

Article

Kernel-Independent Component Analysis for Near-Infrared Spectroscopic Prediction of Tannin Content in Sorghum Grains

Wen-Peng Luo ¹, Yue He ², Yu Wei ¹, Zheng-Guang Chen ^{2,*} and Bing Li ^{1,*}¹ College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150006, China; luowenpeng@hrbeu.edu.cn (W.-P.L.); weiyu@hrbeu.edu.cn (Y.W.)² College of Information and Electrical Engineering, Heilongjiang Bayi Agricultural University, Daqing 163319, China; heuy@byau.edu.cn

* Correspondence: ruzee@byau.edu.cn (Z.-G.C.); libing265@hrbeu.edu.cn (B.L.)

Abstract

To eliminate the complex nonlinear mixing relationships among spectral features in near-infrared (NIR) quantitative analysis, and to overcome the limitations of principal component analysis (PCA), which relies solely on covariance structure and linear assumptions and is therefore incapable of effectively handling nonlinear signals, this study employs kernel-independent component analysis (KICA), for nonlinear feature extraction from NIR spectra, combined with a regression model to achieve rapid detection of tannin content in sorghum grains. KICA effectively separates nonlinearly mixed source signals by mapping spectral data into a high-dimensional feature space via the kernel trick. The prediction model built on KICA-extracted features and support vector regression (SVR) consistently delivered the highest test-set prediction accuracy and exhibited the smallest training-to-test R^2 gap among all evaluated models across repeated random splits, confirming its superiority over PCA-based feature extraction methods and standalone SVR, and its competitive performance relative to ICA-based methods, in both predictive accuracy and generalization capability. Additionally, KICA yielded a lower reconstruction error for the original spectra, indicating its ability to more completely retain the nonlinear informative content of the spectral data. By calculating the mean absolute coefficient of each independent component, it was found that the component with the highest contribution was strongly correlated with the wavelength range near the characteristic absorption peaks of tannin, thereby enhancing the chemical interpretability of the features. On a publicly available corn NIR dataset, the proposed method also achieved superior prediction results compared with benchmark methods, validating its generalization capability across different sample types and quality attributes. This study confirms the feasibility of introducing nonlinear blind source separation via KICA into NIR quantitative analysis, offering a promising approach for spectral feature extraction in the rapid quality assessment of agricultural products with complex matrices.



Academic Editor: Roberto Alves Braga Júnior

Received: 11 May 2026

Revised: 29 June 2026

Accepted: 30 June 2026

Published: 2 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: nonlinear feature extraction; kernel-independent component analysis; tannin content; near-infrared spectroscopy

1. Introduction

Sorghum is one of the world's important cereal crops, widely cultivated in China, primarily for brewing, animal feed, and food processing [1]. Tannin is a characteristic phenolic compound found in sorghum grains, mainly located in the seed coat. Its content

varies significantly due to differences in variety, origin, and cultivation conditions. Tannin has a dual impact on sorghum quality: during brewing, it can be converted into flavor compounds such as syringic acid and syringaldehyde, which impart a unique aroma to spirits, inhibit the growth of spoilage bacteria, and increase alcohol yield [2]; however, when used as animal feed, it acts as an antinutritional factor, forming indigestible complexes with proteins and starches, thereby reducing the absorption and utilization of nutrients by animals, and even affecting palatability [3]. Therefore, accurately measuring tannin content in sorghum is crucial for variety breeding, processing utilization, and quality control. Currently, the determination of tannin content primarily relies on traditional chemical methods, such as the vanillin method, ferric ammonium citrate method, and Prussian blue method. These methods are time-consuming, labor-intensive, sample-destructive, and reliant on chemical reagents, making them unsuitable for large-scale rapid screening [4]. There is an urgent need for a fast, non-destructive detection technique.

Near-infrared spectroscopy (NIRS) is based on the vibration absorption characteristics of hydrogen-containing groups (such as C-H, O-H, N-H) in molecules [5]. By measuring the absorption or reflection spectra of samples in the near-infrared region, it is possible to obtain information about the chemical composition of the sample. This technique is widely used in the agricultural and food industries, enabling rapid quantitative or qualitative analysis of sample quality by revealing spectral features of the molecular structures of major components such as moisture and proteins [6]. However, NIRS data itself has high dimensionality, with individual samples typically containing hundreds to thousands of wavelength variables, and severe multicollinearity between these variables, leading to potential overfitting when directly used for modeling [7]. Measurement interferences, such as light scattering, particle size differences, and temperature fluctuations, can also introduce significant nonlinear distortions in the spectra, making the relationship between the spectral signals and tannin content far from simple linear correspondence. Furthermore, the measured spectra are often contaminated by overlapping information from other components such as moisture and proteins [8]. Effectively extracting feature information that is truly related to the target content from this high-dimensional, nonlinear, and interference-prone spectral data is a significant challenge in using near-infrared spectroscopy for the analysis of agricultural products and food quality.

In recent years, researchers have proposed various dimensionality reduction strategies to address the high-dimensional collinearity problem in NIR spectroscopy, primarily including feature selection and feature extraction methods. Feature selection methods involve selecting a subset of the original wavelengths based on specific criteria, offering advantages such as simplicity and strong interpretability [9]. However, these methods carry the risk of losing collaborative information across the entire spectral range and may not achieve optimal prediction performance. Therefore, this study adopts a feature extraction strategy, aiming to retain more complete effective information from the spectra. Principal component analysis (PCA) is one of the most widely used feature extraction methods [10]. It extracts linearly independent principal components through orthogonal transformation, retaining the primary information in the order of variance contribution, with simple calculation and intuitive results. PCA has been extensively applied in various qualitative and quantitative analysis tasks [11]. However, PCA relies solely on the covariance structure of the data, which cannot effectively distinguish between signal and noise, and its linear assumptions are limited when dealing with the nonlinear relationships commonly present in NIR spectra. When effects caused by factors such as light scattering and particle size differences are pronounced, the principal components extracted by PCA have difficulty accurately characterizing the true association between the target analyte and the spectra, thereby compromising the accuracy of subsequent quantitative analysis. Independent

component analysis (ICA), based on the assumptions of non-Gaussianity and statistical independence, is a linear blind source separation technique [12]. Unlike PCA, which pursues orthogonal components, ICA seeks statistically independent components, which is more aligned with the actual process of NIR spectra, where multiple chemical components (such as water, proteins, and tannins) are mixed. Theoretically, ICA is better suited for separating target signals from background interferences. Moreover, ICA is more sensitive to effective signals that deviate from Gaussian distribution, helping to enhance the signal-to-noise ratio. ICA has demonstrated reliable signal separation capabilities in fields such as speech, EEG, and image separation [13,14], but its applicability and effectiveness in qualitative and quantitative analysis of NIR spectra have not been fully validated. Furthermore, as a linear method, ICA is based on the fundamental assumption that the source signals are mixed linearly during the observation process [15]. When the aforementioned distortions are present in real spectra, this method has difficulty correctly separating the source signals. Since such effects are prevalent in real NIR spectra, it is necessary to introduce methods capable of handling nonlinear mixing. To address this, this study further introduces kernel-independent component analysis (KICA) to explore the feasibility of nonlinear feature extraction methods in NIR quantitative analysis.

Kernel-independent component analysis (KICA) is a nonlinear blind source separation method developed on the basis of ICA [16]. It first implicitly maps the original data into a high-dimensional feature space via a kernel function, transforming the nonlinear mixing problem into a linear one in that space, and then performs independent component analysis therein to separate the nonlinearly mixed source signals. KICA has been widely applied in fields involving nonlinear mixing, such as image processing, biomedical signal analysis, and fault diagnosis, demonstrating superior processing capabilities compared with traditional ICA [17]. When introduced into NIR spectroscopic analysis, KICA theoretically offers better handling of nonlinear effects caused by factors such as light scattering and particle size differences, thereby overcoming the limitations of linear methods in complex spectral interpretation. In the field of chemometrics and spectral analysis, several studies have explored the use of kernel-based methods for nonlinear calibration [18], yet the specific application of KICA to NIR quantitative analysis—particularly for cereal grains—has received limited attention compared with linear methods such as PCA and PLS. This gap motivates the present investigation into the potential of KICA for feature extraction from sorghum NIR spectra. The independent components extracted by KICA can be further analyzed for their correlations with the original spectral wavelengths, allowing the contribution of different wavelengths to each independent component to be evaluated. This approach can, to a certain extent, alleviate the issue of poor interpretability in the dimensionality reduction of NIR spectra, providing a chemical basis for the prediction results.

Therefore, to address the issues of high dimensionality, nonlinearity, and mixed interference in NIR quantitative analysis, this study takes different varieties of sorghum grains as the research object and proposes a quantitative analysis framework that integrates wavelet transform denoising and kernel-independent component analysis (KICA) for nonlinear feature extraction. Specifically, wavelet transform is employed to filter out high-frequency noise from the spectra and achieve data dimensionality reduction. Owing to its excellent time–frequency localization characteristics, wavelet transform can preserve the main spectral profile while suppressing noise, making it well suited for processing non-stationary NIR signals [19]. KICA is used to separate nonlinearly mixed source signals and obtain independent component features. The two methods are functionally complementary: wavelet transform serves as a preprocessing step, providing denoised low-dimensional input for KICA, and KICA further separates the nonlinearly mixed source signals on this basis [20]. To evaluate the quality of the KICA-extracted features, a regression model is

built on the extracted independent components, and the model performance is compared with that of two linear feature extraction methods, principal component analysis (PCA) and independent component analysis (ICA), to assess the advantages of the KICA-based model in terms of predictive performance and feature representation. In addition, correlation analysis between the independent components and the original spectral wavelengths is conducted to preliminarily explore their relationship with spectral features, aiming to alleviate, to a certain extent, the common “black-box” problem in NIR modeling.

This study aims to verify the feasibility and effectiveness of KICA in NIR quantitative analysis so as to provide a new methodological reference for the quality assessment of samples with complex matrices. The specific objectives are as follows: (1) to verify the feasibility of KICA for nonlinear feature extraction from sorghum NIR spectra; (2) to compare the differences in feature extraction quality among KICA, PCA, and ICA and to evaluate their impact on prediction accuracy through quantitative modeling; (3) to investigate the chemical interpretability of the independent components through wavelength correlation analysis; and (4) to validate the generalization capability of the proposed feature extraction method on a publicly available corn dataset.

2. Materials and Methods

2.1. Sample Selection and Preparation

This study used sorghum samples harvested in 2022 from Heilongjiang Bayi Agricultural University as the sample source. The samples encompassed 65 varieties, with 1–7 samples selected per variety, resulting in a total of 305 samples. Near-infrared spectra were obtained using a TANGO FT-NIR spectrometer (Bruker Optik GmbH, Ettlingen, Germany) with a spectral resolution of 8 cm^{-1} and a spectral range of $11,542\text{--}3940\text{ cm}^{-1}$. Each sorghum sample was scanned 32 times to obtain an average spectrum. After scanning the complete grain spectra of each sorghum sample, the grains were ground and sieved through a 40-mesh sieve. The tannin content was then determined using the ferric ammonium citrate method.

2.2. Determination of Tannin Content

The tannin content was determined according to the national standard (GB/T 15686-2008). Before the determination, the moisture content of the samples was measured using the direct drying method, with moisture content calculated as follows [17]:

$$H = \frac{M_1 - M_2}{M_1} \times 100$$

where H is the moisture content of the sample (%); M_1 and M_2 are the mass of the sample before and after drying (g), respectively.

The procedure for tannin determination was as follows: A 1.0 g portion of the ground sorghum sample (40-mesh) was weighed, 20 mL of dimethylformamide solution was added, and the mixture was magnetically stirred for 60 min [21]. The extract was then centrifuged at $3000 \times g$ for 10 min. Two 1 mL aliquots of the supernatant were taken: one aliquot was mixed with ferric ammonium citrate solution and ammonia solution for color development, while the other was mixed with water and ammonia solution, with water replacing the ferric ammonium citrate solution as the sample blank. The absorbance was measured at 525 nm against a water blank. A standard curve was constructed using tannic acid as the standard, with a correlation coefficient $R^2 \geq 0.995$. Tannin content was expressed on a dry

matter basis. All samples were determined in triplicate, and the results were averaged. The tannin content was calculated as follows:

$$X = \frac{2C}{M} \times \frac{100}{100 - H}$$

where C is the tannic acid concentration (g/L) obtained from the standard curve for the sample extract, and M is the mass of the sample (g).

2.3. Kernel-Independent Component Analysis

Independent component analysis (ICA) is a blind source separation technique based on higher-order statistics, aiming to recover statistically independent source signals from observed mixed signals. Unlike principal component analysis (PCA), which uses second-order statistics to remove correlations between variables, ICA maximizes non-Gaussianity measures to extract independent components that carry more intrinsic meaning. The linear ICA model can be expressed as:

$$X = AS$$

where $X \in R^{m \times n}$ is the observed data matrix, A is the mixing matrix, and S is the unknown independent source signal matrix. The goal of ICA is to find the separation matrix W such that:

$$\hat{S} = WX$$

The output components $\hat{S}$ are as statistically independent as possible. Common solution methods include the FastICA algorithm, based on maximizing negative entropy, with the iterative update formula:

$$w \leftarrow E\{xg(w^T x)\} - E\{g'(w^T x)\}w$$

where $g(\cdot)$ is a nonlinear function, and in this study, the commonly used function $g(u) = \tanh(u)$ is selected.

Traditional ICA methods assume that the observed data follows a linear mixing model. However, in near-infrared spectroscopy (NIRS) data, nonlinear relationships are commonly present, such as spectral responses influenced by factors like light scattering effects and molecular interactions. These nonlinearities make it difficult for linear ICA to effectively extract intrinsic features. To overcome this limitation, kernel-independent component analysis (KICA) introduces kernels into the ICA framework. By using nonlinear mapping, KICA implicitly projects the original data into a high-dimensional feature space, where linear ICA is performed, thus enabling the separation of nonlinear mixed signals.

Let the original input space sample set be $X = \{x_1, x_2, \dots, x_m\} \subset R^n$, and the samples are projected into the feature space $F: R^n \rightarrow F$ through a nonlinear mapping ϕ . In this space, the kernel function is defined as the inner product between samples:

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_F$$

Common kernel functions include Gaussian radial basis function (RBF), polynomial kernels, etc., and the choice of kernel directly influences the geometric structure of the feature space. In this study, the Gaussian radial basis function (RBF) kernel is used for kernel-independent component analysis (KICA), and the kernel width parameter σ is determined through cross-validation. The Gaussian RBF kernel is expressed as:

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

where σ is the kernel width parameter, controlling the local range of the function's influence.

Let the kernel matrix $K \in R^{m \times m}$ have elements $K_{ij} = k(x_i, x_j)$. To eliminate the shift effect in the feature space, the kernel matrix is centered:

$$\tilde{K} = K - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m^T K - \frac{1}{m} K \mathbf{1}_m \mathbf{1}_m^T + \frac{1}{m^2} (\mathbf{1}_m^T K \mathbf{1}_m) \mathbf{1}_m \mathbf{1}_m^T$$

where $\mathbf{1}_m$ is an m -dimensional column vector with all elements equal to 1.

The core of the KICA algorithm is performing ICA in the feature space F , typically involving two stages: whitening and independent component extraction. First, perform eigen-decomposition on the centered kernel matrix $\tilde{K}$:

$$\tilde{K} = V \Lambda V^T$$

where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m)$ is the diagonal matrix of eigenvalues, and V is the matrix of corresponding eigenvectors. The top d eigenvalues and eigenvectors are selected to construct the whitening transformation matrix:

$$W_{pca} = V_d \Lambda_d^{-1/2}$$

The whitened data in the feature space is then represented as:

$$Z = \tilde{K} W_{pca}$$

The whitening operation ensures that the components in the feature space have unit variance and are uncorrelated, laying the foundation for subsequent independent component extraction. Then, the FastICA algorithm, based on negative entropy maximization, is used to extract independent components from the whitened data. Let the separating matrix be W_{ica} , and the optimization objective is to maximize the non-Gaussianity measure. The iterative update rule is:

$$W_{ica} \leftarrow E\{zg(W_{ica}^T z)\} - E\{g'(W_{ica}^T z)\} W_{ica}$$

where $g(\cdot)$ is a nonlinear function. After iterative convergence, the independent component matrix is:

$$S = Z W_{ica}^T$$

Through the above steps, KICA maps the original nonlinear data into a low-dimensional independent component space, implicitly performing nonlinear transformations using the kernel function, and simultaneously extracting statistically independent features through ICA, providing a more robust feature representation for subsequent analysis.

2.4. Modeling and Evaluation

This study aims to predict tannin content in sorghum grains using near-infrared (NIR) spectroscopy. First, the raw NIR spectra were preprocessed, and the dataset was randomly split into a training set and a test set at a ratio of 4:1. The modeling process included the following steps: (1) wavelet transform was applied to denoise the preprocessed spectra, while the approximation coefficients were extracted for dimensionality reduction; (2) the prediction performance of different regression models was compared on the wavelet-transformed spectra, and the optimal regression model was selected; (3) three feature extraction methods—PCA, ICA, and KICA—were separately applied to extract features from the compressed spectra; (4) the extracted features were input into the optimal

regression model to construct a predictive model for sorghum tannin content, and the test set samples were predicted and evaluated.

The performance of the modeling strategy was comprehensively evaluated on the test set in terms of prediction accuracy, robustness, and feature extraction quality. The specific evaluation metrics were as follows: (1) The coefficient of determination (R^2) evaluated the degree of fit between predicted and reference values, with a value closer to 1 indicating a better fit. (2) Root mean squared error (RMSE) and ratio of performance to deviation (RPD) assessed prediction accuracy and model robustness, respectively: a lower RMSE indicated higher prediction accuracy, and an RPD greater than 2.5 suggested that the model possessed good quantitative predictive capability [22]. (3) Reconstruction error was measured as the root mean squared error between the original spectra and the spectra reconstructed from the extracted features; a lower value indicated that the feature extraction process more completely preserved the original spectral information.

3. Results and Discussion

3.1. Statistics of Tannin Content

The tannin content of the 305 sorghum samples was determined, and the results are shown in Figure 1. The tannin content ranged from 0.46% to 3.27%, with a mean value of 1.75% and a standard deviation of 0.41%, covering a wide range from low to high levels, thereby providing a sound sample basis for subsequent quantitative modeling.

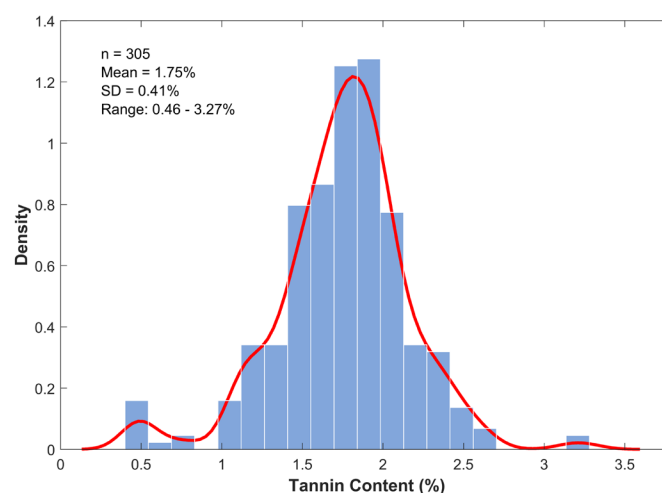


Figure 1. Distribution of tannin content in 305 sorghum samples.

3.2. Sorghum Near-Infrared Spectroscopy

The raw near-infrared (NIR) spectra of sorghum grains are presented in Figure 2. The main absorption peaks are located at approximately 10,070, 8293, 6797, 5770, 5177, 4715, 4315, and 4245 cm^{-1} . Among these, the peak at 10,070 cm^{-1} is attributed to the second overtone of phenolic O–H stretching vibration; the peak at 8293 cm^{-1} is associated with the second overtones of methyl and methylene C–H stretching vibrations; the peak at 6797 cm^{-1} corresponds to the first overtone of hydrogen-bonded O–H stretching in alcohols or phenols; the peak at 5770 cm^{-1} is related to the first overtones of antisymmetric and symmetric C–H stretching of methyl groups attached to aromatic rings; and the peaks at 5177, 4715, 4315, and 4245 cm^{-1} are all combination bands involving aromatic structures, corresponding respectively to C–H stretching combination, cooperative C–C and C–H combination, and alkyl–aromatic C–H coupling combination [23].

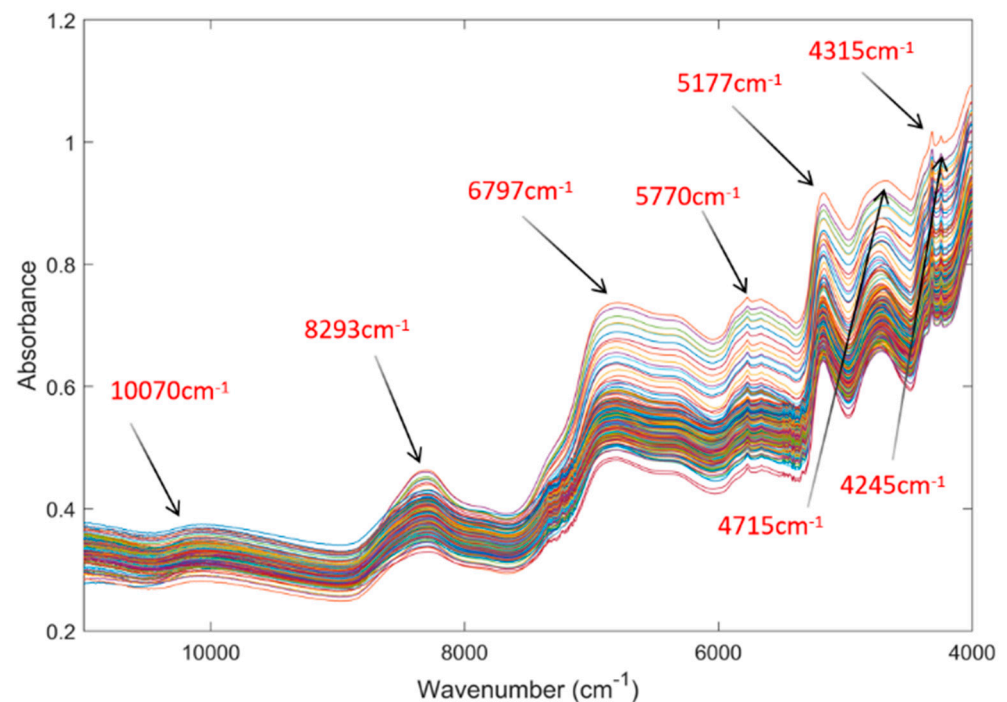


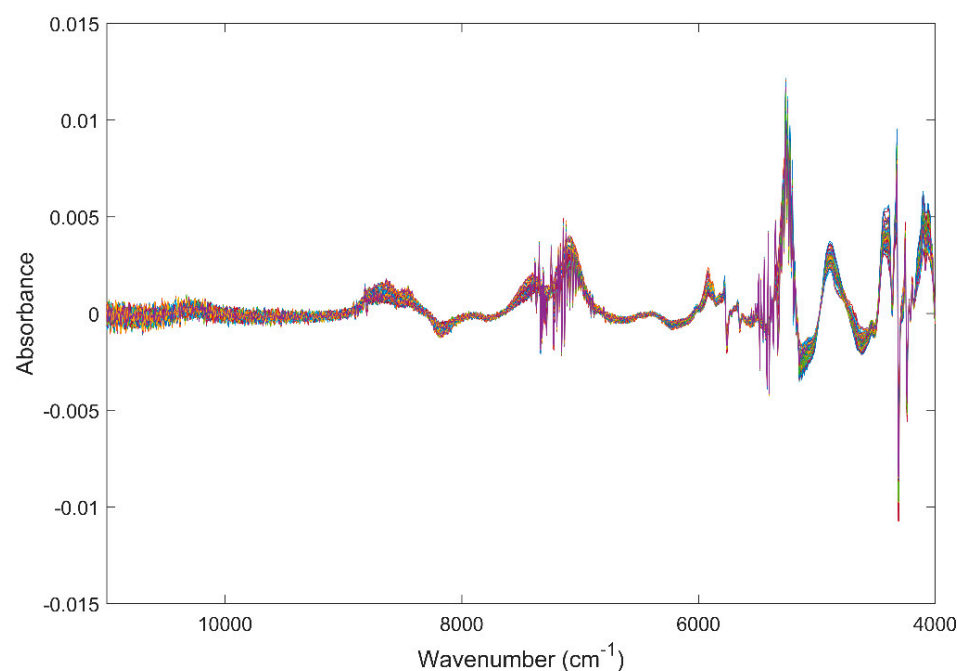
Figure 2. Raw NIR spectra of sorghum grains.

Due to variations in sample particle size and packing, the raw spectra exhibited marked baseline drift and tilt. To correct these physical interferences, which are unrelated to chemical composition, Savitzky–Golay (SG) smoothing combined with a first-order derivative was applied for spectral preprocessing. The first-order derivative eliminated constant baseline offsets and partially corrected linear tilts, while also resolving adjacent shoulders near 5408 and 5177 cm^{-1} . SG smoothing was employed to suppress the amplification of high-frequency noise inherent in the derivative operation [24].

To evaluate the effect of the SG smoothing window size on spectral fidelity, the peak attenuation of two characteristic peaks at 5408 and 6797 cm^{-1} was monitored across different window sizes for optimal window selection. The peak at 5408 cm^{-1} corresponds to the methyl C–H stretching vibration of the aromatic ring, exhibiting a narrow shape that represents fine spectral features. The peak at 6797 cm^{-1} , assigned to the first overtone of O–H stretching, displays a broader profile that captures the dominant spectral intensity. As shown in Table 1, increasing the window size from 3 to 7 resulted in a 25.6% attenuation rate for the 5408 cm^{-1} peak, compared to only 4.2% for the 6797 cm^{-1} peak, indicating that narrow spectral peaks (0.34) are considerably more sensitive to smoothing than broader peaks, and that excessively large window sizes (0.55) can obscure fine spectral features and lead to the loss of critical structural information [25]. The attenuation rate reflects the degree of peak intensity reduction with increasing smoothing, and a smaller value indicates more complete signal retention. To prioritize the preservation of narrow peak information, an SG smoothing window size of 3 was selected. The preprocessed spectrum (Figure 3) demonstrates significant baseline correction; however, due to the high-pass characteristics of the derivative operation, high-frequency noise remains visible in certain local regions.

Table 1. Attenuation of first-derivative peak intensity of characteristic peaks with different smoothing window sizes.

Absorption Peaks (cm ^{−1})	Peak Height (×10 ^{−3}) Window Size 3	Attenuation Rate (%) Window Size 7	Attenuation Rate (%) Window Size 11
5408	0.34	25.6	31.2
6797	0.55	4.2	8.4

**Figure 3.** Preprocessed near-infrared spectra of sorghum grains.

3.3. Spectral Data Compression Based on Wavelet Transform

After preprocessing, high-frequency noise remains in the spectra (Figure 3), and direct modeling with the 1845-dimensional data leads to feature redundancy. To suppress such noise and reduce the data dimensionality, discrete wavelet transform (DWT) is used for denoising and dimensionality reduction. Due to its excellent time–frequency localization characteristics, wavelet transform has been widely used in NIR spectroscopy for denoising and data compression [26]. In this study, the db3 wavelet basis was employed for a four-level decomposition, which separates the signal into high-frequency and low-frequency components. The high-frequency components, mainly noise, were discarded, while the low-frequency components, which carry the main spectral profile, were retained for subsequent processing. The reconstructed spectrum using the low-frequency components (Figure 4a) shows a high degree of similarity with the preprocessed spectrum (Figure 3), with a reconstruction error of only 3.52% of the dynamic range. After wavelet transformation, the noise was effectively suppressed, the spectral curve became smoother, and the main spectral profile was preserved, indicating that the effective information in the spectrum was more prominent [27]. Meanwhile, the data dimensionality was reduced from 1845 to 120 (Figure 4b), providing a low-noise and low-dimensional input for subsequent analysis.

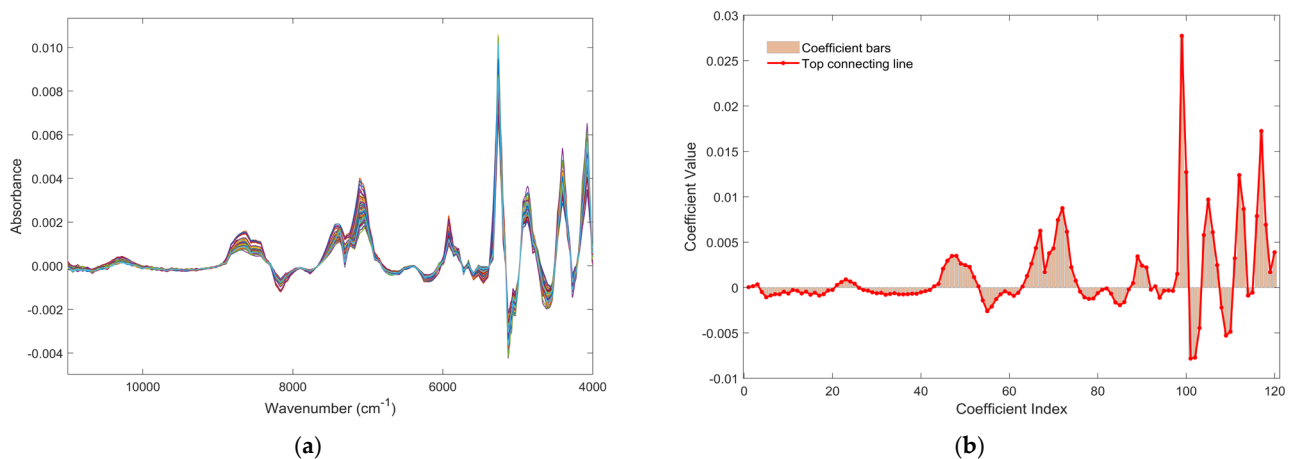


Figure 4. Results of wavelet transform denoising and dimensionality reduction: (a) spectra of all training samples reconstructed after wavelet denoising; (b) 120 approximation coefficients for a single sample, reduced from 1845 to 120 dimensions.

3.4. Model Development

3.4.1. Feature Extraction Using KICA

In this study, KICA was used to extract features from the preprocessed and wavelet-compressed spectra. By mapping the data into a high-dimensional space via the kernel trick and subsequently performing ICA, 11 independent components (ICs) were obtained from the training set, whereby the feature dimensionality was reduced from 120 to 11. The ICA in this study was based on a single random split to obtain clearly interpretable independent components for spectral interpretation.

To assess the importance of each independent component, the mean absolute coefficient of each component across all training samples was calculated (Figure 5a). This coefficient reflects the overall activity level, or contribution, of the component among the samples [28]. Of the 11 components, IC9 and IC6 showed the two highest contributions, with mean absolute coefficients of 0.847 and 0.815, respectively, markedly surpassing those of the remaining components. These two components were therefore selected as axes, and the projection score of each sample onto them was computed. The scores for the training and test sets were then plotted in a two-dimensional scatter plot (Figure 5b). From the plot, it is evident that the sample points from the two sets substantially overlap and exhibit a consistent distribution pattern. The Mann–Whitney U test was further applied to compare the score distributions, yielding *p*-values of 0.840 and 0.563 for IC9 and IC6, respectively. Both are well above 0.05, indicating no significant difference. These results demonstrate that the features extracted by KICA maintain a consistent distribution pattern across the training and test sets with good generalization capability, suggesting that the method can effectively isolate the nonlinearly mixed redundant information in the spectral data and that the reduced-dimension features remain stable and reliable, thereby offering practical value for subsequent modeling.

To further elucidate the physicochemical significance of the extracted independent components, an analysis was performed on the two components with the highest contribution, IC9 and IC6. The correlation between each independent component and wavenumber was plotted (Figure 6a). As shown in Figure 6a, the two curves exhibit markedly different characteristics: IC9 displays multiple sharp, high-intensity absorption peaks in the NIR spectral region of 4000–7000 cm⁻¹, and its overall profile closely resembles that of the preprocessed sorghum NIR spectra. In contrast, IC6 presents a substantially smoother profile, characterized by weak baseline fluctuations across the entire spectral range without any prominent sharp peaks.

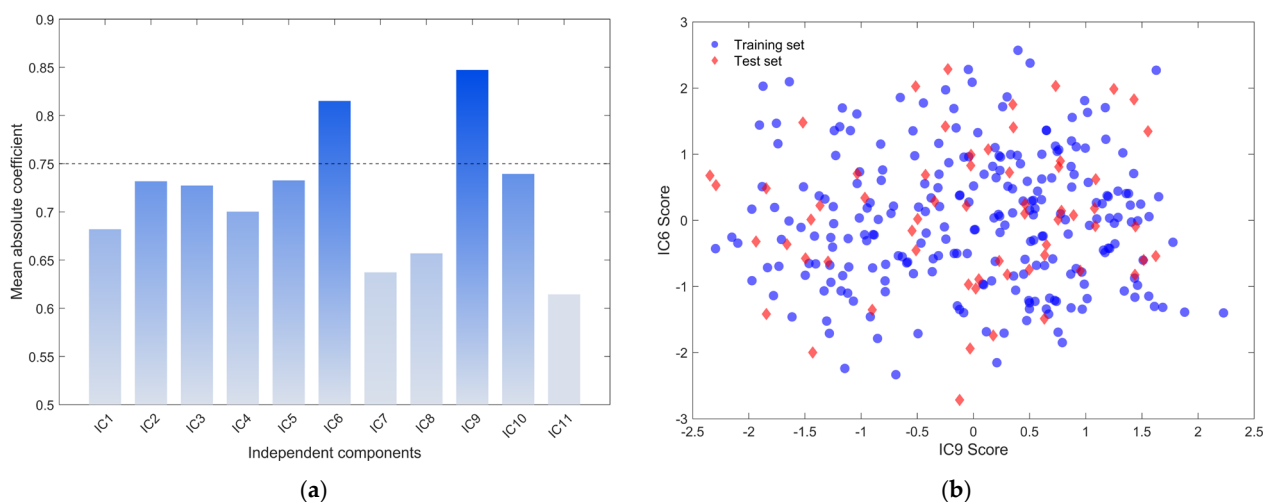


Figure 5. Results of KICA independent component analysis: (a) mean contribution of 11 independent components; (b) 2D score plot based on KICA for training and test sets.

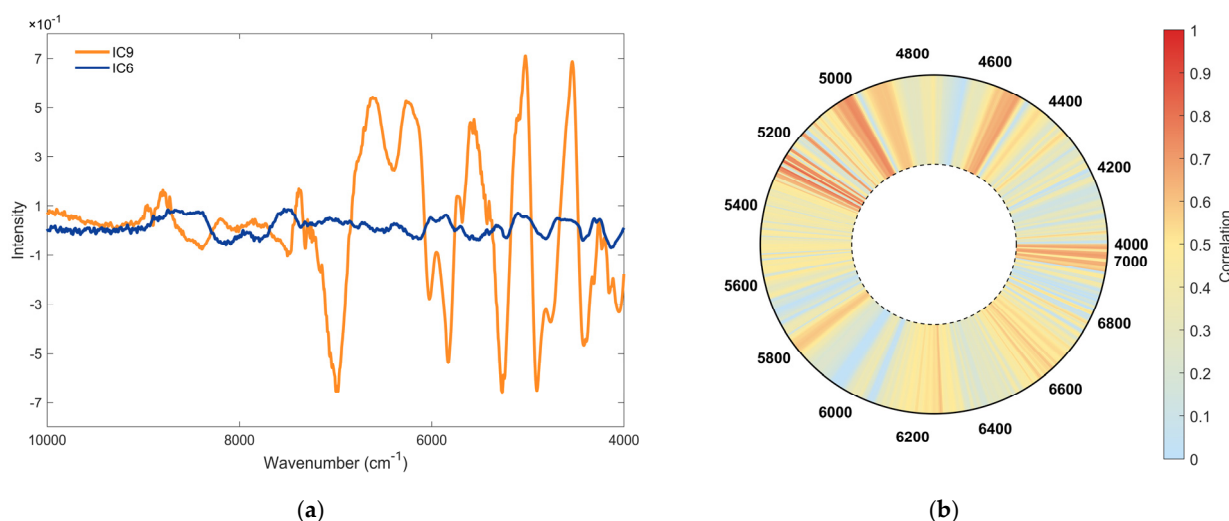


Figure 6. Spectral interpretation of the independent components: (a) correlation of IC9 and IC6 with wavenumber; (b) circular correlation heatmap of IC9.

Based on the peak assignments of the raw NIR spectra of sorghum grains, the 4000–7000 cm^{-1} region represents the core characteristic response range of tannin-related substances. This region encompasses key absorption features, including phenolic O–H stretching vibrations, aromatic C–H stretching vibrations, aromatic C–C skeletal vibrations, and various combination and overtone bands, and serves as a crucial spectral interval for characterizing the structure and content of major chemical constituents in sorghum. To further investigate the intrinsic relationship between the independent components and spectral features, and to pinpoint the specific distribution of tannin-sensitive wavelength bands, IC9, which exhibited the highest contribution, was selected for in-depth analysis. Based on the distribution of the correlation coefficients between each spectral variable and the IC9 score, a circular correlation heatmap of IC9 was generated (Figure 6b). The heatmap clearly shows that key wavelength bands around 4000, 4500, 5000, 5200, and 7000 cm^{-1} exhibit strong positive correlations, with colors that are highly saturated red. This indicates that the prominent characteristic peaks of IC9 are concentrated precisely within the common spectral regions of tannins, aligning closely with the distribution of tannin characteristic absorption peaks in the raw NIR spectra of sorghum grains [29]. This suggests that IC9 is closely linked to the structural information of target substances in sorghum and

can effectively capture the spectral signals of core functional groups potentially related to tannins, such as aromatic rings and phenolic hydroxyl groups. Consequently, IC9 can be regarded as a characteristic source signal highly correlated with variations in tannin content. Additionally, minor shifts are observed between the characteristic peak positions of IC9 and the measured peaks in the raw spectra. This phenomenon is not an anomaly but rather a result of KICA, which, through blind source separation, effectively removes the redundant overlapping interference between different spectral components and noise. This process purifies and reconstructs the mixed spectra, bringing the reconstructed characteristic peaks closer to the true absorption structures likely associated with tannin-related substances [30].

In contrast, as shown in Figure 6a, IC6 does not exhibit any pronounced characteristic response within the tannin-specific spectral region; its smooth profile mainly reflects background information or nonspecific spectral variations. The marked difference between IC9 and IC6 in this characteristic region provides direct evidence that KICA has successfully separated statistically independent source signals from the mixed spectra. As the independent component with the higher contribution, IC9 is closely associated with tannin-related substances in terms of characteristic peak positions and functional group responses, thereby carrying key chemical information strongly linked to tannins. IC6, on the other hand, carries background interferences or other non-target information. The two components are mutually non-redundant and informationally complementary, which directly echoes the core principle of blind source separation underlying KICA.

In this study, KICA extracted a total of 11 independent components, each with a contribution exceeding 0.5 (Figure 5a), indicating that all components carry meaningful spectral information without significant redundancy. The number of retained independent components was determined using the threshold of a mean absolute coefficient greater than 0.5, and all components meeting this threshold were retained for subsequent modeling. Because these components are statistically independent and correspond to different functional groups or compositional characteristics, selecting only those with the highest contributions could risk losing valuable information and thereby compromise model accuracy. Therefore, all 11 independent components were used as input variables for the subsequent quantitative prediction modeling of sorghum tannin content.

3.4.2. Model Comparison

Before applying feature extraction methods, this study first compared the performance of three regression models—principal component regression (PCR), partial least squares regression (PLSR), and support vector regression (SVR)—using the wavelet-compressed spectra without further feature extraction. The modeling in this section was based on the same single random split described in the preceding section. As shown in Table 2, the regression performance of the three methods ranked from best to worst as SVR > PLSR > PCR: PCR constructs its model solely from the directions of maximal spectral variance, resulting in weak predictive performance [31]; PLSR substantially improves prediction by maximizing the covariance between the spectra and the target analyte; and SVR further enhances model performance by leveraging its kernel function to capture the inherent nonlinear relationships within the spectral data. The scatter plots of predicted versus reference values for the training and test sets (Figure 7a–c) provide an intuitive visual comparison: the PCR model exhibits a high degree of dispersion with poor overall fit, the PLSR model noticeably reduces prediction error and improves data clustering, and the SVR model yields the most tightly clustered predictions with the smallest overall deviation. This superiority can be attributed to SVR's strong nonlinear modeling capacity, which enables it to better capture the complex nonlinear information latent in NIR spectral data [28]. Accordingly, SVR was selected as the baseline regression model for the subsequent comparison of different

feature extraction methods, consistent with the overarching aim of this study to analyze the nonlinear structure of spectral data.

Table 2. Comparison of modeling strategies for sorghum tannin content prediction.

Model	Number of Features	Training Set		Test Set		
		R ²	RMSE	R ²	RMSE	RPD
None-PCR	120	0.5867	0.2471	0.4875	0.3502	1.41
None-PLSR	120	0.8690	0.1382	0.8159	0.2132	2.35
None-SVR	120	0.9796	0.0549	0.8726	0.1746	2.83
PCA-SVR	14	0.9190	0.1094	0.8731	0.1743	2.83
ICA-SVR	10	0.9641	0.0728	0.9138	0.1436	3.43
KICA-SVR	11	0.9614	0.0755	0.9248	0.1341	3.68

(Note: Hyperparameters of all models were optimized on the training set via five-fold cross-validation. PCA-SVR: kernel scale = 12.191 (searched [8, 15]), box constraint = 2.568 (searched [1, 5]), ϵ = 0.0278 (searched [0.01, 0.1]); ICA-SVR: kernel scale = 3.72 (searched [1, 6]), box constraint = 5.51 (searched [1, 10]), ϵ = 0.035 (searched [0.01, 0.1]); KICA-SVR: kernel scale = 5.0 (searched among {0.5, 1, 2, 3, 4, 5}), box constraint = 10.0 (searched among {0.1, 1, 10, 100, 500}), ϵ = 0.05 (searched among {0.01, 0.05, 0.1, 0.2, 0.5}), kernel width σ = 0.75 (searched 0.10–2.00, step 0.05); SVR (without feature extraction): C = 0.8 (searched [0.3, 2.0]), γ = 0.007 (searched [0.003, 0.01]), ϵ = 0.05 (searched [0.03, 0.1]). The search ranges were determined independently for each method based on the dimensionality of their input features and their respective parameter sensitivity.)

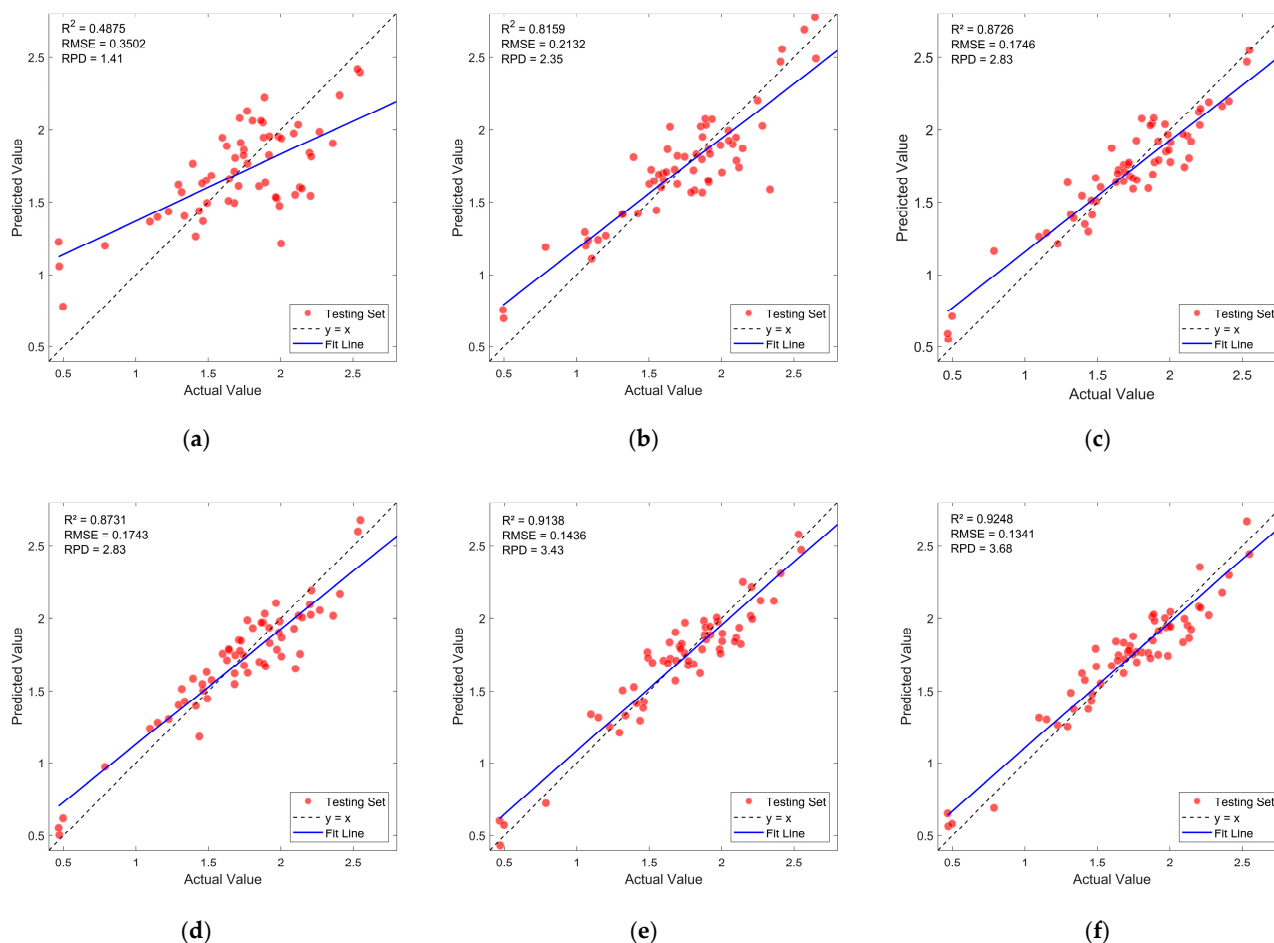


Figure 7. Scatter plots of predicted versus reference values for different modeling strategies: (a) PCR model without feature extraction; (b) PLSR model without feature extraction; (c) SVR model without feature extraction; (d) PCA-SVR model; (e) ICA-SVR model; (f) KICA-SVR model.

(1) Model Performance Comparison

To investigate the effect of different feature extraction methods on the performance of the SVR model, this study established SVR models based on three feature extraction

methods—PCA, ICA, and KICA—with the SVR model without feature extraction serving as the baseline for comparison. The key performance indicators of all models are summarized in Table 2.

As shown in Table 2, when modeling used only the 120 wavelet-compressed features, the SVR model achieved high fitting accuracy on the training set ($R^2 = 0.9796$) but an R^2 of 0.8726 on the test set, revealing a noticeable generalization gap. NIR spectral variables are inherently multicollinear, and high-dimensional redundant features tend to introduce noise into the model. The scatter plot in Figure 7c further illustrates this limitation: the predicted values still deviate noticeably from the reference values, with the sample points being relatively dispersed. Moreover, the error range remains fairly wide, indicating that the consistency between predicted and reference values could still be improved. These observations suggest that relying solely on wavelet-compressed features for modeling has inherent limitations, underscoring the necessity of incorporating feature extraction to refine the model input.

PCA is a commonly used linear dimensionality reduction method in spectral analysis, which extracts orthogonal principal components based on the principle of maximizing variance. After comparing different cumulative contribution rate thresholds, 14 principal components were retained for modeling, corresponding to a cumulative contribution rate of 90%. Following PCA transformation, the test set R^2 reached 0.8731 with an RMSE of 0.1743, indicating improved model stability. As shown in Figure 7d, the predicted values exhibit tighter clustering around the reference values, and the overall prediction error becomes more stabilized, suggesting that PCA reduces the multicollinearity of the spectral data to a certain degree. However, because the principal components extracted by PCA are linear combinations of the original variables, the method has inherent difficulty in effectively separating informative signals from interfering components. Consequently, some redundant information remains in the model, limiting further improvement in prediction accuracy.

To further separate informative signals from interfering components, ICA was employed for feature extraction. Unlike PCA, the components extracted by ICA are statistically independent from one another. In this study, 10 independent components were extracted for modeling, raising the test set R^2 to 0.9138, reducing the RMSE to 0.1436, and yielding an RPD of 3.43. At this stage, the model outperformed all preceding methods on the test set. As shown in Figure 7e, the predicted values align more closely with the reference values and the data points are more tightly clustered, indicating that the correspondence between predicted and reference values has become more consistent and that the prediction error has stabilized. These improvements suggest that ICA enhances the extraction of effective spectral information to a certain degree while attenuating noise interference. Notably, compared with the PLS-based sorghum NIR model reported by Yu et al. [29] ($R^2 = 0.8841$), the proposed KICA–SVR framework achieved a higher prediction accuracy of $R^2 = 0.9248$, suggesting that nonlinear feature extraction via KICA offers a modest improvement over conventional linear modeling for tannin quantification in sorghum.

Although ICA outperforms PCA in terms of predictive performance, it remains fundamentally a linear method. In NIR spectroscopy, the relationship between the spectra and the target analyte content often exhibits a degree of nonlinearity, which may arise from factors such as light scattering effects, matrix differences, or molecular interactions—factors that are inherently insufficiently fully captured by linear methods. To address this, KICA was further employed for feature extraction. By expanding the data space via the kernel trick, KICA extracted 11 nonlinear independent components for modeling. Under this approach, the model performance was further enhanced, achieving a test set R^2 of 0.9248, an RMSE of 0.1341, and an RPD of 3.68. As shown in Figure 7f, the predicted values align even more closely with the reference values and the sample points are more tightly

clustered. Moreover, the deviations between predicted and reference values are smaller and the error range is markedly narrowed. These results indicate that KICA can, to a certain extent, capture the nonlinear information embedded in the spectra, thereby enhancing the model's predictive capability. This also confirms that KICA represents the optimal feature extraction strategy in this study for adapting to spectral data and improving the quantitative prediction of target components.

(2) Comparison of Feature Extraction Quality

In evaluating feature extraction quality, reconstruction error serves as an intuitive and effective comparative metric [32]. After the original high-dimensional data are transformed into low-dimensional features, a reconstruction model is employed to map these features back to the original data space, and the discrepancy between the reconstructed and original data is then calculated. If the dimensionality reduction process preserves the essential information of the data, the reconstruction error should be small; conversely, if substantial information is lost, the reconstruction error will be large. By comparing the reconstruction errors of different dimensionality reduction methods, their respective capacities to retain the original spectral information can be objectively assessed, thereby providing a basis for comparing feature extraction quality.

The reconstruction strategy varies with the dimensionality reduction algorithm. For PCA, the dimensionality reduction is a linear projection, and reconstruction is directly accomplished via the inverse transform of the loading matrix. For ICA, which is also linear, the original features can likewise be reconstructed from the independent components through a pseudo-inverse operation. For KICA, however, the dimensionality reduction process is nonlinear, making direct inversion difficult; therefore, a linear regression model was adopted as a bridge for reconstruction. It should be noted that PCA and ICA possess exact linear inverse transforms, whereas KICA, owing to the nonlinear nature of its dimensionality reduction, requires a linear regression model for reconstruction, and its reconstruction error therefore includes an additional approximation error. Consequently, the reconstruction advantage of KICA may be underestimated, and its true information retention capability may be better than the reported values. This linear reconstruction approach places all three methods under a common benchmark for comparison, while also avoiding the overfitting risk associated with nonlinear reconstruction. The reconstruction error was uniformly quantified using the root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{1}{N \times D} \sum_{i=1}^N \sum_{j=1}^D (x_{ij} - \hat{x}_{ij})^2}$$

where x_{ij} and $\hat{x}_{ij}$ represent the j -th original and reconstructed compressed feature value of the i -th sample, respectively; N is the number of samples; and D is the dimensionality of the compressed features.

All three methods performed dimensionality reduction on the same set of wavelet-compressed features. Based on the model performance optimization described earlier, PCA retained 14 principal components, ICA extracted 10 independent components, and KICA extracted 11 independent components; their reduced dimensionalities were therefore comparable (10–14 dimensions). The reconstruction errors shown in Table 3 indicate that PCA produced the largest reconstruction error, followed by ICA, whereas KICA yielded the smallest. The fundamental reason for this discrepancy lies in the distinct statistical information exploited by each method. PCA relies exclusively on second-order statistics (the covariance matrix) and is unable to capture the nonlinear structures inherent in the spectral data; consequently, its reconstruction accuracy remains limited even with a relatively large number of principal components. ICA, in contrast, exploits higher-order

statistical information by maximizing non-Gaussianity, and its independent components can represent the original signals more intrinsically, resulting in better reconstruction than PCA at a comparable dimensionality. KICA further incorporates the kernel trick, extracting independent components in an implicit high-dimensional feature space and thereby more fully leveraging the nonlinear information embedded in the spectra. This conclusion aligns with the regression modeling results presented earlier: among the three approaches, the KICA-based SVR model achieved the highest predictive accuracy, followed by the ICA-based model, with the PCA-based model being the lowest, further validating the advantages of higher-order statistics and kernel methods in spectral feature extraction.

Table 3. Reconstruction error comparison of different feature extraction methods.

Feature Extraction Method	Number of Features	Training Set	Test Set
		Reconstruction RMSE	Reconstruction RMSE
PCA	14	0.000094	0.000101
ICA	10	0.000066	0.000074
KICA	11	0.000062	0.000069

(3) Model Stability Analysis

To avoid the contingency introduced by a single data split, 20 repeated experiments were conducted based on the original data partition, each with an independently executed complete modeling pipeline. Figure 8 presents the distributions of R^2 and RMSE on the training and test sets for each model across the 20 replicates.

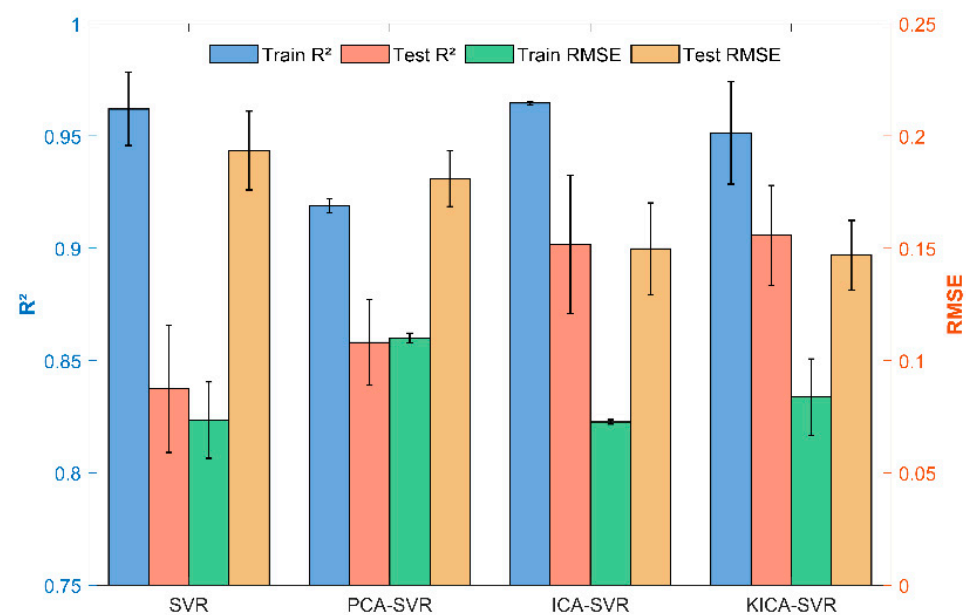


Figure 8. Comparison of R^2 and RMSE on the training and test sets for each model over 20 repeated experiments.

On the test set, KICA–SVR achieved a mean R^2 of 0.9057, higher than those of ICA–SVR (0.9017), PCA–SVR (0.8581), and SVR (0.8374), indicating that KICA attained the best average prediction accuracy. In terms of variability, the standard deviation of R^2 for KICA was 0.0222, lower than that of ICA (0.0307) but slightly higher than that of PCA (0.0191), suggesting that KICA not only delivered the best average performance across repeated runs but also exhibited relatively small fluctuations, reflecting favorable consistency in predictive performance. A similar trend was observed for RMSE: the mean test-set RMSE

of KICA was 0.1469, lower than those of ICA (0.1498), PCA (0.1810), and SVR (0.1935). Paired *t*-test results showed that KICA–SVR significantly outperformed PCA–SVR and SVR ($p < 0.001$), whereas no significant difference was found between KICA–SVR and ICA–SVR ($p = 0.6571$). In certain splits, ICA even slightly surpassed KICA, which may be related to the preprocessing pipeline—after first-order derivative preprocessing and wavelet transform, some nonlinear effects in the original spectra were attenuated, enabling linear ICA to achieve performance comparable to that of KICA in those splits.

On the training set, SVR and ICA yielded the highest mean R^2 values (0.9620 and 0.9646, respectively), while KICA (0.9514) was slightly lower, although the difference was marginal. Examining the gap between training-set and test-set R^2 , SVR showed the largest discrepancy (0.1246), indicating a pronounced overfitting problem, which can be attributed to SVR modeling directly on high-dimensional raw features without effective dimensionality reduction. The R^2 gaps for PCA and ICA were 0.0610 and 0.0629, respectively, whereas KICA exhibited the smallest gap among the four models (0.0457), suggesting the strongest generalization capability. With regard to error variability, the standard deviation of the training-set RMSE for KICA (0.0170) was markedly larger than those of PCA (0.0022) and ICA (0.0010), indicating that KICA was more sensitive to the composition of the training set—variations in individual training samples may propagate through the kernel matrix and affect the structure of the entire high-dimensional feature space, thereby amplifying the fluctuation of training error. On the test set, however, the standard deviation of RMSE for KICA (0.0155) was actually lower than that of ICA (0.0204), demonstrating that while KICA was sensitive to sample variation on the training set, its generalization stability on unseen samples was not compromised. These results further validate the combined advantages of KICA–SVR in terms of prediction accuracy and generalization stability.

(4) Validation of KICA Applicability

To validate the applicability of the proposed method, a generalization experiment was conducted on the publicly available corn NIR dataset from Eigenvector (<https://eigenvector.com/resources/data-sets/>, accessed on 25 March 2026). Specifically, spectra acquired with the M5 instrument were used, covering the wavelength range of 1100–2498 nm with 700 wavelength points, and fat and protein contents were set as the prediction targets. The identical analytical pipeline as that applied to the sorghum data was adopted, namely wavelet compression, feature extraction, and SVR modeling. The predictive performance of three feature extraction methods—PCA, ICA, and KICA—was compared, and the results are presented in Table 4. As shown in the table, for both constituents, the KICA-based model outperformed the PCA- and ICA-based models on the test set. These results are consistent with the findings obtained for the sorghum tannin data in this study.

Table 4. Comparison of prediction performance of different feature extraction methods on the corn dataset.

Target	Model	Training Set		Test Set		
		R^2	RMSE	R^2	RMSE	RPD
Protein	PCA-SVR	0.9658	0.0874	0.9030	0.1684	3.32
	ICA-SVR	0.9382	0.1234	0.9223	0.1355	3.71
	KICA-SVR	0.9873	0.0546	0.9311	0.1193	3.94
Fat	PCA-SVR	0.9043	0.0511	0.8571	0.0788	2.73
	ICA-SVR	0.9032	0.0560	0.8715	0.0569	2.88
	KICA-SVR	0.9247	0.0488	0.9525	0.0361	4.74

For sorghum tannin, corn protein, and corn fat, KICA consistently achieved the best quantitative prediction performance on two independent datasets, fully demonstrating

its effectiveness and generalization capability in NIR quantitative analysis. These results further indicate that KICA holds considerable potential for application in agricultural NIR testing, particularly for quality assessment of samples with complex matrices. By adopting a blind source separation framework and leveraging the kernel trick to handle nonlinear mixing effects, KICA can more fully exploit the latent features embedded in the spectra and effectively separate informative signals from interferences, thereby substantially improving both the accuracy and robustness of the prediction model.

These results indicate that KICA achieved superior prediction performance across two independent datasets and three different constituents, which can be closely attributed to its feature extraction mechanism. The improvement in KICA over linear methods in both prediction accuracy and reconstruction quality stems from its combined use of the kernel trick and blind source separation. Unlike PCA, which relies solely on second-order statistics, KICA maximizes non-Gaussianity in an implicit high-dimensional feature space and is capable of separating nonlinearly mixed source signals, thereby more fully extracting effective information related to the target constituents from the spectra. Compared with ICA, which can only handle linear mixtures, KICA maps the original data into a high-dimensional space via a kernel function and performs linear ICA in that space, which is equivalent to constructing a feature representation that adapts to the nonlinear structure of the data. Consequently, when confronting nonlinear spectral distortions caused by factors such as light scattering and particle size differences, KICA can more effectively capture the underlying nonlinear information, reducing the information loss induced by linear assumptions.

4. Conclusions

This study introduced kernel-independent component analysis (KICA) into NIR spectral modeling and combined it with support vector regression (SVR) to achieve accurate prediction of tannin content in sorghum grains. The results demonstrated that KICA–SVR outperformed PCA and ICA in terms of prediction accuracy, and its stability was validated through 20 repeated experiments ($R^2 = 0.906 \pm 0.022$). Reconstruction error analysis indicated that KICA can more completely preserve the nonlinear informative content of the original spectra, while wavelength correlation analysis enhanced the interpretability of the model. Generalization experiments on a corn dataset further validated the applicability of the proposed method.

The application of KICA in NIR quantitative analysis of cereal grains remains relatively underexplored compared with linear methods, and this study provides a systematic evaluation of its potential in this context. The main limitations of KICA are its sensitivity of kernel parameters to the composition of the training set and its relatively high computational cost, both of which can be effectively managed through cross-validation. Future work may incorporate prior physicochemical information on the samples, leverage the inherent advantages of KICA as a blind source separation technique to further investigate the correspondence between independent components and specific chemical constituents, and validate the method in more complex real-world detection scenarios. Overall, KICA can serve as a promising approach for rapid NIR-based quality assessment of agricultural products with complex matrices.

Author Contributions: W.-P.L.: conceptualization, investigation, writing—review and editing; Y.H.: formal analysis, software, validation, visualization, writing—original draft; Y.W.: funding acquisition; Z.-G.C.: methodology, supervision, resources; B.L.: data curation, project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by the National Natural Science Foundation of China (No. 32472000).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhao, W.; Liu, Y.; Zhou, K.; Li, K.; Shen, J. How sorghum grain composition affects the quality of Chinese baijiu—a comprehensive review. *J. Food Compos. Anal.* **2024**, *134*, 106512.
2. Shi, L.; Du, Y.; Nie, J.; Tan, H.; Li, S.; Li, T. Relationship between the Sorghum's Physicochemical Attributes and the Flavor Components of Qing Stubble Daqu Baijiu. *Sci. Technol. Food Ind.* **2025**, *46*, 231–238.
3. Pan, L.; Feng, S.; Li, W.; Zhu, W. Sorghum tannin extract impedes in vitro digestibility and fermentability of nutrients in the simulated porcine gastrointestinal tract. *J. Anim. Sci.* **2023**, *101*, skad126. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Palacios, C.E.; Nagai, A.; Torres, P.; Rodrigues, J.A.; Salatino, A. Contents of tannins of cultivars of sorghum cultivated in Brazil, as determined by four quantification methods. *Food Chem.* **2021**, *337*, 127970. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Beć, K.B.; Grabska, J.; Huck, C.W. Interpretability in near-infrared (NIR) spectroscopy: Current pathways to the long-standing challenge. *TrAC Trends Anal. Chem.* **2025**, *189*, 118254.
6. Zhu, J.; Ji, G.; Chen, B.; Yan, B.; Ren, F.; Li, N.; Zhu, X.; He, S.; Mu, Z.; Liu, H. High-throughput near-infrared spectroscopy for detection of major components and quality grading of peas. *Front. Nutr.* **2024**, *11*, 1505407. [\[PubMed\]](#)
7. Chakraborty, S. Bayesian multiple response kernel regression model for high dimensional data and its practical applications in near infrared spectroscopy. *Comput. Stat. Data Anal.* **2012**, *56*, 2742–2755. [\[CrossRef\]](#)
8. Xiao, X.U.; Lijuan, X.I.E.; Yibin, Y. Factors influencing near infrared spectroscopy analysis of agro-products: A review. *Front. Agric. Sci. Eng.* **2019**, *6*, 105. [\[CrossRef\]](#)
9. Han, J.; Guo, J.; Zhang, Z.; Yang, X.; Shi, Y.; Zhou, J. The Rapid Detection of Trash Content in Seed Cotton Using Near-Infrared Spectroscopy Combined with Characteristic Wavelength Selection. *Agriculture* **2023**, *13*, 1928. [\[CrossRef\]](#)
10. Abdi, H.; Williams, L.J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459. [\[CrossRef\]](#)
11. Cruz-Tirado, J.P.; Oliveira, M.; de Jesus Filho, M.; Godoy, H.T.; Amigo, J.M.; Barbin, D.F. Shelf life estimation and kinetic degradation modeling of chia seeds (*Salvia hispanica*) using principal component analysis based on NIR-hyperspectral imaging. *Food Control* **2020**, *1423*, 107777. [\[CrossRef\]](#)
12. Hyvärinen, A.; Oja, E. Independent component analysis: Algorithms and applications. *Neural Netw.* **2000**, *13*, 411–430. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Pati, R.; Pujari, A.K.; Gahan, P.; Kumar, V. Independent component analysis: A review with emphasis on commonly used algorithms and contrast function. *Comput. Sist.* **2021**, *25*, 97–115. [\[CrossRef\]](#)
14. Sun, D.; Gong, X.; Chen, Y. Integrating Canonical Variate Analysis and Kernel Independent Component Analysis for Tennessee Eastman Process Monitoring. *J. Chem. Eng. Jpn.* **2020**, *53*, 126–133. [\[CrossRef\]](#)
15. Huang, J.-Z.; Liu, Y.-J.; Qi, H.; Chen, S.-X.; Li, D.-S. A novel nonlinear ICA method for structural health monitoring under nonlinear environmental or operational influences. *Structures* **2025**, *71*, 108118.
16. Diallo, A.R.; Homri, L.; Dantan, J.-Y.; Bonnet, F.; Boeuf, T. Enhancing Fault Diagnosis in Process Industries with Internal Variables of Model Predictive Control. *IFAC-PapersOnLine* **2024**, *58*, 538–543. [\[CrossRef\]](#)
17. Thiex, N.; Richardson, C. Challenges in measuring moisture content of feeds. *J. Anim. Sci.* **2003**, *81*, 3255–3266. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Rossi, G.B.; Lozano, V.A.; Olivieri, A.C. Spectral pre-processing and non-linear calibration with convolutional kernel partial least-squares. Teaching new tricks to an old dog. *Chemom. Intell. Lab. Syst.* **2023**, *233*, 104736.
19. Maywald, D.; Vimalajeewa, D. Utilizing Wavelet Transform in the Analysis of Scaling Dynamics for Milk Quality Evaluation. In *Time Series and Wavelet Analysis: Festschrift in Honor of Pedro A. Morettin*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 261–279.
20. Weidong, Z. Removal of noise and ECG artifact from EEG based on wavelet transform and independent component analysis. *J. Shandong Univ. Health Sci.* **2003**, *2003*, 116–119.
21. Wang, X.; Han, X.; Li, L.; Zheng, X. Optimization for quantification of sorghum tannins by Ferric ammonium citrate assay. *Grain Oil Sci. Technol.* **2020**, *3*, 146–153. [\[CrossRef\]](#)
22. Simeone, M.L.F.; Pimentel, M.A.G.; Queiroz, V.A.V.; Santos, F.; Brito, A.; Aquino, L.F.M.; da C Filho, J.C.E.; de Menezes, C.B.; Paes, M.C.D.; Tibola, C.S.; et al. Portable near-infrared (NIR) spectroscopy and multivariate calibration for reliable quality control of maize and sorghum grain chemical composition. *J. Food Compos. Anal.* **2024**, *134*, 106502. [\[CrossRef\]](#)

23. Workman, J., Jr.; Weyer, L. *Practical Guide to Interpretive Near-Infrared Spectroscopy*; CRC Press: Boca Raton, FL, USA, 2007.
24. Fan, C.; Liu, Y.; Cui, T.; Qiao, M.; Yu, Y.; Xie, W.; Huang, Y. Quantitative Prediction of Protein Content in Corn Kernel Based on Near-Infrared Spectroscopy. *Foods* **2024**, *13*, 4173. [[CrossRef](#)] [[PubMed](#)]
25. Zimmermann, B.; Kohler, A. Optimizing Savitzky–Golay Parameters for Improving Spectral Resolution and Quantification in Infrared Spectroscopy. *Appl. Spectrosc.* **2013**, *67*, 892–902. [[CrossRef](#)] [[PubMed](#)]
26. Fu, C.; Cao, S.; Tian, A. Estimation of soil chromium content using visible and near-infrared spectroscopy coupled with discrete wavelet transform and long short-term memory model. *Adv. Space Res.* **2025**, *75*, 5386–5401.
27. Wang, M.; An, H.; Cai, W.; Shao, X. Wavelet transform makes water an outstanding near-infrared spectroscopic probe. *Chemosensors* **2023**, *11*, 37. [[CrossRef](#)]
28. Chuang, Y.-K.; Yang, I.-C.; Lo, Y.M.; Tsai, C.-Y.; Chen, S. Integration of independent component analysis with near-infrared spectroscopy for analysis of bioactive components in the medicinal plant *Gentiana scabra* Bunge. *J. Food Drug Anal.* **2014**, *22*, 336–344. [[CrossRef](#)] [[PubMed](#)]
29. Yu, S.; Huang, Z.; Wu, Q.; Jia, J.; Wang, H.; Wang, S.; Shen, C. Constructing rapid and undamaged detection models for main physicochemical indexes of brewing sorghum based on near infrared spectrum. *Sci. Technol. Food Ind.* **2023**, *44*, 311–319.
30. Shao, X.; Wang, G.; Wang, S.; Su, Q. Extraction of mass spectra and chromatographic profiles from overlapping GC/MS signal with background. *Anal. Chem.* **2004**, *76*, 5143–5148. [[CrossRef](#)] [[PubMed](#)]
31. Yang, Q.; Liu, Q.; Li, J.; Zhang, Z.; Li, Y.; Zhang, M.; Wang, X.; Xie, G. Optimization of starch and soluble sugar models for complex biomass samples using near-infrared spectroscopy. *J. Food Compos. Anal.* **2025**, *148*, 108530. [[CrossRef](#)]
32. Li, J.; Li, M.; Cao, Q.; Liu, S.; Wei, C. Dimensionality reduction of spectral reflectance by dividing the error space of principal component analysis. *J. Imaging Sci. Technol.* **2021**, *66*, 020410-1–020410-7.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.