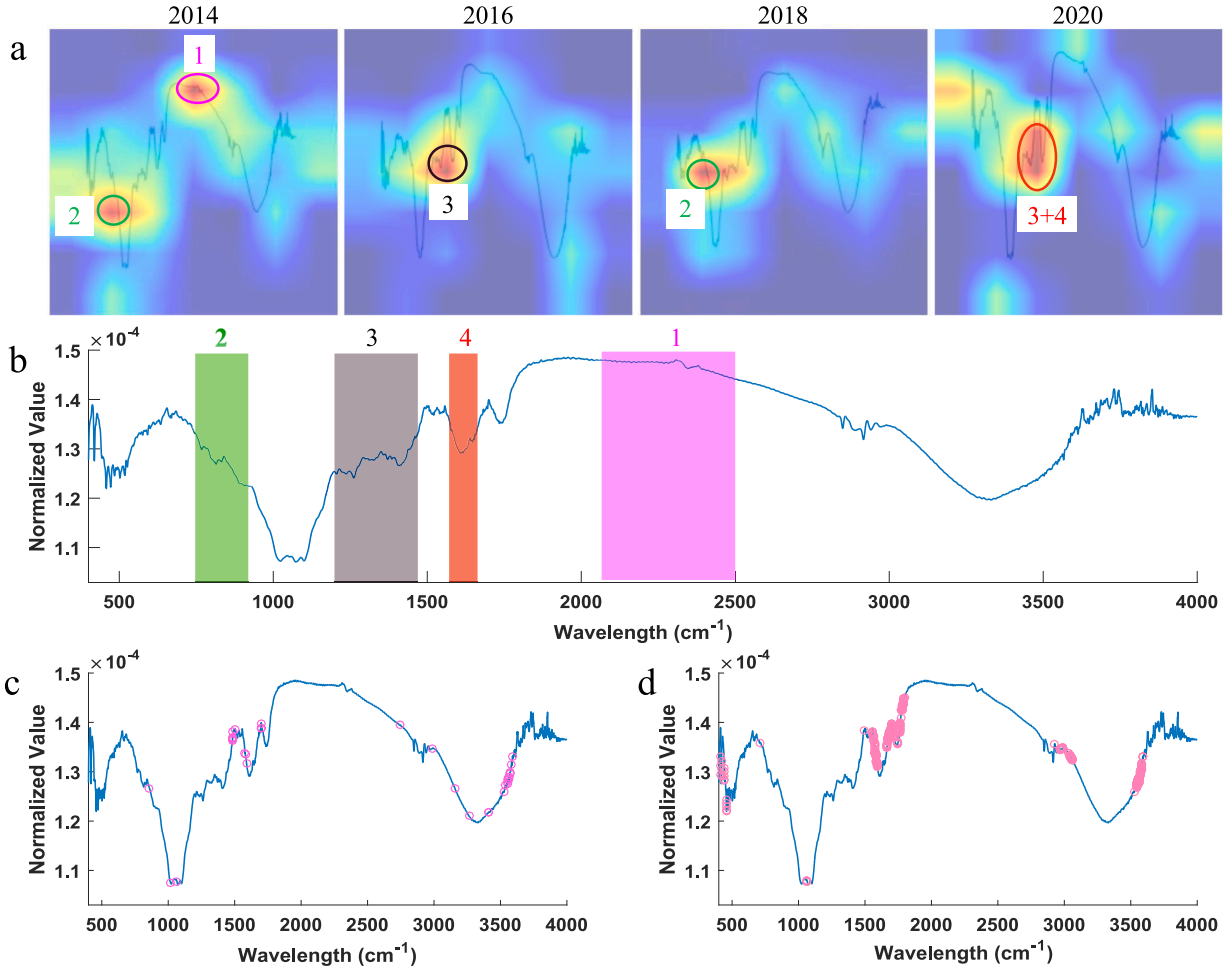


Fig. 6. (a) Grad-CAM++ visualization of 2D-CNN identified features, (b) combination of 2D-CNN extracted characteristic features, (c) AdaBoost extracted features, (d) Boruta extracted features.

Table 2
The discrimination performance of machine learning models.

Feature selection	Model methods	Accuracy (%)	Precision (%)	Recall (%)	F1_score (%)
–	GBDT	70.21	74.67	70.21	70.89
AdaBoost		91.48	92.85	91.48	91.59
Boruta		87.23	88.96	87.23	87.43
2D-CNN*		97.87	98.05	97.87	97.86
–	LR	87.13	88.05	86.67	86.21
AdaBoost		85.11	84.85	83.64	83.94
Boruta		95.74	95.83	95.45	95.23
2D-CNN*		97.87	98.21	98.21	98.21
–	DT	91.49	93.49	91.49	91.51
AdaBoost		93.62	94.81	93.62	93.44
Boruta		93.62	95.21	93.62	93.78
2D-CNN*		93.62	95.09	93.62	93.53
–	XGBoost	93.62	94.23	93.91	93.47
AdaBoost		95.74	95.58	95.58	95.58
Boruta		80.85	84.00	81.57	80.80
2D-CNN*		91.49	91.49	92.15	91.47
–	KNN	87.23	88.01	87.23	87.27
AdaBoost		87.23	88.01	87.23	87.27
Boruta		82.98	84.15	82.96	82.94
2D-CNN*		91.49	91.49	91.49	91.91
–	PLS-DA	85.11	87.17	85.11	84.89
AdaBoost		72.34	74.87	72.34	72.57
Boruta		68.09	67.23	68.09	66.73
2D-CNN*		80.85	81.96	80.85	80.23

2D-CNN*: features identified by 2D-CNN.

Overall, this method stands out from traditional approaches by combining a fast and non-destructive spectral technique, a deep learning algorithm, and easy-to-understand results, making the analysis more thorough and insightful.

5. Conclusions

The present study has illustrated the success of using 2D-CNN combined with FTIR spectral images for the discrimination of Chenpi age, achieving the classification accuracy of 97.92%. Moreover, the features identified by the 2D-CNN were visualized using the Grad-CAM++ technique, which helps to avoid the "black box" phenomenon commonly associated with deep learning models. To further validate the effectiveness of the features identified by the 2D-CNN, they were input into six different machine learning methods to establish classification models. The results indicated that the 2D-CNN model established on the FTIR spectral images not only outperformed other used methods but also demonstrated that the features selected by 2D-CNN show good interpretability and classification accuracy across various machine learning models. Therefore, our research lays a solid foundation for further research and the optimisation of deep learning based feature extraction methods.

Acknoledgement and funding

This research was funded by "Development and Research on Methods for Identifying the Age and Origin of Xinhui Chenpi" (Grant number: 34220002).

CRediT authorship contribution statement

Li Jun Tang: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Xin Kang Li**: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yue Huang**: Investigation, Formal analysis, Data curation. **Xiang-Zhi Zhang**: Writing – review & editing, Supervision. **Bao Qiong Li**: Writing – review & editing, Visualization, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fochx.2024.101759.

References

Adegun, A. A., Viriri, S., & Tapamo, J.-R. (2023). Review of deep learning methods for remote sensing satellite images classification: Experimental survey and comparative analysis. *Journal of Big Data, 10*(1), 93. https://doi.org/10.1186/s40537-023-00772-x

Akturk, B., Beyaztas, U., Shang, H. L., & Mandal, A. (2024). Robust functional logistic regression. *Adv. Data Anal. Classif.*. https://doi.org/10.1007/s11634-023-00577-z

Ali, N., Neagu, D., & Trundle, P. (2019). Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. *SN Applied Sciences, 1*(12), 1559. https://doi.org/10.1007/s42452-019-1356-9

Cai, Z., Huang, Z., He, M., Li, C., Qi, H., Peng, J., Zhou, F., & Zhang, C. (2023). Identification of geographical origins of Radix Paeoniae Alba using hyperspectral imaging with deep learning-based fusion approaches. *Food Chemistry, 422*, Article 136169. https://doi.org/10.1016/j.foodchem.2023.136169

Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V. N., & IEEE. (2018). *Grad-CAM plus plus : Generalized gradient-based visual explanations for deep convolutional networks. In 18th IEEE winter conference on applications of computer vision (WACV), (pp. 839-847)*. Nv.

Chen, W., Li, Y., Xue, W. F., Shahabi, H., Li, S. J., Hong, H. Y., … Bin Ahmad, B. (2020). Modeling flood susceptibility using data-driven approaches of naive Bayes tree, alternating decision tree, and random forest methods. *Science of the Total Environment, 701*, Article 134979. https://doi.org/10.1016/j.scitotenv.2019.134979

Cunningham, P., & Delany, S. J. (2021). K-nearest neighbour classifiers - a tutorial. *ACM Computing Surveys, 54*(6), 128. https://doi.org/10.1145/3459665

Dong, Y. (2024). Convolutional neural networks for sensitive identification of tea species using electrochemical sensors. *Journal of Food Measurement and Characterization.*. https://doi.org/10.1007/s11694-024-02615-6

Dou, B. Z., Zhu, Z. L., Merkurjev, E., Ke, L., Chen, L., Jiang, J., … Wei, G. W. (2023). Machine learning methods for small data challenges in molecular science. *Chemical Reviews, 123*(13), 8736–8780. https://doi.org/10.1021/acs.chemrev.3c00189

Hao, Z., Jikai, W., Zonghai, C., Shiqi, L., Peng, B., & Meng, X. (2023). Interpretability-mask: A label-preserving data augmentation scheme for better classification. *Signal, Image and Video Processing, 17*(6), 2799–2808.

Jin, B. C., Zhang, C., Jia, L. Q., Tang, Q. Z., Gao, L., Zhao, G. W., & Qi, H. N. (2022). Identification of Rice seed varieties based on near-infrared hyperspectral imaging technology combined with deep learning. *ACS Omega, 7*, 4735–4749. https://doi.org/10.1021/acsomega.1c04102

Khalifa, N. E., Loey, M., & Mirjalili, S. (2022). A comprehensive survey of recent trends in deep learning for digital images augmentation. *Artificial Intelligence Review, 55*(3), 2351–2377. https://doi.org/10.1007/s10462-021-10066-4

Kim, K. (2016). A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree. *Pattern Recognition, 60*, 157–163. https://doi.org/10.1016/j.patcog.2016.04.016

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software, 36*(11), 1–13. https://doi.org/10.18637/jss.v036.i11

Li, S. Z., Guan, X.-M., Gao, Z., Lan, H.-C., Yin, Q., Chu, C., … Zhou, P. (2019). A simple method to discriminate Guangchenpi and Chenpi by high-performance thin-layer chromatography and high-performance liquid chromatography based on analysis of dimethyl anthranilate. *Journal of Chromatography B, 1126-1127*, Article 121736. https://doi.org/10.1016/j.jchromb.2019.121736

Li, Y., Zhao, W., Qian, M., Wen, Z., Bai, W., Zeng, X., Wang, H., Xian, Y., & Dong, H. (2024). Recent advances in the authentication (geographical origins, varieties and aging time) of tangerine peel (Citri reticulatae pericarpium): A review. *Food Chemistry, 442*, Article 138531. https://doi.org/10.1016/j.foodchem.2024.138531

Liang, S., Wen, Z., Tang, T., Liu, Y., Dang, F., Xie, T., & Wu, H. (2022). Study on flavonoid and bioactivity features of the pericarp of Citri Reticulatae 'chachi' during storage. *Arabian Journal of Chemistry, 15*(3), Article 103653. https://doi.org/10.1016/j.arabjc.2021.103653

Liang, W. Z., Luo, S. Z., Zhao, G. Y., & Wu, H. (2020). Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms. *Mathematics, 8*(5), 765. https://doi.org/10.3390/math8050765

Liu, Z., Su, W., Ao, J., Wang, M., Jiang, Q., He, J., Gao, H., Lei, S., Nie, J., Yan, X., Guo, X., Zhou, P., Hu, H., & Ji, M. (2022). Instant diagnosis of gastroscopic biopsy via deep-learned single-shot femtosecond stimulated Raman histology. *Nature Communications, 13*(1), 4050. https://doi.org/10.1038/s41467-022-31339-8

Luo, S. H., Wang, W. L., Zhou, Z. F., Xie, Y., Ren, B., Liu, G. K., & Tian, Z. Q. (2022). Visualization of a machine learning framework toward highly sensitive qualitative analysis by SERS. *Analytical Chemistry, 94*(28), 10151–10158. https://doi.org/10.1021/acs.analchem.2c01450

Moujahid, H., Cherradi, B., Al-Sarem, M., Bahatti, L., Eljialy, A., Alsaeedi, A., & Saeed, F. (2022). Combining CNN and grad-cam for COVID-19 disease prediction and visual explanation. *Intelligent Automation and Soft Computing, 32*(2), 723–745. https://doi.org/10.32604/iasc.2022.022179

Naeim Mohamad Asri, M., Verma, R., Arafat Mahat, N., Azman Mohd Nor, N., Nur Syuhaila Mat Desa, W., & Ismail, D. (2022). Raman spectroscopy with self-organizing feature maps and partial least squares discriminant analysis for discrimination and source correspondence of red gel ink pens. *Microchemical Journal, 175*, Article 107170. https://doi.org/10.1016/j.microc.2021.107170

Pan, S., Zhang, X., Xu, W., Yin, J., Gu, H., & Yu, X. (2022). Rapid on-site identification of geographical origin and storage age of tangerine peel by near-infrared spectroscopy. *Spectrochimica Acta Part A, 271*, Article 120936. https://doi.org/10.1016/j.saa.2022.120936

Pokhrel, D. R., Sirisomboon, P., Khurnpoon, L., Posom, J., & Saechua, W. (2023). Comparing machine learning and PLSDA algorithms for durian pulp classification using inline NIR spectra. *Sensors, 23*(11), 5327. https://doi.org/10.3390/s23115327

Qin, Y., Zhao, Q., Zhou, D., Shi, Y., Shou, H., Li, M., Zhang, W., & Jiang, C. (2024). Application of flash GC e-nose and FT-NIR combined with deep learning algorithm in preventing age fraud and quality evaluation of pericarpium citri reticulatae. *Food Chemistry: X, 21*, Article 101220. https://doi.org/10.1016/j.fochx.2024.101220

Qiu, Y. G., Zhou, J., Khandelwal, M., Yang, H. T., Yang, P. X., & Li, C. Q. (2022). Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration. *Engineering with Computers, 38* (Suppl. 5), 4145–4162. https://doi.org/10.1007/s00366-021-01393-9

Qu, X., Li, H., Yang, X., Tan, M., Ao, H., & Wang, J. (2015). Artificial neural network analysis of Xinhui Pericarpium *Citri Reticulatae* using gas chromatography - mass spectrometer - automated mass spectral deconvolution and identification system. *Tropical Journal of Pharmaceutical Research, 14*(11), 2071. https://doi.org/10.4314/tjpr.v14i11.17

Sheridan, R. P., Wang, W. M., Liaw, A., Ma, J. S., & Gifford, E. M. (2016). Extreme gradient boosting as a method for quantitative structure-activity relationships. *Journal of Chemical Information and Modeling, 56*(12), 2353–2360. https://doi.org/10.1021/acs.jcim.6b00591

Shi, J. J., Peng, L. H., Chen, W. X., Qiao, W. L., Wang, K., Xu, Y. Y., & Cheng, J. L. (2024). Evaluation of chemical components and quality in Xinhui Chenpi (*Citrus reticulata* "Chachi") with two different storage times by GC-MS and UPLC. *Food Science & Nutrition, 00*, 1–16. https://doi.org/10.1002/fsn3.4154

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Sun, X. M., Deng, H. D., Shan, B. J., Shan, Y. Q., Huang, J. Y., Feng, X. S., … Yang, Q. (2023). Flavonoids contribute most to discriminating aged Guang Chenpi (*Citrus reticulata* Chachi') by spectrum-effect relationship analysis between LC-Q-Orbitrap/MS fingerprint and ameliorating spleen deficiency activity. *Food Science & Nutrition, 11*(11), 7039–7060. https://doi.org/10.1002/fsn3.3629

Tang, J. Y., Henderson, A., & Gardner, P. (2021). Exploring AdaBoost and random forests machine learning approaches for infrared pathology on unbalanced data sets. *Analyst, 146*(19), 5880–5891. https://doi.org/10.1039/d0an02155e

Wang, H., Chen, G., Fu, X., & Liu, R. H. (2016). Effects of aging on the phytochemical profile and antioxidative activity of *Pericarpium Citri Reticulatae* "Chachiensis". *RSC Advances, 6*(107), 105272–105281. https://doi.org/10.1039/c6ra22082g

Wang, Q., Qiu, Z. Y., Chen, Y. L., Song, Y. Q., Zhou, A. M., Cao, Y., … Song, M. Y. (2023). Review of recent advances on health benefits, microbial transformations, and authenticity identification of *Citri reticulatae* Pericarpium bioactive compounds. *Critical Reviews in Food Science and Nutrition, 1-29*. https://doi.org/10.1080/10408398.2023.2222834

Wu, W. M., Wang, J. X., Huang, Y. S., Zhao, H. Y., & Wang, X. T. (2021). A novel way to determine transient heat flux based on GBDT machine learning algorithm. *International Journal of Heat and Mass Transfer, 179*, Article 121746. https://doi.org/10.1016/j.ijheatmasstransfer.2021.121746

Yang, M., Jiang, Z., Wen, M., Wu, Z., Zha, M., Xu, W., & Zhang, L. (2022). Chemical variation of Chenpi (Citrus peels) and corresponding correlated bioactive compounds by LC-MS metabolomics and multibioassay analysis. *Fronters in Nutrion, 9*, Article 825381. https://doi.org/10.3389/fnut.2022.825381

Yu, Y., Yu, Q. H., Luo, R. S., Chen, S., Yang, J. B., & Yan, F. W. (2024). Degradation and polarization curve prediction of proton exchange membrane fuel cells: An interpretable model perspective. *Applied Energy, 365*, Article 123289. https://doi.org/10.1016/j.apenergy.2024.123289

Zhang, X., Gao, Z., Yang, Y., Pan, S., Yin, J., & Yu, X. (2022). Rapid identification of the storage age of dried tangerine peel using a hand-held near infrared spectrometer and machine learning. *Journal of Near Infrared Spectroscopy, 30*(1), 31–39. https://doi.org/10.1177/09670335211057232

Zhao, J., Pan, F. J., Li, Z. M., Lan, Y. B., Lu, L. Q., Yang, D. J., & Wen, Y. T. (2021). Detection of cotton waterlogging stress based on hyperspectral images and convolutional neural network. *International Journal of Agricultural and Biological Engineering, 14*(2), 167–174. https://doi.org/10.25165/j.ijabe.20211402.6023