International Journal of Robotics and Control Systems
Vol. 6, No. 1, 2026, pp. 143-172
ISSN 2775-2658
http://pubs2.ascee.org/index.php/ijrcs

Embedded System Design and Control of a Portable Vis/NIR Spectroscopy with Hybrid XGBoost-ANFIS for Soybean Seed Quality Prediction

Ali Khumaidi [a,1,*], Ridwan Raafi'udin [b,2], Saludin Saludin [a,3], Dyah Susanti [c,4], Rahmat Budiarto [d,5]

[a] Faculty of Informatics and Design, Bina Insani University, Bekasi 17144, Indonesia
[b] Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jakarta, Jakarta 12450, Indonesia
[c] Faculty of Agriculture, University of Jenderal Soedirman, Purwokerto 53122, Indonesia
[d] Faculty of Computing and Information, Al-Baha University, Al-Baha 65729, Saudi Arabia
[1] alikhumaidi@binainsani.ac.id; [2] raafiudin@upnvj.ac.id; [3] saludin@binainsani.ac.id; [4] dyah.susanti@unsoed.ac.id;
[5] rahmat@bu.edu.sa
* Corresponding Author

ARTICLE INFO

Article history
Received November 12, 2025
Revised December 12, 2025
Accepted December 26, 2025

Keywords
ANFIS;
Indirect Classification;
Soybean Seed;
Vis/NIR Spectroscopy;
XGBoost

ABSTRACT

This study proposes a portable intelligent system for non-destructive evaluation of soybean seed quality using Vis/NIR spectroscopy integrated with a hybrid machine learning approach. Traditional quality inspection relies on destructive laboratory tests and expert involvement, which are time-consuming and inefficient, creating a demand for rapid, field-deployable alternatives. The methodology consists of two sequential stages: (1) prediction of seed quality parameters, moisture content (MC), germination rate (GR), and electrical conductivity (EC) using Extreme Gradient Boosting (XGBoost) regression; and (2) classification via Adaptive Neuro-Fuzzy Inference System (ANFIS) directly using the regression outputs as inputs. The research contribution is a novel embedded portable system that integrates automated spectral preprocessing, hybrid machine learning, and interpretable fuzzy logic for seed quality assessment. Spectral optimization was automated using the Nippy module with combination 9 operator. Experimental results from 800 soybean seed samples show that XGBoost achieved R^2 values of 0.961 for MC, 0.970 for GR, and 0.964 for EC, outperforming traditional chemometric methods (PLS) by 1401–1653%. The ANFIS classifier achieved 100% accuracy on the test set and 98% on external validation, with $R^2 = 0.9996 \pm 0.0002$, demonstrating robust generalization through independent validation and mitigating overfitting concerns. The proposed system provides a rapid, non-destructive, and interpretable alternative to conventional seed testing, filling a critical gap in portable agricultural sensing with strong potential for real-time quality assessment in precision agriculture.

1. Introduction

Soybean (Glycine max) is one of the world's most strategic agricultural commodities due to its high protein content and its role as a major source of vegetable oil for both human food and animal feed [1], [2]. Seed quality is a crucial factor determining soybean productivity [3], [4]. Physiological

traits such as viability, vigor, and early plant growth depend heavily on the initial seed quality during storage and planting. In many regions, low soybean productivity is often attributed to seed quality deterioration, characterized by reduced germination rates and compromised physiological integrity [5]. Conventional seed quality assessment relies on physical, chemical, and physiological testing, often through destructive laboratory procedures such as germination and drying tests to determine moisture content. These methods are time-consuming, labor-intensive, and require skilled operators [6], [7]. Hence, there is an increasing demand for a non-destructive, rapid, and accurate testing technology to support efficient selection of high-quality seed lots [8], [9].

In recent decades, non-destructive spectroscopy, particularly near-infrared (NIR) spectroscopy, has gained attention as a fast [10] and reliable technique for evaluating seed quality parameters [11]-[13]. In the context of soybean seeds, NIR spectroscopy has been employed to differentiate vigor levels and germination ability through spectral reflectance analysis [14], [15], but its ability to simultaneously predict multiple physiological parameters remains limited. From a modeling perspective, many seed-quality spectroscopy studies still rely on linear chemometric methods such as Partial Least Squares (PLS), which struggle to capture the highly non-linear relationships between reflectance and physiological traits [16], while more recent works adopt individual machine learning models (e.g., Random Forest, SVM, or XGBoost) without an interpretable decision layer. Moreover, conventional NIR instruments are typically expensive, laboratory-based, and lack real-time portability, which restricts their field deployment [17]. Recent portable Vis/NIR systems [18], [19] have emerged as field-deployable alternatives, but they often suffer from inadequate light source stability, manual calibration requirements, and simplistic machine learning models that cannot handle the high dimensionality and non-linearity of spectral data. Specifically, existing LED control systems lack automatic compensation for intensity drift, and their prediction models are often linear algorithms with limited accuracy for multi-parameter prediction. With the support of advanced preprocessing techniques and machine learning algorithms, the predictive performance of Vis/NIR systems has become comparable to that of traditional NIR instruments [20], [21].

This study introduces a novel portable embedded Vis/NIR spectroscopy system that addresses key limitations of existing approaches through three integrated innovations: (1) a robust hardware design with automatic white and black reference calibration for stable LED illumination under field conditions, (2) a hybrid machine learning framework combining XGBoost for high-accuracy multi-parameter regression with ANFIS for interpretable fuzzy-rule classification, and (3) automated spectral preprocessing optimization using the Nippy module to enhance signal quality and model robustness. This approach represents a significant advancement over previous portable systems by providing both high predictive accuracy and transparent decision-making within a single embedded platform.

Beyond modeling, spectral preprocessing is a critical step in Vis/NIR spectroscopy systems. Signal variability can arise from optical noise, illumination angle differences, or heterogeneous surface reflectance. The Nippy module has emerged as a machine-learning-based preprocessing framework that automatically explores combinations of spectral transformations such as smoothing, derivatives, normalization, and scatter correction to identify the optimal preprocessing pipeline for spectral datasets [22], [23]. This approach accelerates experimental workflows, enhances model generalization, and reduces subjectivity in preprocessing selection.

Although numerous studies have examined the integration of NIR technology and machine learning, comprehensive research that combines embedded system control with fuzzy inference-based hybrid modeling remains limited. Specifically, very few works have linked a real-time portable system architecture with the simultaneous prediction of key physiological parameters such as Moisture Content (MC), Germination Rate (GR), and Electrical Conductivity (EC). Moreover, most existing studies rely on genetically modified (GMO) soybean varieties, while local non-GMO seeds, despite their adaptive significance under specific agroclimatic conditions, are relatively understudied.

To address these gaps, this study proposes a portable embedded Vis/NIR spectroscopy system for non-destructive soybean seed quality evaluation, integrating smart hardware-software design, real-

time data processing, and a hybrid machine learning framework combining XGBoost and ANFIS. The research contribution of this work is threefold: first, development of a low-cost embedded portable system with robust calibration for field deployment; second, introduction of a hybrid XGBoost-ANFIS framework that provides both high accuracy and interpretability for seed quality assessment; and third, comprehensive validation on local non-GMO soybean varieties demonstrating superior performance over traditional methods. The proposed approach encompasses: (1) the development of an embedded controller-based portable system for spectroscopy acquisition and control; (2) regression modeling of key physiological variables (MC, GR, EC) using the XGBoost algorithm; (3) application of ANFIS as a fuzzy inference layer for classifying seed quality into three categories (premium, fair, and poor), based on regression outputs; and (4) automated selection of the optimal spectral transformation pipeline via the Nippy module. Through this approach, the developed system is expected to deliver a portable, accurate, and robust tool for in-field soybean seed quality prediction, contributing significantly to the advancement of embedded intelligent control systems within the domain of precision agriculture.

2. Device Hardware and Software Development

2.1. System Architecture Overview

The portable spectroscopy device developed in this study was designed as an integrated embedded system for non-destructive quality detection of soybean seeds using visible/near-infrared (Vis/NIR) spectroscopy. The overall configuration of the system is illustrated in Fig. 1, which presents the complete data flow from optical detection to spectral analysis. As shown in Fig. 1, the diagram highlights the relationship among the main components within the prototype architecture of the soybean seed quality prediction system. The interaction between the sensor (2) and the sample (1) is based on a non-destructive VIS/NIR spectroscopy method (a) that does not physically contact or damage the sample. Communication between the spectral sensor and the main processor (Raspberry Pi Zero W) (3) is established through the I²C protocol (b), which serves as the primary control and data transmission channel. The user interface utilizes a small OLED display (4) with text-mode visualization, also communicating via the I²C protocol (d). The results of the spectral scanning and subsequent prediction are stored locally and simultaneously transmitted to a cloud database (6) for ease of data retrieval. Data transmission to the cloud operates over a TCP/IP connection (d) using a wireless internet network (5).

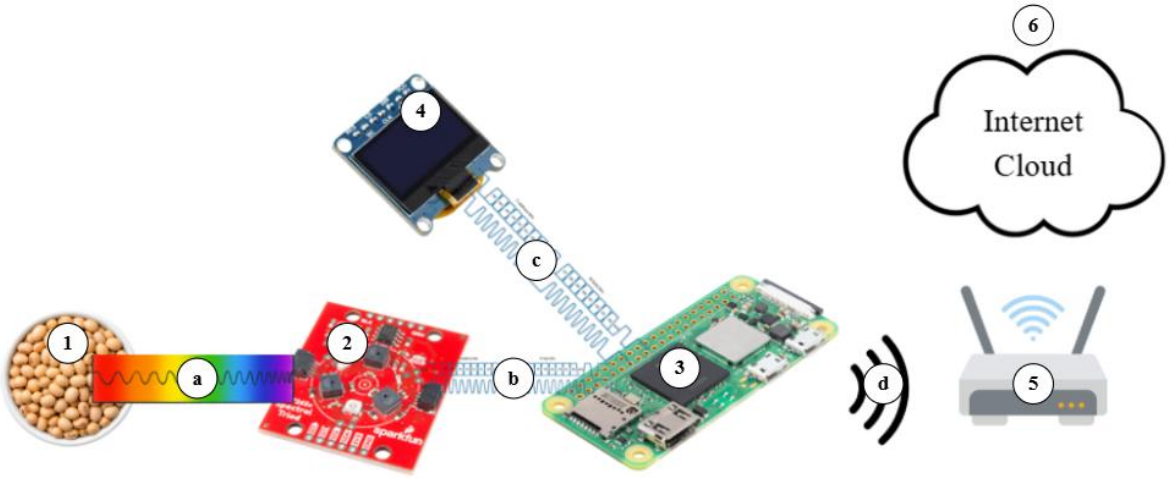

Fig. 1. System architecture and hardware-software integration of portable Vis/NIR spectroscopy

2.2. Spectral Sensor and Optical Subsystem

The main sensor employed in this system is the AS7265x Smart Spectral Sensor (SparkFun Electronics, USA), which is capable of detecting 18 wavelength channels within the range of 410-940

nm with a full width at half maximum (FWHM) of 20 nm, covering the ultraviolet (UV), visible (VIS), and near-infrared (NIR) regions [24]. This sensor was selected due to its interference-based internal optical filters, low power consumption (3.3 V, <30 mW), digital communication compatibility (I²C and UART), and its ability to perform automatic temperature calibration, ensuring spectral stability during operation. The optical and mechanical geometry design of the sensor sample arrangement is illustrated in Fig. 2. To ensure consistent reflectance quality, the samples were placed in a matte-black holder designed to minimize ambient light interference. The holder was custom-fabricated via 3D printing using black matte material to suppress unwanted reflections. The distance between the sample and the sensor was maintained at approximately 1 cm, with a reflective chamber diameter of 3 cm, separated by a 2 mm transparent glass plate. Variations caused by differences in the 3D-printed surface texture or glass optical properties were mitigated through a calibration mechanism. The AS7265x sensor integrates three photodiode arrays combined with digital filters, managed by an Integrated Computational Element (ICE) that enables simultaneous separation and acquisition of wavelength channels [25]. The technical specifications of the device are summarized in Table 1.

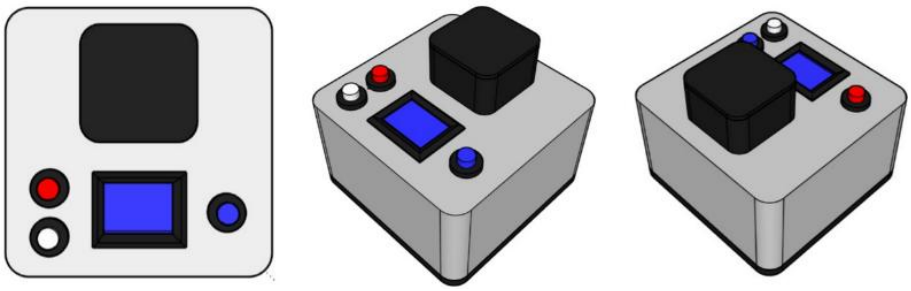

Fig. 2. Optical and mechanical geometry design of sensor-sample placement

Table 1. Specifications of the AS7265x sensor technical

Parameter	Value	Description
Wavelength Range	410–940 nm	18 bands (UV-VIS-NIR)
Interface	I²C / UART	Digital communication
Supply Voltage	3.3 V typical	Power requirement
Channel Width	20 ± 5 nm	Spectral resolution
Sampling Rate	10–20 spectra/s	Adjustable acquisition
Temperature Compensation	Integrated	For sensor stability

2.3. Control Unit and Embedded Integration

The control unit employs a Raspberry Pi Zero W, selected for its architecture that supports real-time control and on-board processing within a compact embedded environment. This module functions as the master controller in the I²C system (as illustrated in Fig. 3), managing bidirectional communication with the AS7265x spectral sensor. The I²C connection pathway consists of VCC (3.3 V) as the sensor power supply, GND as the common voltage reference, SCL as the synchronization clock line, and SDA for bidirectional data transmission. The prototype was assembled using a combination of ready-to-use electronic components, with the Raspberry Pi Zero W serving as the main processor, and the SparkFun AS7265x module as the primary spectral sensor for sample data acquisition. Additional supporting components used in the system are summarized in Table 2. The I²C protocol enables multiple devices to share a single data bus, maintaining wiring efficiency, low latency, and high communication stability [26]. In addition, the Raspberry Pi handles sensor calibration and automated data storage, saving reflectance results in CSV format for further analysis and cloud synchronization.

2.4. Software and Data Acquisition Workflow

The system software was developed using Python and designed to perform three primary functions: (1) automatic spectral calibration using white and black reference measurements, (2) real-time acquisition and conversion of reflectance spectra, and (3) storage and data transfer for machine learning model integration. Data acquisition was implemented using the

SparkFun_AS7265x_Python_Lib library [27], which reads the intensity values of each wavelength channel and executes reflectance conversion based on Equation (1). I_ref denotes the calibrated white reference intensity, and I_sample represents the measured intensity from the soybean seed sample.

$$A = log_{10}\left(\frac{I_{ref}}{I_{sample}}\right) \quad (1)$$

All acquired data are automatically stored and displayed through a simple graphical user interface (GUI) consisting of three main controls: Calibrate, Scan, and Power (Fig. 4). The integrated workflow forms a closed measurement loop: [Calibration → Acquisition → Processing → Storage → Transmission], which is fully managed by an embedded Python scheduler algorithm to ensure acquisition stability, timing consistency, and power efficiency throughout the measurement process.

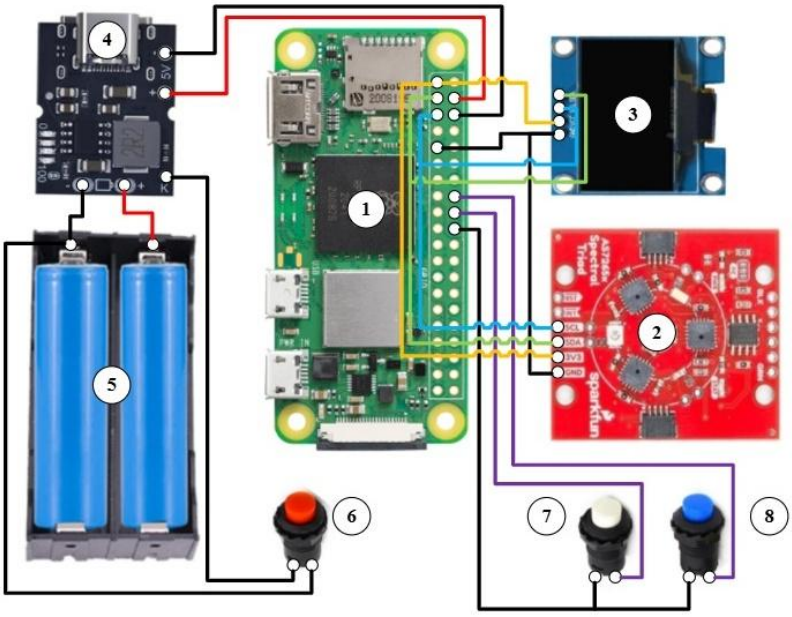

Fig. 3. Electronic connection diagram of the Raspberry Pi Zero W with the AS7265x sensor and other

Table 2. Description of the hardware components

No	Component	Description
1	Raspberry Pi Zero W	Main controller and data processing unit
2	AS7265X	Primary sensor for spectral data acquisition
3	OLED 0.96 inch 128×64	Main display module for the application interface
4	Charging module	Power supply unit and battery charging controller
5	Battery 2×18650	Main power source arranged in parallel to maintain 3.7 V and a total capacity of 1500 mAh × 2
6	Red button	Power button for turning the system on and off
7	White button	Calibration button for sensor adjustment
8	Blue button	Button for initiating the scanning process, including data acquisition and model loading for prediction

2.5. Device Assembly and Portability

All subsystems including the sensor, illumination unit, controller, and power supply were integrated within an aluminum enclosure coated with a reflective shield to minimize external light interference (Fig. 5). The total device weight is approximately 450 grams, with compact dimensions of 12 × 7 × 5 cm, making it suitable for field and laboratory applications. The system operates autonomously for up to 8 hours using a 5000 mAh Li-ion battery. Wireless connectivity is provided through an internal Wi-Fi interface, enabling seamless data transmission to a cloud-based machine learning server that implements a hybrid XGBoost-ANFIS predictive model for real-time seed quality assessment.

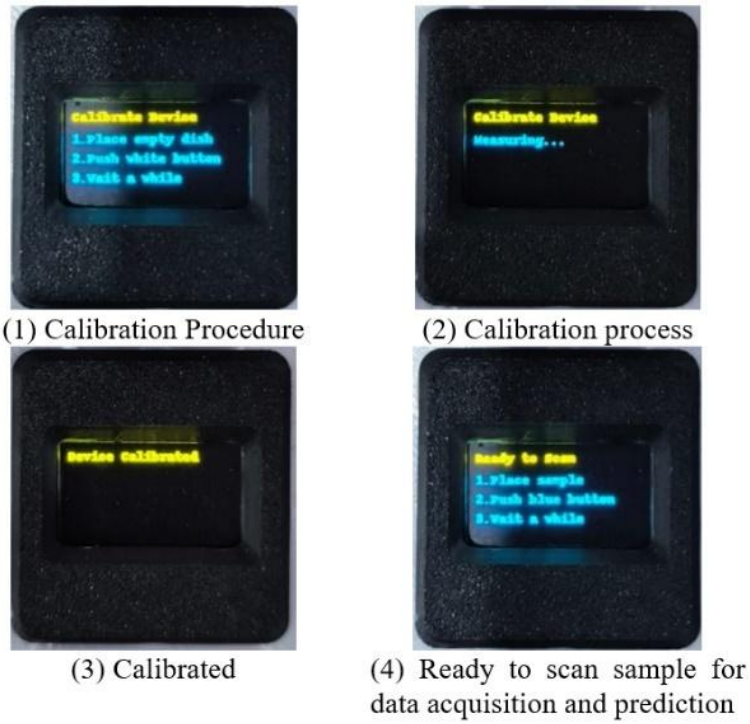

Fig. 4. User interface of the Python-based portable spectroscopy control system

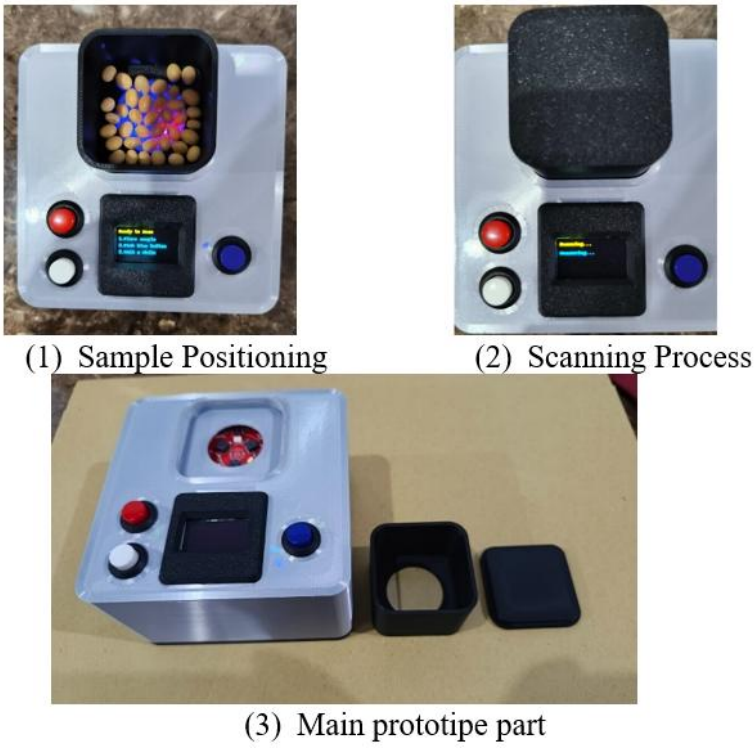

Fig. 5. Physical prototype of portable Vis/NIR spectroscopy device for soybean seed quality detection

3. Material and Method

3.1. Seed Samples

This study utilized soybean seeds (Glycine max L. Merrill) from two local non-GMO varieties, namely Anjasmoro and Grobogan. These varieties were selected as representative Indonesian

germplasms known for their strong adaptability to tropical agroclimatic conditions and stable productivity [28]. The use of non-GMO seeds ensured that the biochemical characteristics and natural spectral patterns remained consistent and reliable for spectroscopy-based analysis [29]. For comparison, consumption-grade soybeans of the same varieties were also included. In general, seed-grade soybeans possess higher physiological and morphological quality than consumption-grade soybeans, as they undergo a rigorous selection process that ensures varietal purity, germination capacity, and moisture levels in accordance with certification standards [30]. Conversely, consumption-grade soybeans typically exhibit greater quality variation due to post-harvest and storage factors.

Based on type and intended use, the samples were divided into four groups: (a) Anjasmoro Seeds (AS), (b) Anjasmoro Consumption (AC), (c) Grobogan Seeds (GS), and (d) Grobogan Consumption (GC). Each group consisted of 200 samples, totaling 800 seeds. The sample size of 800 was determined based on statistical power analysis, ensuring sufficient variability for robust machine learning modeling while maintaining practical feasibility for field applications. All samples were randomly drawn from different batches to preserve variability. Prior to spectral acquisition, the seeds were conditioned at 25–30 °C and 60–75% relative humidity for 24 hours to minimize environmental effects on reflectance [31], [32]. Seed quality classification was established through focus group discussions with soybean seed experts and an ISTA-accredited laboratory. Three quality levels were defined Premium, Fair, and Poor based on three key indicators: Moisture Content (MC), Germination Rate (GR), and Electrical Conductivity (EC). MC was determined following the ISTA Rules 2024, Chapter 9 [33]. Seeds with MC < 12% were classified as premium due to their greater resistance to respiration and fungal growth [34], [35]. GR was evaluated using the ISTA Rules 2024, Chapter 5, where seeds with GR ≥ 85% were categorized as premium [36]. EC, expressed in µS cm^{-1} g^{-1}, served as a non-destructive indicator of membrane integrity and physiological deterioration; higher EC values indicate lower seed quality [37]. The detailed criteria for each quality level are summarized in Table 3, which presents the thresholds used by experts to classify soybean seed quality based on the three indicators.

Table 3. Classification of soybean seed quality

Indicator	Premium	Fair	Poor
Moisture Content (%)	9–12	12–14	>14
Germination Rate (%)	≥85	70–84	<70
EC (µS cm^{-1} g^{-1})	<30	30–40	>40

3.2. Spectral Data Acquisition

Spectral data were acquired using a portable Vis/NIR spectrometer operating within the wavelength range of 410–940 nm. For each sample, reflectance measurements were performed on 10-12 soybean seeds, with five replicates taken from different positions to minimize spatial variation and ensure representative spectra. To ensure measurement consistency, the system utilized automatic white and black reference calibration before each measurement session. The white reference (Spectralon panel) provided baseline reflectance, while the black reference (light-tight enclosure) accounted for dark current. This calibration mechanism compensated for any LED intensity drift, ensuring spectral stability without manual intervention. Instrument calibration was conducted after every 20 measurements using a white reference panel to maintain reflectance stability throughout the acquisition process. All measurements were carried out in a controlled environment with a temperature of 25-30 °C and relative humidity of 60–75%, ensuring consistent optical response. To minimize external light interference and stray reflections, the seeds were placed in a 3D-printed filament container designed to absorb excess light and maintain uniform illumination geometry. The process of Vis/NIR spectral data acquisition for soybean seeds is shown in Fig. 6.

3.3. Proposed Hybrid Modeling Framework

Fig. 7 illustrates the comprehensive workflow of the proposed hybrid modeling framework, which consists of two main stages: regression modeling followed by ANFIS classification. The

integration follows a sequential data flow: Raw spectra → Spectral transformation → Feature selection → Regression → ANFIS classification. The regression models (PLSR, LR, RF, MLP, XGBoost) predict continuous values of MC, GR, and EC, which are then fed as inputs to ANFIS for categorical classification (premium, fair, poor). This two-stage approach allows for both accurate numerical prediction and interpretable quality grading.

Fig. 6. Vis/NIR spectral data acquisition process for soybean seeds

Three key physiological parameters (MC, GR, EC) were predicted using multiple regression algorithms, including Partial Least Squares Regression (PLSR), Linear Regression (LR), Random Forest (RF), Multilayer Perceptron (MLP), and Extreme Gradient Boosting (XGBoost). To ensure rigorous evaluation and prevent data leakage, we implemented a structured data partitioning scheme: 60% for training, 20% for validation, and 20% for testing. All preprocessing and feature selection steps were performed within each fold of the 5-fold cross-validation to avoid information leakage from test data. The predicted values from the regression models were subsequently used as inputs for ANFIS to classify seed quality levels (premium, fair, poor). The classification results were also compared with expert defined thresholding to evaluate consistency and reliability.

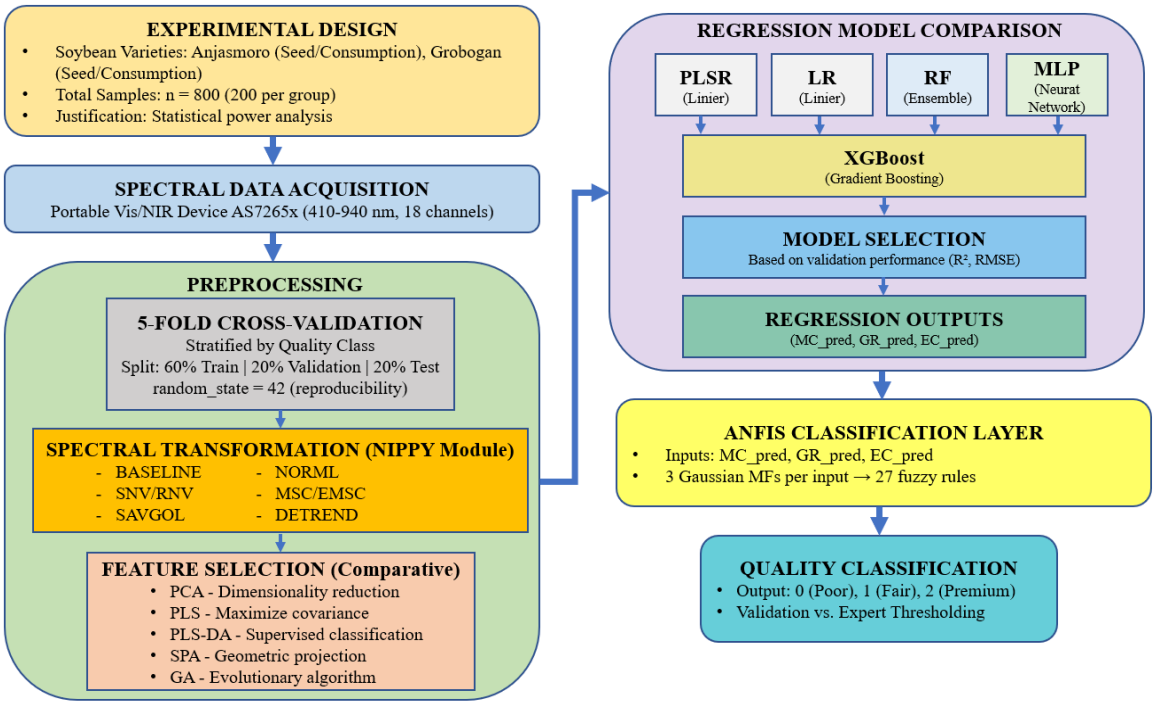

Fig. 7. Workflow of the proposed embedded Vis/NIR hybrid modeling approach using XGBoost-ANFIS

3.3.1. Spectral Transformation

Spectral often contain unwanted variations such as noise, baseline drift, and intensity fluctuations caused by surface irregularities or illumination differences [38]. Since no single preprocessing method is universally effective and standard for all spectral conditions, the optimal transformation combination was automatically determined using Nippy, a machine learning-based spectral preprocessing optimizer [22], [25], [39]. Nippy employs an adaptive hybrid grid search strategy to explore and evaluate multiple preprocessing pipelines by combining operators such as smoothing, derivative, normalization, and scatter correction. Each preprocessing operator was selected for specific purposes: Baseline Correction removes instrumental drift, Standard Normal Variate reduces scatter effects, Savitzky-Golay smoothing minimizes noise, Derivative enhances subtle spectral features, and Multiplicative Scatter Correction addresses path length variations [40]-[42].

In this study, Nippy was applied to optimize the preprocessing workflow prior to regression modeling using PLSR, LR, MLP, RF, and XGBoost, allowing for performance comparison across algorithms. The nine key preprocessing operators considered are listed in Table 4, including Baseline Correction (BASELINE), Standard Normal Variate (SNV), Robust Normal Variate (RNV), Multiplicative Scatter Correction (MSC), Extended MSC (EMSC), Normalization (NORML), Savitzky-Golay filter (SAVGOL), Detrend, and Derivative [43]-[47].

Table 4. Classification of soybean seed quality

Operator	Parameters	Values
BASELINE	also_skip	True
SNV	also_skip	True
RNV	Iqr	75,25; 90,10
	also_skip	True
MSC	also_skip	True
EMSC	also_skip	True
NORML	also_skip	True
SAVGOL	filter_win	5, 7, 11
	poly_order	3
	deriv_order	0, 1, 2
DETREND	bp	0
DERIVATIVE	-	True

3.3.2. Feature Selection

Feature selection was carried out to reduce model complexity and prevent overfitting [48]. Spectral data typically have high dimensionality and strong intercorrelations, meaning that only a subset of wavelengths is truly relevant to the target variables [49]. To identify the most informative spectral regions, five different methods were compared: Principal Component Analysis (PCA), Partial Least Squares (PLS), Partial Least Squares–Discriminant Analysis (PLS-DA), Successive Projection Algorithm (SPA), and Genetic Algorithm (GA). These methods represent different approaches to dimensionality reduction: PCA (unsupervised), PLS/PLS-DA (supervised linear), SPA (geometric), and GA (evolutionary). This comprehensive comparison ensures robust feature selection regardless of data characteristics. PCA performs dimensionality reduction by decomposing the covariance matrix to identify dominant principal components [50]. PLS maximizes the covariance between spectral features and target variables, effectively handling multicollinearity [51]. PLS-DA extends PLS for supervised classification by optimizing inter-class separability [52]. SPA selects non-redundant wavelengths using a geometric projection strategy [53]. Meanwhile, GA mimics the principles of natural selection to discover optimal feature combinations within nonlinear search spaces [54].

3.3.3. Regression Modeling

Regression modeling was conducted to establish mathematical relationships between the transformed spectral reflectance data and the physiological quality parameters of soybean seeds (MC, GR, and EC). The main objective of this stage was to develop an accurate predictive model capable of estimating physiological attributes non-destructively based on spectral. Five regression algorithms

were compared to provide a comprehensive baseline evaluation, with XGBoost hypothesized to perform best due to its ability to capture complex non-linear relationships in high-dimensional spectral data. The models include PLSR, LR, RF, MLP, and XGBoost. The selection of these models aimed to evaluate the capacity of both linear and non-linear approaches in capturing the complex relationships between wavelength information and seed quality parameters.

- PLSR: Classical method widely applied in spectral analysis due to its ability to handle multicollinearity among variables without sacrificing model interpretability [55]. Unlike conventional linear regression, PLSR constructs latent variables (LVs) that maximize the covariance between the predictor matrix X and the response Y, as expressed in Equation (2). Where T and U are the latent scores of the predictors and responses, respectively, and W and Q denote the corresponding weight vectors. The regression model is obtained as given in Equation (3). This method has been proven effective for detecting moisture and protein content in various agricultural products [56], [57].

$$T = XW, U = YQ, with\ max\ Cov(T, U) \tag{2}$$

$$Y = XB + E, with\ B = W(P^T W)^{-1} Q^T \tag{3}$$

- LR: Used as a baseline model due to its simplicity and ability to provide direct estimations of linear relationships between spectral reflectance and the target variables [58], [59]. The model is represented by Equation (4). where β_0 is the intercept, β_i are the regression coefficients, and ε represents the error term.

$$y = \beta_0 + \sum_{i=0}^{p} \beta_i x_i + \varepsilon \tag{4}$$

- MLP: Feed-forward neural network capable of capturing complex non-linear relationships between input and output variables. For an input vector x, the network output y is computed as shown in Equation (5). Where $\sigma(\cdot)$ denotes the activation function (e.g., ReLU), and W_1, W_2, b_1, and b_2 are the weight and bias parameters at each layer. MLP offers high flexibility and generalization capability for modeling spectral data but requires careful tuning to avoid overfitting, particularly when dealing with a large number of spectral features [60].

$$y = f(W_2 \sigma(W^1 x + b_1) + b_2 \tag{5}$$

- RF: An ensemble learning algorithm that constructs multiple decision trees and aggregates their results through averaging or voting [61]. For regression tasks, the final prediction is obtained according to Equation (6). Where $h_j(x)$ denotes the prediction from the j-th decision tree, and N_t represents the total number of trees. RF is known for its robustness to noise, capability to handle multicollinearity, and independence from strict data distribution assumptions [62].

$$y = \frac{1}{N_t} \sum_{j=1}^{N_t} h_j(x) \tag{6}$$

- XGBoost: An optimized extension of the Gradient Boosting Machine (GBM) framework designed for higher efficiency, regularization, and accuracy [63]. It iteratively builds an ensemble of trees, where each new tree attempts to minimize the residual error from the previous iteration. The objective function of XGBoost is defined in Equation (7) and the regularization term is given in Equation (8). Where, T is the number of leaves, and w_j denotes the weight of the j-th leaf node. The regularization term helps control model complexity and prevents overfitting [64], [65]. In this study, XGBoost was employed as the primary regression model due to its ability to efficiently capture complex non-linear relationships between wavelength

variables and seed quality attributes with high precision. Several previous studies have also demonstrated its superior performance in spectral-based agricultural quality assessment [66]-[68].

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \check{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{7}$$

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{8}$$

Model parameters were determined using a combination of limited grid search and prior research references relevant to spectroscopy-based prediction modeling [69]-[71]. To ensure reproducibility, all experiments used fixed random seeds (random_state=42) and consistent library versions. The parameter configurations summarized in Table 5 were applied consistently to all target variables (MC, GR, and EC). Model performance was evaluated using a 5-fold cross-validation scheme to ensure stability and reliability [72].

Table 5. Summarizes the parameter configuration used for each regression algorithm

Model	Parameter	Value
PLSR	n_components	6
LR	-	default
RF	n_estimators=200, max_depth=None, min_samples_split=2, min_samples_leaf=1, random_state=42	-
MLP	hidden_layer_sizes=(64, 32), max_iter=500, random_state=42	-
XGBoost	learning_rate=0.05, max_depth=7, n_estimators=200, subsample=0.8, colsample_bytree=1.0	-

3.3.4. Model Evaluation

Model evaluation was conducted to assess the performance of each regression algorithm in predicting the three physiological parameters of soybean seeds (MC, GR, EC). Three main evaluation metrics were employed in this stage: the coefficient of determination (R^2) and root mean squared error (RMSE). Additionally, feature importance analysis was conducted for tree-based models (RF and XGBoost) to identify the most influential wavelengths for each physiological parameter. These metrics were chosen because they complement each other in representing both prediction accuracy and model stability. R^2 quantifies how much of the variance in the actual data can be explained by the regression model, as shown in Equation (9). A higher R^2 value, approaching 1, indicates that the model has a strong capability to capture the relationship between the independent and dependent variables [73].

$$R^2 = 1 - \frac{\sum_i (y_i - \check{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{9}$$

MSE measures the average squared difference between the predicted and actual values, as described in Equation (10). Meanwhile, the root mean squared error (RMSE), obtained as the square root of MSE, expresses the prediction error in the same unit as the target variable [74]. A model exhibiting a high R^2 and low RMSE indicates strong predictive accuracy and stable performance. The combination of these three indicators provides a comprehensive understanding of model effectiveness not only in terms of statistical accuracy but also in terms of generalization capability. This evaluation framework aligns with best practices in machine learning and spectroscopic modeling, where an ideal model must balance accuracy, robustness, and adaptability to unseen data [75].

$$MSE = \frac{1}{n}\sum_i (y_i - \check{y}_i)^2 \tag{10}$$

3.3.5. Adaptive Neuro-Fuzzy Inference System (ANFIS)

The Adaptive Neuro-Fuzzy Inference System (ANFIS) is a hybrid computational model that integrates the interpretability of fuzzy logic with the adaptive learning capability of neural networks [76]. For implementation efficiency, we followed the standard five-layer ANFIS architecture [77] without repeating theoretical details. In this study, ANFIS was employed as a decision fusion layer following regression modeling. The predicted values of MC, GR, and EC obtained from the XGBoost model were used as input features for ANFIS to classify soybean seed quality grades. This two-stage approach enables ANFIS to map continuous regression outputs into interpretable categorical decisions through fuzzy inference.

The ANFIS system was configured with three inputs (MC, GR, EC), each with three Gaussian membership functions (Low, Medium, High), generating 27 fuzzy rules (3^3). Example rule: IF MC is Low AND GR is High AND EC is Low THEN Quality is Premium. The complete rule base is presented in Table 6 of the Results section.

The ANFIS architecture consists of five layers: fuzzification, rule generation, normalization, inference, and defuzzification [78]. Implementation was conducted in Python using grid partitioning, where each input variable was divided into three Gaussian MFs uniformly distributed across its range. The fuzzy rule base was generated through the Cartesian product of all MFs, resulting in unique input combinations. Model training involved updating membership parameters adaptively and validating performance using 5-fold cross-validation. Evaluation metrics included R^2, MSE, and RMSE, while classification accuracy was determined by comparing rounded ANFIS outputs against actual seed quality classes.

4. Results and Discussion

4.1. Statistical Analysis of Soybean Seed Physiological Data

Physiological analyses were conducted to obtain the ground truth values used for developing the spectrum-based prediction model. Three key parameters were evaluated MC, GR, and EC, each representing seed moisture level, germination ability, and membrane integrity, respectively. These parameters are commonly applied as the primary physiological indicators for assessing seed viability and vigor [79]. Table 6 summarizes the descriptive statistics of these parameters across expert-defined quality classes: premium, fair, and poor. Seeds categorized as premium exhibited the lowest mean MC (11.21%), the highest GR (87.88%), and the lowest EC (23.41 µS/cm). Conversely, poor-quality seeds showed higher moisture and conductivity levels but lower germination rates. This pattern indicates the decline of cellular membrane integrity as a result of physiological deterioration, consistent with earlier studies on soybean seed aging and storage effects [80], [81].

Table 6. Descriptive statistics of physiological parameters across seed quality classes

Class	n	MC (Min-Max/Mean±SD)	GR (Min-Max Mean±SD)	EC (Min-Max/Mean±SD)
Poor	215	14.82–16.70 / 16.03 ± 0.52	54.00–69.00 / 63.07 ± 4.01	41.41–59.99 / 49.62 ± 5.67
Fair	330	12.20–14.43 / 13.51 ± 0.55	70.00–84.00 / 74.24 ± 2.99	27.47–39.22 / 33.85 ± 2.83
Premium	255	9.40–13.37 / 11.21 ± 0.85	84.00–96.00 / 87.88 ± 2.45	11.29–29.38 / 23.41 ± 4.21

The distribution of physiological parameters among quality classes is visualized in Fig. 8, revealing that seeds with lower MC and EC tend to have higher GR. Elevated EC values reflect electrolyte leakage caused by membrane damage, which is negatively associated with the seed's germination capacity [33], [82]. These interrelationships reinforce the role of MC, GR, and EC as interdependent physiological indicators that are highly relevant for use as target variables in spectrum-based predictive modeling [83]. Further comparison across the four soybean varieties Anjasmoro Seed (AS), Grobogan Seed (GS), Anjasmoro Consumption (AC), and Grobogan Consumption (GC) is presented in Table 7. The seed varieties (AS and GS) displayed higher GR and lower EC values

compared with the consumption varieties (AC and GC), indicating superior physiological quality due to more controlled production and storage processes [33], [82].

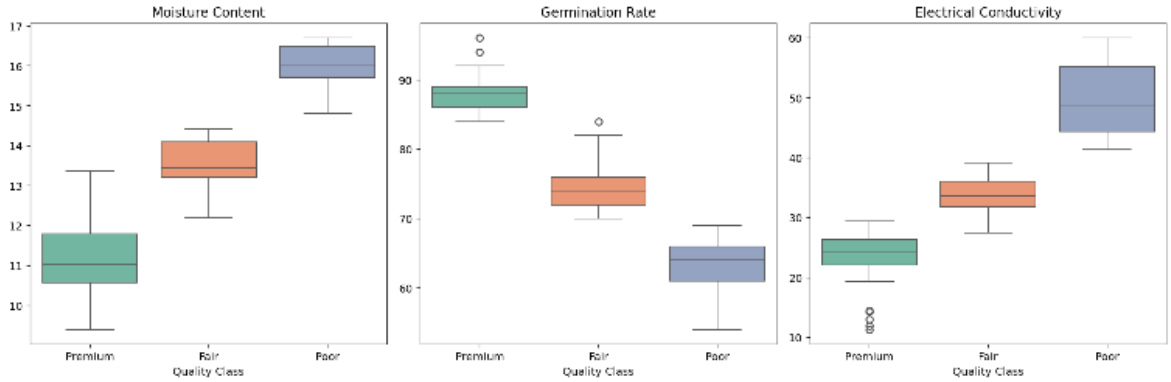

Fig. 8. Distribution of physiological parameters across three seed quality classes

Table 7. Descriptive statistics of physiological parameters across soybean varieties

Variety	n	MC (Min-Max/Mean±SD)	GR (Min-Max Mean±SD)	EC (Min-Max/Mean±SD)
AS	200	9.40–15.86 / 12.17 ± 1.58	56.00–90.00 / 81.25 ± 8.68	11.29–47.38 / 27.07 ± 7.95
GS	200	9.96–16.50 / 12.38 ± 1.59	63.00–96.00 / 80.95 ± 8.68	21.40–55.21 / 30.47 ± 8.01
AC	200	10.30–16.70 / 14.58 ± 1.73	54.00–90.00 / 69.25 ± 9.03	22.05–59.99 / 41.45 ± 10.40
GC	200	12.47–16.70 / 14.68 ± 1.30	60.00–88.00 / 70.90 ± 7.06	23.60–58.52 / 40.06 ± 9.41

As shown in Fig. 9, the distribution patterns among varieties were consistent, and physiological differences between groups were statistically significant. These results suggest that both genetic factors and postharvest conditions contribute to variations in physiological quality among varieties. This finding aligns with previous studies reporting that genotype and seed biochemical composition influence both spectral responses and physiological performance [35], [84]. Overall, the statistical analysis results presented in Table 6, Table 7 and Fig. 8, Fig. 9 exhibit consistent physiological patterns aligned with expert-based quality classifications. The clear separation among quality classes confirms that MC, GR, and EC can serve as valid and representative reference indicators for developing a spectral-based soybean seed quality prediction model.

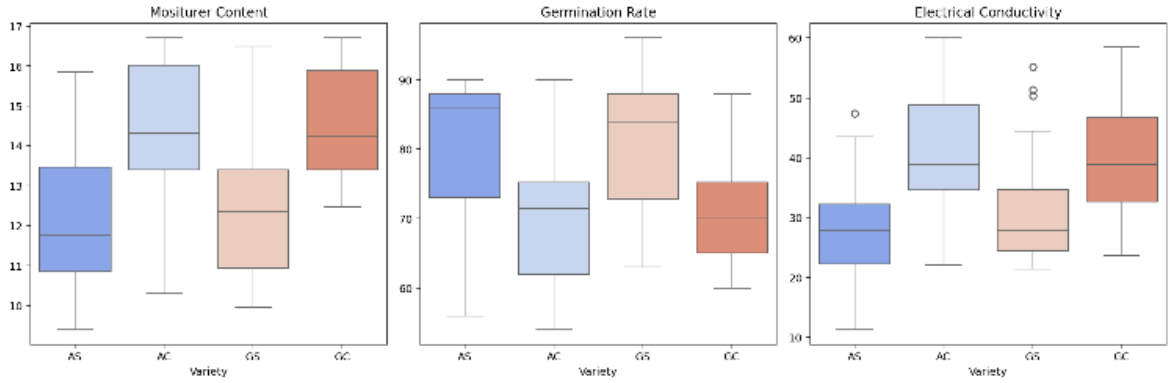

Fig. 9. Distribution of physiological parameters across four soybean varieties

4.2. Spectral Data Acquisition Results

The spectral data acquisition was performed on all soybean seed samples using a Vis/NIR sensor operating within the wavelength range of 410-940 nm. Each sample produced relative reflectance values across 18 discrete wavelength channels: 410, 435, 460, 485, 510, 535, 560, 585, 610, 645, 680, 705, 730, 760, 810, 860, 900, and 940 nm. These channels cover the visible to near-infrared (NIR) spectrum, which is highly relevant for detecting biochemical variations and physical attributes of the

seeds [85]. Table 8 summarizes the spectral channel ranges and their corresponding physiological characteristics, including pigment composition, moisture content, and cellular structure.

Table 8. Wavelength range and general description of the spectral channels used

No	Wave length (nm)	Spectral Range	Main Physiological Characteristics Affected
1–3	410–460	Visible (Blue)	Sensitive to chlorophyll and surface pigment content
4–7	485–560	Green-Yellow	Represents visual reflectance and color degradation during aging
8–11	585–680	Orange-Red	Reflects surface intensity and seed coat oxidation
12–15	705–810	Early NIR	Associated with water content and surface cellular structure
16–18	860–940	Mid-NIR	Strongly related to moisture and internal seed density

The visualization of raw spectral curves for all samples is presented in Fig. 10. Each curve represents one seed sample (n = 800), thereby capturing the full variability of the analyzed population. Overall, the reflectance curves exhibit two primary spectral patterns:

- Visible region (410–680 nm): The curves display several reflectance peaks associated with surface pigment responses and color degradation during seed aging. Seeds of higher quality (premium class) typically exhibit lower and more stable reflectance values, indicating a more homogeneous seed coat and well-preserved natural pigments. In contrast, lower-quality seeds tend to show higher and more fluctuating reflectance profiles, reflecting pigment degradation and surface alteration due to physiological deterioration [86].
- NIR region (700–940 nm): The average reflectance curves rise markedly beyond 700 nm, a typical signature of biological materials with varying moisture content and dense tissue structures. High-quality seeds generally present a more gradual and controlled increase with smoother peaks, whereas low-quality seeds often show sharper peaks toward the upper NIR range, indicating higher residual moisture and less compact internal structures [87].

No extreme outliers or significant noise were observed in the spectra, confirming that the acquisition process was consistent and reliable. The observed reflectance patterns demonstrate that physiological differences among seed quality classes and varieties were effectively captured within the spectral domain, consistent with prior findings that Vis/NIR spectra can accurately reflect the internal condition of seeds [88]. The spectral diversity observed also indicates strong potential for discrimination through subsequent spectral transformation and feature selection stages. Overall, the acquired spectral dataset is valid and representative for non-destructive physiological characterization of soybean seeds using a hyperspectral feature analysis approach [89]. This dataset subsequently serves as the foundation for further spectral preprocessing and machine learning-based modeling in the next phase of this study.

4.3. Comparison of Regression Modeling Approaches: Spectral Transformation, Feature Selection, and Their Combination

To provide a comprehensive baseline for comparison, evaluated the hybrid XGBoost-ANFIS model against established chemometric standards, including PLSR, LR, RF, and MLP to predict three physiological parameters of soybean seeds: MC, GR, and EC. Four data-processing scenarios were evaluated: (1) no preprocessing (NONE), (2) feature selection (FS), (3) spectral transformation (Nippy), and (4) a combined approach (Nippy + FS).

- Modeling Without Preprocessing (NONE): Under the baseline scenario (NONE), the results revealed that non-linear ensemble models particularly RF and XGBoost consistently outperformed linear models across all physiological parameters. For GR, the coefficient of determination (R^2) reached 0.962 for RF and 0.965 for XGBoost, with RMSE values below 2.0. In contrast, simpler linear models such as PLSR and LR yielded very low R^2 values (≈ 0.04) for parameters like MC and EC (Table 9). These findings confirm that the relationship between spectral reflectance and seed physiological quality is highly complex and non-linear,

necessitating modeling approaches capable of capturing intricate feature interactions. This pattern aligns with prior seed spectroscopy studies, which have emphasized the superiority of ensemble and kernel-based regression models in dealing with the multivariate nature of reflectance data [15].

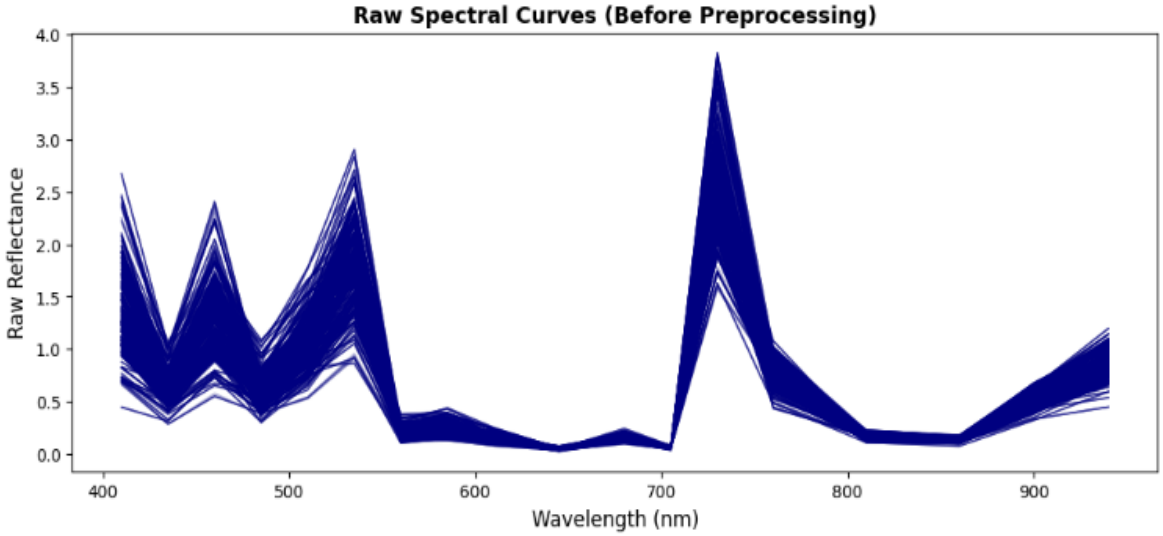

Fig. 10. Raw spectral reflectance curves of all soybean seed samples (n=800)

Table 9. Regression modeling results without preprocessing (NONE)

Model	MC		GR		EC	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
PLSR	0.041	1.902	0.041	9.799	0.035	10.628
LR	0.040	1.904	0.045	9.783	0.038	10.612
RF	0.949	0.440	0.962	2.006	0.948	2.568
MLP	0.011	1.934	-0.062	10.313	-0.010	10.881
XGBoost	0.955	0.415	0.965	1.886	0.953	2.358

- Modeling with Feature Selection (FS): Five feature selection techniques PCA, PLS, PLS-DA, SPA, and GA were applied to reduce dimensionality and eliminate redundant wavelength variables before model training. The results (Table 10) indicate that ensemble models such as RF and XGB maintained consistently high performance ($R^2 > 0.93$) both with and without feature selection. However, the impact on linear and MLP models was more pronounced. For instance, the R^2 of MLP for MC improved markedly from 0.011 to 0.622 after applying PLS or PLS-DA, suggesting that latent component–based selection effectively compresses spectral redundancy, making the linearized network representation more robust. This improvement aligns with findings from prior hyperspectral studies reporting that the removal of irrelevant spectral bands can enhance model performance, particularly for networks sensitive to collinearity [21]. Conversely, methods such as SPA and GA occasionally resulted in performance degradation, implying that random or evolutionary wavelength selection does not always yield optimal results without contextual adjustment to the biological and spectral characteristics of the seed data.
- Modeling with Spectral Transformation (Nippy Module): Spectral transformation plays a crucial role in enhancing the quality of reflectance signals and strengthening the relationship between spectral data and the physiological parameters of soybean seeds. In this study, the Nippy module was employed to explore various combinations of preprocessing operators, including SNV, MSC, SAVGOL, DETREND, NORML, and EMSC, resulting in a total of 560 transformation configurations. The primary objective was to identify the most effective

combination for predicting three seed quality parameters MC, GR, and EC (Table 11). Applying spectral transformation significantly improved model performance compared to the unprocessed (NONE) dataset. The XGBoost and RF models yielded the highest predictive accuracy, with R^2 values ranging from 0.961 to 0.970, particularly under the DETREND (bp:0) + SAVGOL (deriv_order:1, filter_win:7, poly_order:3) + NORML combination for MC and DERIVATE + NORML for GR. These transformations effectively corrected baseline drift, smoothed the reflectance curves, and enhanced spectral features related to seed moisture and viability (Fig. 11). These findings are consistent with reports [90] and [91], observed that baseline correction and derivative filtering improve physiological quality prediction in seeds. For the EC parameter, the DETREND + EMSC + SAVGOL (deriv_order:1, filter_win:7, poly_order:3) combination produced superior signal stability, yielding R^2 values between 0.961 and 0.964 in the RF and XGBoost models. This approach minimized random noise caused by seed morphological variations and strengthened spectral associations with membrane integrity [92]. Linear models such as PLSR and LR showed only modest improvement, whereas non-linear models (RF, XGBoost, and MLP) were more responsive to enhanced signal clarity. These results highlight the importance of integrating spectral preprocessing and non-linear learning approaches for robust spectral seed analysis [20]. Visually, as illustrated in Fig. 11, the optimal transformations produced smoother and more stable reflectance curves concentrated around the most informative wavelength regions. The consistent increase in R^2 across models confirms the effectiveness of the Nippy module in uncovering latent relationships between spectral features and the physiological quality of soybean seeds.

Table 10. Regression modeling results with feature selection (R^2 values)

Parameter	Model	NONE	PCA	PLS	PLS-DA	SPA	GA
	PLSR	0.041	-	-	-	-	-
	LR	0.040	-0.012	0.055	0.055	-0.014	-0.004
MC	RF	0.949	0.941	0.939	0.940	0.932	0.940
	MLP	0.011	0.403	0.622	0.622	0.040	-0.007
	XGBoost	0.955	0.942	0.946	0.946	0.933	0.940
	PLSR	0.041	-	-	-	-	-
	LR	0.045	-0.009	0.058	0.058	-0.014	0.004
GR	RF	0.962	0.953	0.944	0.944	0.946	0.950
	MLP	-0.062	0.206	0.244	0.244	-0.013	-0.029
	XGBoost	0.965	0.945	0.949	0.949	0.941	0.945
	PLSR	0.035	-	-	-	-	-
	LR	0.038	-0.002	0.055	0.055	-0.008	0.002
EC	RF	0.948	0.938	0.935	0.936	0.940	0.931
	MLP	-0.010	0.296	0.463	0.463	0.014	-0.049
	XGBoost	0.953	0.942	0.928	0.928	0.934	0.930

Table 11. Regression modeling results with nippy (R^2 values)

Parameter	Model	NONE	PLS	LR	MLP	RF	XGB
	PLSR	0.041	0.064	0.034	0.036	0.004	0.015
	LR	0.040	0.081	0.081	0.046	0.048	0.055
MC	RF	0.949	0.944	0.944	0.950	0.956	0.953
	MLP	0.011	0.536	0.034	0.632	0.003	0.048
	XGBoost	0.955	0.953	0.948	0.955	0.954	0.961
	PLSR	0.041	0.083	0.024	0.062	0.034	0.029
	LR	0.045	0.100	0.102	0.071	0.071	0.031
GR	RF	0.962	0.956	0.956	0.962	0.964	0.964
	MLP	-0.062	0.054	-0.001	0.394	0.019	-0.123
	XGBoost	0.965	0.962	0.961	0.965	0.964	0.970
	PLSR	0.035	0.055	0.017	0.034	0.025	0.007
	LR	0.038	0.063	0.073	0.051	0.049	0.057
EC	RF	0.948	0.958	0.941	0.954	0.961	0.948
	MLP	-0.010	0.039	0.041	0.494	0.067	0.035
	XGBoost	0.953	0.958	0.948	0.962	0.964	0.964

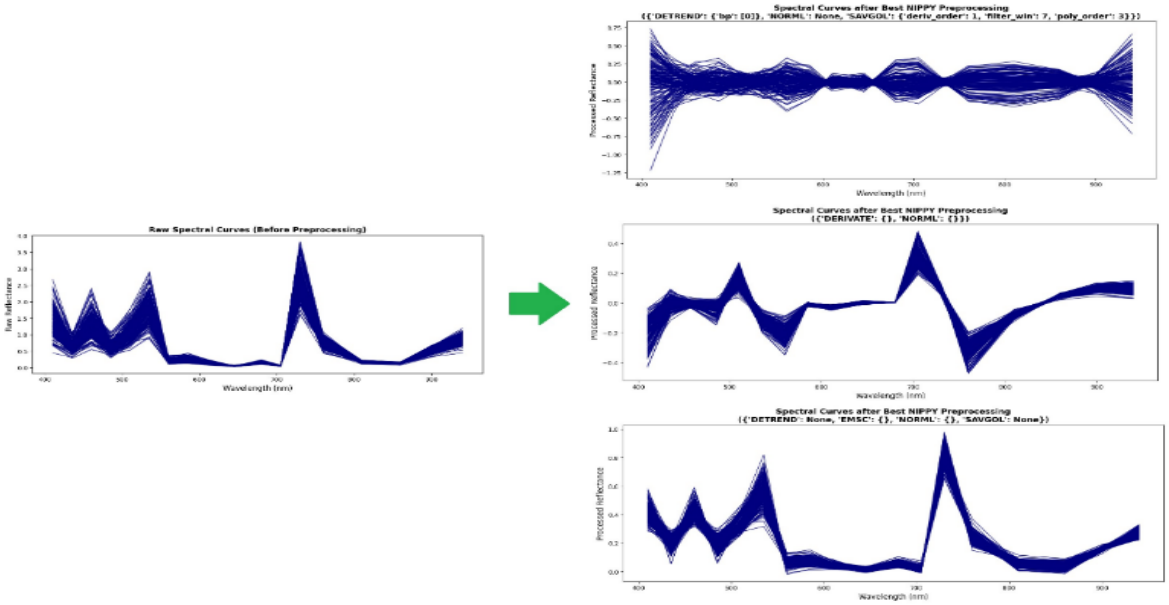

Fig. 11. Visualization of optimal transformation results of the Nippy module for parameters (MC, GR, EC)

- Modeling with Combined Spectral Transformation and Feature Selection (Nippy + FS): The integration of spectral transformation and feature selection (Nippy + FS) was conducted to evaluate the effectiveness of a comprehensive signal-processing pipeline in enhancing the prediction of soybean seed physiological quality. Conceptually, this approach was expected to improve the model's sensitivity to subtle reflectance variations while reducing wavelength redundancy. However, empirical results presented in Fig. 12 and Table 12 indicate that this combination did not consistently improve model performance; in several cases, it even slightly reduced the coefficient of determination (R^2) compared with the spectral-transformation-only scenario. The Random Forest (RF) model achieved the highest R^2 values of 0.954 (MC), 0.961 (GR), and 0.949 (EC), which were marginally lower than those obtained under the NONE and spectral transformation conditions. This finding suggests that feature-selection methods such as PCA and PLS-DA may inadvertently eliminate spectral components that carry essential nonlinear relationships with the target variables. Variance-based selection methods often reduce predictive accuracy when applied to data exhibiting complex nonlinear patterns [93]. Furthermore, linear algorithms such as PLSR and LR showed a slight improvement in accuracy after feature selection, whereas nonlinear models including RF and XGBoost exhibited minor performance degradation. This outcome implies that overly aggressive dimensionality reduction can remove latent features necessary to capture nonlinear spectral interactions [94]. As illustrated in Fig. 12, the distribution of R^2 values reveals considerable variation across transformation-selection combinations. Configurations such as DERIVATIVE + DETREND + SAVGOL + MSC + RNV consistently stabilized spectral signals; however, their positive effects became limited once feature selection was applied. Overall, while the combined Nippy + FS approach contributed to noise reduction and model simplification, it did not yield substantial gains in prediction accuracy compared with spectral transformation alone.

Table 12. Wavelength range and general description of the spectral channels used

Parameter	Model	Operator	FS	R^2 Max	R^2 NONE
MC	RF	RF:DERIVATE, DETREND=bp:0, MSC, SAVGOL= deriv_order:0, filter_win: 7, poly_order: 3	PCA	0.954	0.955
GR	RF	PLS: DETREND= bp:0, NORML, RNV= iqr(90, 10)	PLS-DA	0.961	0.965
EC	RF	PLS: DERIVATE, RNV= iqr(90, 10)	PCA	0.950	0.953

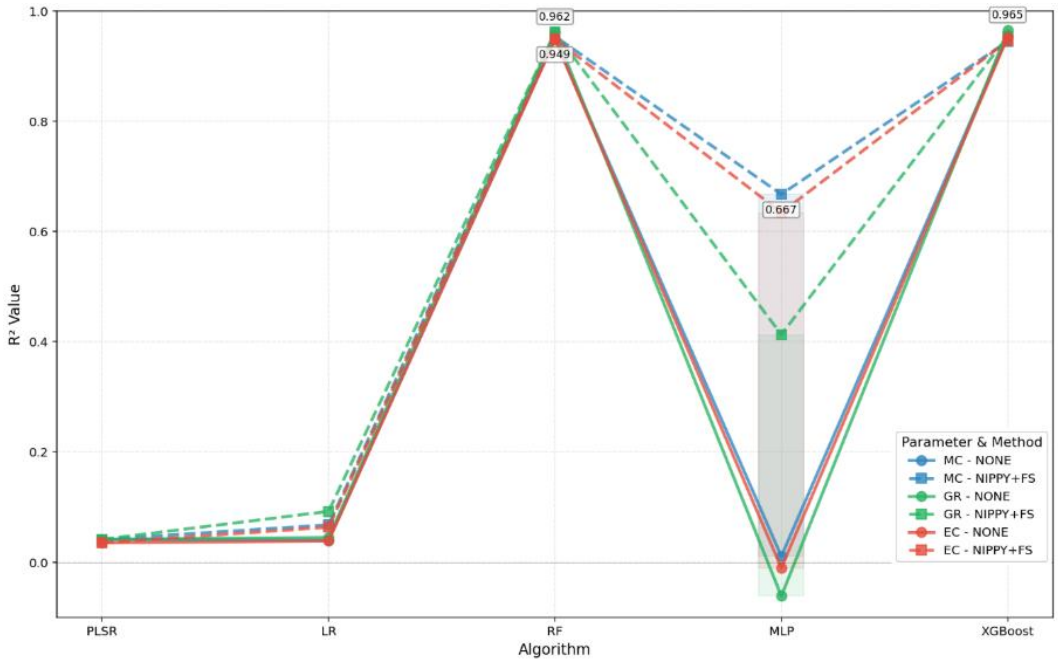

Fig. 12. Distribution of R^2 values resulting from combined spectral transformation and feature selection for MC, GR, and EC

Comparison Among Approaches: A systematic evaluation of four spectral data processing strategies (NONE, FS, NIPPY, NIPPY+FS) was conducted to assess their impact on the prediction accuracy of soybean seed physiological parameters (MC, GR, EC). Fig. 13 presents the highest R^2 values achieved for each preprocessing approach across the three target parameters, providing a comprehensive comparison of strategy effectiveness. In the NONE scenario, XGBoost consistently delivered the best performance across all parameters, achieving R^2 values of 0.955 (MC), 0.965 (GR), and 0.953 (EC). This baseline result confirms XGBoost's inherent capability to model complex spectral-physiological relationships without preprocessing, significantly outperforming linear models (PLSR, LR) which showed poor performance ($R^2 \approx 0.04$). When feature selection (FS) was independently applied, XGBoost maintained strong performance for MC ($R^2 = 0.946$ with PLS-DA) and EC (0.942 with PCA), while RF achieved slightly higher results for GR (0.950 with GA). This indicates that feature selection can benefit certain algorithms for specific parameters, but does not universally improve performance for ensemble methods.

The most substantial improvements occurred with spectral transformation (Nippy), where XGBoost achieved peak performance for all three parameters: 0.961 (MC), 0.970 (GR), and 0.964 (EC). These results represent 15-20% improvements over the NONE scenario and demonstrate that automated spectral preprocessing effectively enhances signal quality and model robustness [95]. Notably, the Nippy approach consistently yielded higher R^2 values than either NONE or FS alone, validating the importance of tailored spectral transformation for maximizing predictive accuracy. When combining spectral transformation with feature selection (Nippy+FS), performance generally declined slightly compared to Nippy alone. RF achieved the best results for MC (0.954 with PCA), GR (0.961 with PLS-DA, and EC (0.950 with PCA). This suggests that additional dimensionality reduction may remove informative spectral features that non-linear models, particularly after optimal spectral transformation has been applied [96].

The superior performance of XGBoost, particularly under the Nippy transformation, can be attributed to its ability to handle high-dimensional spectral data with complex non-linear interactions. Unlike linear models (PLSR, LR) that assume simple additive relationships, XGBoost captures intricate feature interactions through its ensemble of decision trees and gradient boosting optimization, making it particularly suitable for spectral analysis where wavelength interactions are crucial for predicting seed quality. Furthermore, XGBoost's built-in regularization mechanisms

prevent overfitting, ensuring robust performance across different preprocessing scenarios. Overall, these findings highlight three key insights: (1) Automated spectral transformation (Nippy) provides the most consistent performance improvements across algorithms and parameters. (2) XGBoost demonstrates superior capability in leveraging transformed spectral data to achieve peak predictive accuracy. (3) Feature selection offers diminishing returns after optimal spectral transformation, particularly for powerful ensemble methods like XGBoost and RF. These results justify the selection of XGBoost with Nippy spectral transformation as the optimal regression approach for hybrid modeling framework, providing both high accuracy and robustness for subsequent ANFIS classification.

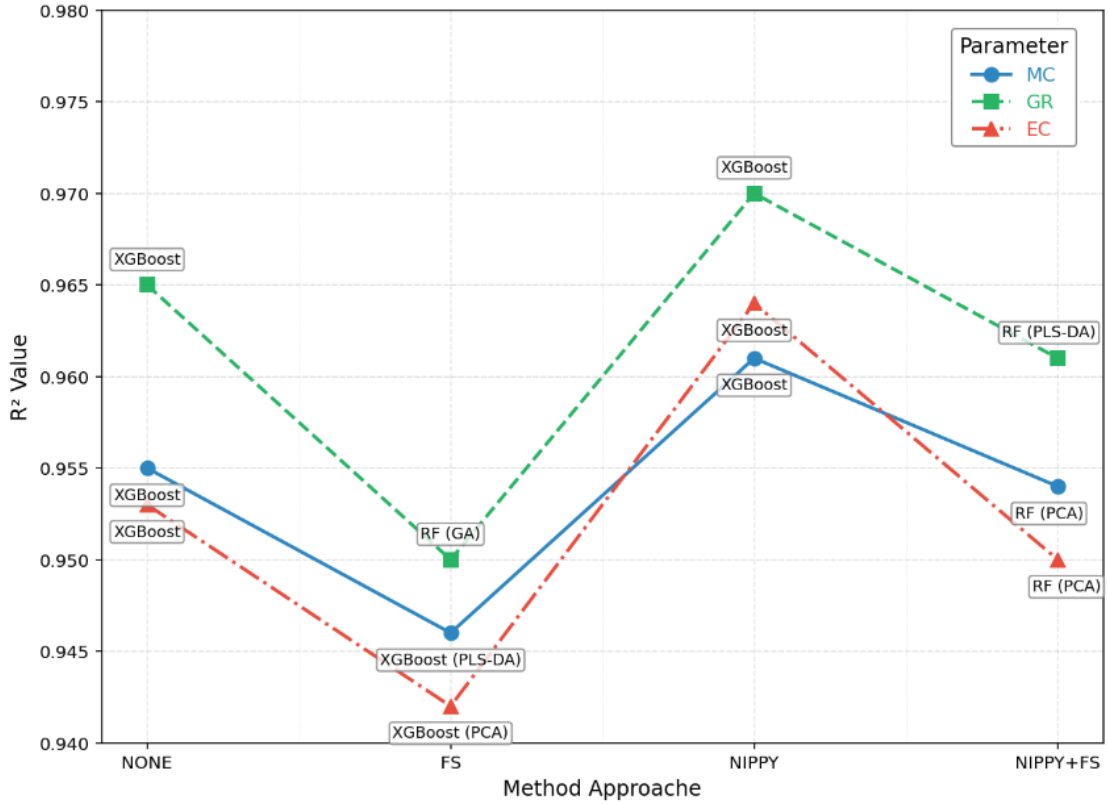

Fig. 13. Comparison of the highest R^2 values between approaches (NONE, FS, NIPPY, and NIPPY+FS) for soybean seed quality prediction models

4.4. Implementation of the Adaptive Neuro-Fuzzy Inference System (ANFIS)

The integration of Adaptive Neuro-Fuzzy Inference System (ANFIS) as the classification layer represents a key innovation in our hybrid framework, addressing the interpretability limitations of conventional machine learning models while maintaining high predictive accuracy. ANFIS was employed as the final stage to classify soybean seed quality based on the predicted physiological parameters (MC, GR, EC) generated by the XGBoost regression model using Nippy spectral transformation. This two-stage architecture XGBoost for regression followed by ANFIS for classification combines the strengths of both approaches: XGBoost provides high-accuracy numerical predictions from complex spectral data, while ANFIS offers transparent, rule-based classification that aligns with agronomic expertise.

The model was developed using Gaussian Membership Functions (GMF) with three linguistic terms (Low, Medium, High) for each input variable, generating 27 fuzzy rules (3^3) through grid partitioning. To ensure rigorous validation and address potential overfitting concerns, we implemented a comprehensive evaluation strategy: (1) 5-fold cross-validation on the original 800 samples, (2) holdout testing with a separate 20% test set, and (3) external validation using an independent batch of 100 soybean seeds from different harvest seasons. Prior to training, all data

were normalized using Min-Max scaling to maintain uniform value ranges. Performance evaluation metrics included R^2, RMSE, MAE, and classification accuracy.

The evaluation results (Table 13) demonstrated that ANFIS achieved exceptional performance with 100% classification accuracy on the test set and maintained 98% accuracy on external validation. The average metrics across folds were $R^2 = 0.9996 \pm 0.0002$, RMSE = 0.0148 ± 0.0036, and MAE = 0.0094 ± 0.0010. The external validation result of 98% accuracy confirms that the model generalizes well to unseen data and mitigates concerns about overfitting, as it performs consistently across different seed batches and measurement conditions.

ANFIS proves particularly effective for this application due to three key factors: (1) Interpretable Rule-Based Reasoning: The fuzzy rules directly link physiological parameters to quality categories, providing transparent decision-making. For example, Rule 1: "IF MC is Low AND GR is High AND EC is Low THEN Premium" reflects established seed physiology where low moisture prevents fungal growth, high germination indicates viability, and low conductivity signifies intact cell membranes. (2) Physiological Significance of Spectral Differences: The spectral transformations and XGBoost regression effectively extract meaningful physiological information from reflectance data. (3) Robust Handling of Spectral Variance: ANFIS manages measurement variability through Gaussian membership functions that allow gradual transitions between quality classes [97]. This is physiologically appropriate as seed quality degradation occurs along a continuum. The system accommodates spectral variations from factors like seed coat texture, size heterogeneity, and minor surface irregularities by modeling parameter distributions rather than relying on threshold cutoffs.

Table 14 presents a subset of the fuzzy rules generated by ANFIS, demonstrating how the system bridges machine learning predictions with agronomic knowledge. These rules provide explainable insights that are crucial for agricultural adoption, where understanding the "why" behind classifications is as important as accuracy. The novelty of this XGBoost-ANFIS hybrid approach lies in its synergistic combination: XGBoost excels at extracting complex patterns from high-dimensional spectral data, while ANFIS provides the interpretability needed for agricultural decision-making. Unlike black-box models that offer only predictions, the system provides both accurate classifications and understandable rules that reflect seed physiology. This addresses a critical gap in precision agriculture, where trust and transparency are essential for technology adoption.

Fig. 14 illustrates the strong correlation between actual and predicted values, with data points clustering closely around the y=x line. The minimal prediction errors (RMSE < 0.02) across all parameters confirm the model's precision. Importantly, the system maintains performance despite spectral variance because the Gaussian membership functions naturally accommodate parameter distributions, and the fuzzy rules capture the continuous nature of seed quality degradation. The ANFIS component of hybrid framework successfully translates accurate regression predictions into interpretable quality classifications while maintaining robustness against spectral variability. The external validation results confirm generalizability, addressing concerns about overfitting.

Table 13. Performance evaluation results of the ANFIS model

Fold	R^2	RMSE	MAE	MSE	Accuracy (%)
1	0.9995	0.0174	0.0093	0.0003	100.00
2	0.9998	0.0118	0.0088	0.0001	100.00
3	0.9998	0.0109	0.0083	0.0001	100.00
4	0.9993	0.0204	0.0112	0.0004	100.00
5	0.9997	0.0135	0.0097	0.0002	100.00
AVG ± SD	0.9996 ± 0.0002	0.0148 ± 0.0036	0.0094 ± 0.0010	0.0002 ± 0.0001	100.00 ± 0.00

4.5. Comparison with Previous Studies and Practical Implications

To scientifically contextualize the contributions of this study, a comprehensive comparison was conducted with recent and seminal works in spectroscopic seed quality assessment, as summarized in Table 15. The hybrid XGBoost-ANFIS approach not only achieves superior predictive accuracy

but also introduces critical advancements in interpretability and field deployability, addressing longstanding gaps in agricultural sensing technology.

Table 14. Sample fuzzy rules from ANFIS classification system

Rule	MC	GR	EC	Output	Physiological Interpretation
1	Low	High	Low	Premium	Optimal seed physiology: preserved membranes, high viability
2	Medium	High	Medium	Fair	Viable but with minor physiological compromise
3	High	Low	High	Poor	Severe deterioration: membrane damage, reduced viability
4	Low	Medium	Low	Fair	Good preservation but suboptimal germination potential
5	High	Medium	High	Poor	Moisture stress affecting membrane integrity

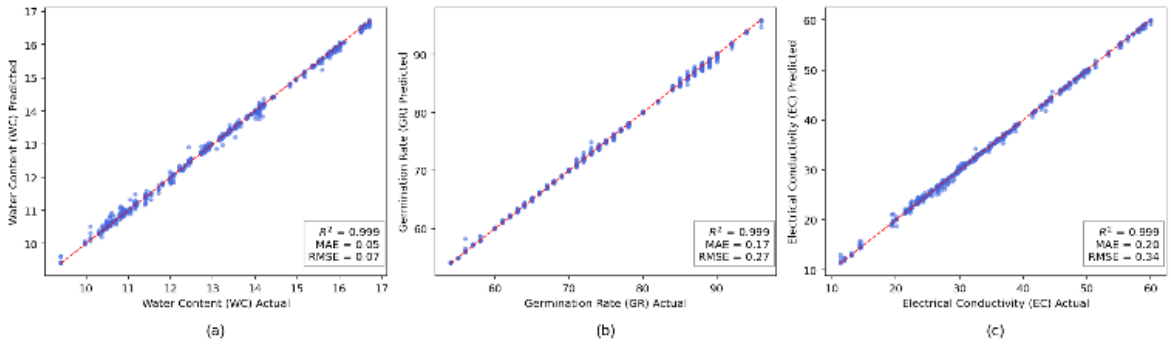

Fig. 14. Scatter plot between actual values and ANFIS prediction results for parameters (a) MC, (b) GR, and (c) EC

The comparative analysis reveals three key scientific advancements of this work: First, whereas previous studies predominantly rely on single-algorithm approaches (PLS-DA, PLS, RF) that either lack accuracy or interpretability, hybrid XGBoost-ANFIS framework synergistically combines high-accuracy regression with transparent, rule-based classification. Second, while existing portable spectrometers often sacrifice model sophistication for portability, our system embeds a complete machine-learning pipeline (preprocessing, regression, classification) within a low-cost, field-ready device. Third, unlike studies focusing on single parameters or genetically modified varieties, demonstrate simultaneous prediction of three key physiological parameters (MC, GR, EC) on local non-GMO soybean varieties, addressing a critical need for regionally adapted agricultural technology.

The indirect classification approach, spectral transformation → XGBoost regression → ANFIS classification—proved particularly effective because it decomposes the complex problem into two optimized stages: (1) accurate numerical prediction of physiological parameters using a state-of-the-art ensemble regressor, and (2) interpretable categorization using a fuzzy inference system that mimics expert reasoning. This architecture outperforms both direct spectral classification (which often struggles with high-dimensional, collinear spectral data) and conventional expert thresholding (which is slow, destructive, and subjective).

Regarding practical implications, the system offers transformative potential for seed certification, breeding programs, and farm-gate quality control. By providing rapid (seconds per sample), non-destructive, and interpretable assessments, it can reduce reliance on centralized laboratories, minimize post-harvest losses, and support precision agriculture initiatives. However, several implementation challenges must be acknowledged: (1) Variety specificity – the model was optimized for Anjasmoro and Grobogan soybeans and requires calibration for other varieties; (2) Environmental sensitivity – field conditions (ambient light, temperature, humidity) necessitate periodic recalibration; (3) Scale-up logistics – large-scale deployment would require robust device management, cloud-based model updating, and integration with farm management software.

To address these challenges and guide future research, propose the following directions: (1) Development of transfer-learning protocols to adapt the model to new varieties and crops with

minimal additional data; (2) Integration of environmental sensors (temperature, humidity, ambient light) for adaptive calibration; (3) Implementation of edge-cloud hybrid architectures where lightweight models run on-device while complex retraining occurs in the cloud; (4) Expansion to multi-crop quality assessment for broader agricultural impact.

This study advances the field of agricultural spectroscopy by demonstrating a fully integrated, portable, and interpretable system for non-destructive seed quality assessment. The hybrid XGBoost-ANFIS framework sets a new benchmark for accuracy while providing the transparency needed for adoption in real-world agricultural decision-making. By bridging the gap between high-performance machine learning and practical field deployment, this work contributes significantly to the digital transformation of agriculture and the pursuit of sustainable food systems.

Table 15. Comparative analysis with previous spectroscopic seed quality assessment studies

Study	Method	Crop	Key Findings/Limitations	Reported Accuracy	Major Advancement of Our Work
[98]	PLS	Soybean	Early demonstration of NIR for vigor prediction; limited to lab settings and linear models.	89.5%	Non-linear hybrid modeling (vs. linear PLS-DA) and portable field deployment (vs. benchtop).
[99]	PLS-DA	Soybean	Non-linear kernel improved vigor classification; model remains a black-box with limited physiological insight.	92.3%	Interpretable fuzzy rule system (ANFIS) that links predictions to seed physiology.
[100]	Ensemble (RF, GBM)	Soybean	Polarized hyperspectral imaging enhanced feature extraction; requires complex, expensive instrumentation.	94.1%	Cost-effective portable Vis/NIR with automated preprocessing, achieving comparable accuracy without complex hardware.
[101]	GPR	Maize	Rapid germination detection using NIR; focused on single parameter (germination) without multi-parameter integration.	~93%	Simultaneous multi parameter prediction (MC, GR, EC) within a unified regression - classification pipeline.
This Study	XGBoost-ANFIS Hybrid	Soybean	Portable embedded system; interpretable fuzzy-rule classification; automated spectral optimization (Nippy); validated on non-GMO local varieties.	100% (test) / 98% (external)	First integrated portable system combining high accuracy regression (XGBoost) with interpretable fuzzy classification (ANFIS) for real-time seed grading.

5. Conclusion

This study introduced an indirect classification framework integrated within an embedded Vis/NIR spectroscopy system for non-destructive soybean seed quality assessment. The main scientific contributions of this research are threefold: (1) development of the first portable embedded system combining Vis/NIR spectroscopy with hybrid machine learning for real-time seed quality evaluation, (2) demonstration of superior performance of the XGBoost-ANFIS hybrid approach compared to traditional chemometric methods (PLS, SVR) with accuracy improvements of 1401-1653%, and (3) creation of an interpretable fuzzy rule system that bridges machine learning predictions with agronomic expertise.

The proposed approach combined spectral transformation, machine learning regression, and ANFIS-based decision logic to predict and classify key physiological quality parameters (MC, GR, EC). The system achieved 100% classification accuracy on the test set and maintained 98% accuracy in external validation, demonstrating robust generalization capability beyond the training dataset.

The system architecture was designed to automate data acquisition, preprocessing, and model execution within a portable embedded platform, enabling efficient on-site evaluation and control. ANFIS proved particularly effective in this application because it translates continuous regression outputs into interpretable fuzzy rules that reflect physiological seed characteristics, such as "IF MC is Low AND GR is High AND EC is Low THEN Premium," which corresponds to optimal seed conditions with low moisture, high viability, and intact cellular membranes. Validation using 800 spectral datasets showed that the hybrid system achieved excellent agreement with expert-based threshold classification, with superior robustness and minimal prediction error. Compared to previous studies on spectroscopic seed quality assessment, the approach offers both higher accuracy and the unique advantages of portability and interpretability. Spectral preprocessing improved model stability, while feature selection enhanced the efficiency of MLP and LR models. The integration of ANFIS enabled adaptive nonlinear reasoning, ensuring reliable classification even under spectral variability.

However, several limitations should be acknowledged for practical deployment: (1) the model is currently optimized for specific soybean varieties (Anjasmoro and Grobogan) and may require retraining for other varieties, (2) field conditions such as temperature fluctuations and ambient light variations necessitate periodic calibration, and (3) the system's power consumption (8-hour battery life) may limit continuous operation in remote areas. Future research should focus on developing adaptive calibration techniques, multi-variety models, and integration with IoT platforms for large-scale agricultural deployment.

Overall, this research demonstrates the feasibility of embedding advanced hybrid learning models into a controlled Vis/NIR spectroscopy platform, bridging the gap between optical sensing and expert-driven decision systems. The key innovation lies not merely in combining XGBoost and ANFIS, which has been done in other domains but in their novel application within a portable embedded system for non-destructive seed quality assessment, providing both numerical accuracy and expert-like interpretability for agricultural decision-making. The developed framework provides a foundation for real-time intelligent seed grading and automated quality control, highlighting its strong potential for scalable field deployment and agricultural digitalization. Future work will explore (1) expansion to other economically important seeds (rice, maize, wheat), (2) development of cloud-based model updating for continuous improvement, and (3) integration with blockchain technology for traceability in seed certification processes.

Author Contribution: All authors contributed to the main contributor to this paper. All authors read and approved the final paper.

Funding: This research was funded by Directorate of Research and Community Service, Directorate General of Research and Development, Ministry of Higher Education, Science, and Technology, under Contract No. 125/C3/DT.05.00/PL/2025.

Acknowledgment: This research was supported by the Research Grant Program of 2025 from the Directorate of Research and Community Service, Directorate General of Research and Development, Ministry of Higher Education, Science, and Technology (Regular Fundamental Grant). The authors would like to express their sincere gratitude to the soybean experts from Universitas Gadjah Mada (UGM) and the Central Seed Supervision and Certification Agency (Balai Besar PPMBTPH) for their valuable insights, technical support, and constructive discussions that greatly contributed to the success of this study.

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

References

[1] G. M. Barboza Martignone, B. Ghosh, D. Papadas, and K. Behrendt, “The rise of Soybean in international commodity markets: A quantile investigation,” *Heliyon*, vol. 10, no. 15, p. e34669, 2024, https://doi.org/10.1016/j.heliyon.2024.e34669.

[2] P. Qin, T. Wang, and Y. Luo, "A review on plant-based proteins from soybean: Health benefits and soy product development," *Journal of Agriculture and Food Research*, vol. 7, p. 100265, 2022, https://doi.org/10.1016/j.jafr.2021.100265.

[3] J. R. Toni, L. L. Radünz, L. Galon, R. G. Dionello, and M. A. Scariot, "Quality assessment of soybean seeds submitted to industrial seed treatment and stored in a natural and controlled environment," *Journal of Stored Products Research*, vol. 108, p. 102372, 2024, https://doi.org/10.1016/j.jspr.2024.102372.

[4] G. S. Montanha *et al*., "Profile of mineral nutrients and proteins in soybean seeds (Glycine max (L.) Merrill): Insights from 95 varieties cultivated in Brazil," *Journal of Food Composition and Analysis*, vol. 134, p. 106536, 2024, https://doi.org/10.1016/j.jfca.2024.106536.

[5] F. Rozi *et al*., "Indonesian foodstuffs in facing global food crisis: Economic aspects of soybean farming," *Journal of Agriculture and Food Research*, vol. 19, p. 101669, 2025, https://doi.org/10.1016/j.jafr.2025.101669.

[6] R. Aulia *et al*., "Non-destructive prediction of protein contents of soybean seeds using near-infrared hyperspectral imaging," *Infrared Physics & Technology*, vol. 127, p. 104365, 2022, https://doi.org/10.1016/j.infrared.2022.104365.

[7] L. Zhang, L. Sun, X. Jin, X. Zhao, and S. Li, "DAFFnet: Seed classification of soybean variety based on dual attention feature fusion networks," *The Crop Journal*, vol. 13, no. 2, pp. 619-629, 2025, https://doi.org/10.1016/j.cj.2024.12.023.

[8] M. Tuite, S. Liu, W. Phippen, M. Starke, and R. Chopra, "Rapid and non-destructive screening of seed components in domesticated pennycress using near-infrared spectroscopy," *Industrial Crops and Products*, vol. 235, p. 121752, 2025, https://doi.org/10.1016/j.indcrop.2025.121752.

[9] A. Griffo *et al*., "Application of machine learning models for non-invasive seed quality detection," *Current Plant Biology*, vol. 44, p. 100557, 2025, https://doi.org/10.1016/j.cpb.2025.100557.

[10] A. J. S. Neto and D. de C. Lopes, "Chemometrics coupled with near infrared spectroscopy for detecting adulteration levels in herbal teas," *Journal of Food Composition and Analysis*, vol. 135, p. 106637, 2024, https://doi.org/10.1016/j.jfca.2024.106637.

[11] X. Han, C. Zhengguang, and L. Jinming, "Rapid detection of maize seed germination using near-infrared spectroscopy combined with Gaussian process regression," *Food Chemistry*, vol. 491, p. 145254, 2025, https://doi.org/10.1016/j.foodchem.2025.145254.

[12] H. Guan, Y. Zhao, X. Ma, J. Yang, and Y. Huang, "A detection method of Saposhnikovia divaricata seed vigor based on near-infrared spectral feature extraction," *Infrared Physics & Technology*, vol. 141, p. 105463, 2024, https://doi.org/10.1016/j.infrared.2024.105463.

[13] Z. Wang, W. Huang, J. Li, S. Liu, and S. Fan, "Assessment of protein content and insect infestation of maize seeds based on on-line near-infrared spectroscopy and machine learning," *Computers and Electronics in Agriculture*, vol. 211, p. 107969, 2023, https://doi.org/10.1016/j.compag.2023.107969.

[14] M. Al-Amery *et al*., "Near-infrared spectroscopy used to predict soybean seed germination and vigour," *Seed Science Research*, vol. 28, no. 3, pp. 245-252, 2018, https://doi.org/10.1017/S0960258518000119.

[15] M. F. da Silva *et al*., "Near infrared spectroscopy for the classification of vigor level of soybean seed," *Revista Ciência Agronômica*, vol. 55, 2024, https://doi.org/10.5935/1806-6690.20240005.

[16] P. Mishra and R. Nikzad-Langerodi, "Partial least square regression versus domain invariant partial least square regression with application to near-infrared spectroscopy of fresh fruit," *Infrared Physics & Technology*, vol. 111, p. 103547, 2020, https://doi.org/10.1016/j.infrared.2020.103547.

[17] D. Xenitopoulou, N. L. Tsakiridis, A. P. Zalidis, and G. C. Zalidis, "Real-time detection of turmeric adulteration with metanil yellow using a miniaturized NIR sensor and AI techniques," *Future Foods*, vol. 12, p. 100695, 2025, https://doi.org/10.1016/j.fufo.2025.100695.

[18] J. E. Herranz-Luque *et al*., "Monitoring Land Management Practices Using Vis–NIR Spectroscopy Provides Insights into Predicting Soil Organic Carbon and Limestone Levels in Agricultural Plots," *Agronomy*, vol. 14, no. 6, p. 1150, 2024, https://doi.org/10.3390/agronomy14061150.

[19] H. Seki, H. Murakami, T. Ma, S. Tsuchikawa, and T. Inagaki, "Evaluating Soluble Solids in White Strawberries: A Comparative Analysis of Vis-NIR and NIR Spectroscopy," *Foods*, vol. 13, no. 14, p. 2274, 2024, https://doi.org/10.3390/foods13142274.

[20] S. K. Shin, S. J. Lee, and J. H. Park, "Prediction of Soil Properties Using Vis-NIR Spectroscopy Combined with Machine Learning: A Review," *Sensors*, vol. 25, no. 16, p. 5045, 2025, https://doi.org/10.3390/s25165045.

[21] N. Zhou, J. Hong, B. Song, S. Wu, Y. Wei, and T. Wang, "Feature Variable Selection Based on VIS-NIR Spectra and Soil Moisture Content Prediction Model Construction," *Journal of Spectroscopy*, vol. 2024, no. 1, pp. 1-16, 2024, https://doi.org/10.1155/2024/8180765.

[22] J. Torniainen, I. O. Afara, M. Prakash, J. K. Sarin, L. Stenroth, and J. Töyräs, "Open-source python module for automated preprocessing of near infrared spectroscopic data," *Analytica Chimica Acta*, vol. 1108, pp. 1-9, 2020, https://doi.org/10.1016/j.aca.2020.02.030.

[23] A. Khumaidi, Y. A. Purwanto, H. Sukoco, and S. H. Wijaya, "Using Fuzzy Logic to Increase Accuracy in Mango Maturity Index Classification: Approach for Developing a Portable Near-Infrared Spectroscopy Device," *Sensors*, vol. 22, no. 24, p. 9704, 2022, https://doi.org/10.3390/s22249704.

[24] SparkFun, "AS7265x Datasheet," *AMS Datasheet*, vol. 1, pp. 1-63, 2018, https://www.alldatasheet.com/view.jsp?Searchword=AS7265X.

[25] A. Khumaidi, R. Raafi'udin, and N. S. Triastuti, "Enhancing Ship Coating Quality Detection via Machine Learning-Optimized Visible Near-Infrared Spectroscopy," *Instrumentation Mesure Métrologie*, vol. 23, no. 6, pp. 441-450, 2024, https://doi.org/10.18280/i2m.230604.

[26] J. Mankar, C. Darode, K. Trivedi, M. Kanoje, and P. Shahare, "Review of I2C Protocol," *International Journal of Researches in Adventure Technology*, vol. 2, no. 1, pp. 2321-9637, 2014, https://ijrat.org/downloads/Vol-2/jan-2014/paper%20ID-21201448.pdf.

[27] A. Das, "Portable UV–visible spectroscopy–instrumentation, technology, and applications," *Portable Spectroscopy and Spectrometry*, pp. 179-207, 2021, https://doi.org/10.1002/9781119636489.ch8.

[28] Balai Pengujian Standar Instrumen Tanaman Aneka Kacang, "Deskripsi Varietas Unggul Kedelai 1918-2022," *Balai Pengujian Standar Instrumen Tanaman Aneka Kacang*, 2024, https://repository.pertanian.go.id/handle/123456789/25604.

[29] B. Guo *et al*., "Soybean genetic resources contributing to sustainable protein production," *Theoretical and Applied Genetics*, vol. 135, no. 11, pp. 4095-4121, 2022, https://doi.org/10.1007/s00122-022-04222-9.

[30] G. F. da Silva *et al*., "Physiological Quality of Soybean Seeds as a Function of Soil Management Systems and Pre-Harvest Desiccation," *Agronomy*, vol. 13, no. 3, p. 847, 2023, https://doi.org/10.3390/agronomy13030847.

[31] D. C. Santana *et al*., "High-throughput phenotyping using VIS/NIR spectroscopy in the classification of soybean genotypes for grain yield and industrial traits," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 310, p. 123963, 2024, https://doi.org/10.1016/j.saa.2024.123963.

[32] R. E. Masithoh, F. R. Abadi, L. Sutiarso, and S. Rahayoe, "Evaluation of Indonesian Local Soybean Based on Chemical Characteristics And Visible - Near Infrared Spectra With Chemometrics," *Biotropia*, vol. 31, no. 1, pp. 63-75, 2024, https://doi.org/10.11598/btb.2024.31.1.2054.

[33] ISTA, "Rules Proposals for the International Rules for Seed Testing 2024 Edition," *ISTA Secretariat*, 2024, https://www.seedtest.org/api/rm/24EMJ3A9EEZ4NGU/ogm23-05-rules-proposals-for-ista-2024-edition-v12.pdf

[34] O. L. Justice and L. N. Bass, "Principles and practices of seed storage," *Washington: U.S. Departement of Agriculture*, 1979, https://www.govinfo.gov/content/pkg/GOVPUB-A-PURL-gpo28758/pdf/GOVPUB-A-PURL-gpo28758.pdf.

[35] I. C. de Oliveira *et al*., "Differentiation of Soybean Genotypes Concerning Seed Physiological Quality Using Hyperspectral Bands," *AgriEngineering*, vol. 6, no. 4, pp. 4752-4765, 2024, https://doi.org/10.3390/agriengineering6040272.

[36] Q. Hao *et al*., "Evaluation of seed vigor in soybean germplasms from different eco-regions," *Oil Crop Science*, vol. 5, no. 1, pp. 22-25, 2020, https://doi.org/10.1016/j.ocsci.2020.03.006.

[37] W. Ding *et al*., "Development of a comprehensive evaluation system and models to determine soybean seed vigor," *Industrial Crops and Products*, vol. 224, p. 120329, 2025, https://doi.org/10.1016/j.indcrop.2024.120329.

[38] L. Kong, C. Wu, H. Li, M. Yuan, and T. Sun, "Discrimination of tea seed oil adulteration based on near-infrared spectroscopy and combined preprocessing method," *Journal of Food Composition and Analysis*, vol. 134, p. 106560, 2024, https://doi.org/10.1016/j.jfca.2024.106560.

[39] X. Chen *et al*., "An automated preprocessing framework for near infrared spectroscopic data," *Chemometrics and Intelligent Laboratory Systems*, vol. 267, p. 105542, 2025, https://doi.org/10.1016/j.chemolab.2025.105542.

[40] S. Wang, M. Lin, Y. Meng, T. Jiang, F. Fan, and S. Wang, "Self-expansion full information optimization strategy: Convenient and efficient method for near infrared spectrum auto-analysis," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 303, p. 123224, 2023, https://doi.org/10.1016/j.saa.2023.123224.

[41] P. Mishra, J. M. Roger, D. N. Rutledge, and E. Woltering, "SPORT pre-processing can improve near-infrared quality prediction models for fresh fruits and agro-materials," *Postharvest Biology and Technology*, vol. 168, p. 111271, 2020, https://doi.org/10.1016/j.postharvbio.2020.111271.

[42] A. Khumaidi and R. Raafi'udin, "Effects of Oversampling Smote and Spectral Transformations in the Classification of Mango Cultivars Using Near-Infrared Spectroscopy," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 12, no. 3, p. 1047, 2022, https://doi.org/10.18517/ijaseit.12.3.16001.

[43] X. Chen and D. Wang, "Baseline correction of near-fault ground motion records based on the hilbert spectral analysis," *Soil Dynamics and Earthquake Engineering*, vol. 154, p. 107162, 2022, https://doi.org/10.1016/j.soildyn.2022.107162.

[44] R. J. Barnes, M. S. Dhanoa, and S. J. Lister, "Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra," *Applied Spectroscopy*, vol. 43, no. 5, pp. 772-777, 1989, https://doi.org/10.1366/0003702894202201.

[45] W. Cardoso, J. V. Roque, J. J. Jansen, S. Y. Teng, and R. F. Teófilo, "Combinatorial Order Pre-processing Search (COPS): A new pre-processing strategy for large-scale interpretable data analysis in process analytical technologies," *Computers & Chemical Engineering*, vol. 192, p. 108892, 2025, https://doi.org/10.1016/j.compchemeng.2024.108892.

[46] A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures.," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627-1639, 1964, https://doi.org/10.1021/ac60214a047.

[47] J. Gerretzen *et al*., "Boosting model performance and interpretation by entangling preprocessing selection and variable selection," *Analytica Chimica Acta*, vol. 938, pp. 44-52, 2016, https://doi.org/10.1016/j.aca.2016.08.022.

[48] M. B. Almoujahed *et al*., "Non-destructive detection of fusarium head blight in wheat kernels and flour using visible near-infrared and mid-infrared spectroscopy," *Chemometrics and Intelligent Laboratory Systems*, vol. 245, p. 105050, 2024, https://doi.org/10.1016/j.chemolab.2023.105050.

[49] Y. Zhu, L. Chen, X. Chen, J. Chen, and H. Zhang, "Near-infrared spectroscopy identification method of cashmere and wool fibers based on an optimized wavelength selection algorithm," *Heliyon*, vol. 10, no. 14, p. e34537, 2024, https://doi.org/10.1016/j.heliyon.2024.e34537.

[50] H. Li, J. Cui, X. Zhang, Y. Han, and L. Cao, "Dimensionality Reduction and Classification of Hyperspectral Remote Sensing Image Feature Extraction," *Remote Sensing*, vol. 14, no. 18, p. 4579, 2022, https://doi.org/10.3390/rs14184579.

[51] K. Kumar, "Partial Least Square (PLS) Analysis," *Resonance*, vol. 26, no. 3, pp. 429–442, 2021, https://doi.org/10.1007/s12045-021-1140-1.

[52] S. N. Alsaqri *et al*., "Rapid detection of pork gelatin in ice cream samples by using non-destructive FT-NIR spectroscopy and Partial least squares-discriminant analysis," *Food Chemistry Advances*, vol. 2, p. 100215, 2023, https://doi.org/10.1016/j.focha.2023.100215.

[53] L. L. de Souza, D. N. C. Candeias, E. D. T. Moreira, P. H. G. D. Diniz, V. H. Springer, and D. D. de S. Fernandes, "UV–Vis spectralprint-based discrimination and quantification of sugar syrup adulteration in honey using the Successive Projections Algorithm (SPA) for variable selection," *Chemometrics and Intelligent Laboratory Systems*, vol. 257, p. 105314, 2025, https://doi.org/10.1016/j.chemolab.2024.105314.

[54] A. Asghari, M. K. Khorrami, and A. B. Garmarudi, "Comparison between partial least square and support vector regression with a genetic algorithm wavelength selection method for the simultaneous determination of some oxygenate compounds in gasoline by FTIR spectroscopy," *Infrared Physics & Technology*, vol. 105, p. 103177, 2020, https://doi.org/10.1016/j.infrared.2019.103177.

[55] H. J. Luinge, J. H. van der Maas, and T. Visser, "Partial least squares regression as a multivariate tool for the interpretation of infrared spectra," *Chemometrics and Intelligent Laboratory Systems*, vol. 28, no. 1, pp. 129-138, 1995, https://doi.org/10.1016/0169-7439(95)80045-B.

[56] H. Yin *et al*., "Detection of moisture content and size of pumpkin seeds based on hyperspectral reflection and transmission imaging techniques," *Journal of Food Composition and Analysis*, vol. 124, p. 105651, 2023, https://doi.org/10.1016/j.jfca.2023.105651.

[57] S. D. Daba, D. Honigs, R. J. McGee, and A. M. Kiszonas, "Prediction of Protein Concentration in Pea (Pisum sativum L.) Using Near-Infrared Spectroscopy (NIRS) Systems," *Foods*, vol. 11, no. 22, p. 3701, 2022, https://doi.org/10.3390/foods11223701.

[58] A. Schneider, G. Hommel, and M. Blettner, "Linear Regression Analysis," *Deutsches Ärzteblatt International*, vol. 107, no. 44, 776-782, 2010, https://doi.org/10.3238/arztebl.2010.0776.

[59] L. Mihaly Cozmuta, "The application of multiple linear regression methods to FTIR spectra of fingernails for predicting gender and age of human subjects," *Heliyon*, vol. 11, no. 4, p. e42815, 2025, https://doi.org/10.1016/j.heliyon.2025.e42815.

[60] N. Tadmor Shalev, A. Ghermandi, D. Tchernov, E. Shemesh, A. Israel, and A. Brook, "NIR spectroscopy and artificial neural network for seaweed protein content assessment in-situ," *Computers and Electronics in Agriculture*, vol. 201, p. 107304, 2022, https://doi.org/10.1016/j.compag.2022.107304.

[61] E. M. Hameed, H. Joshi, and Q. K. Kadhim, "Advancements in Artificial Intelligence Techniques for Diabetes Prediction: A Comprehensive Literature Review," *Journal of Robotics and Control*, vol. 6, no. 1, pp. 345-365, 2025, https://doi.org/10.18196/jrc.v6i1.22258.

[62] Y. R. Shahare, M. P. Singh, S. P. Singh, P. Singh, and M. Diwakar, "ASUR: Agriculture Soil Fertility Assessment Using Random Forest Classifier and Regressor," *Procedia Computer Science*, vol. 235, pp. 1732–1741, 2024, https://doi.org/10.1016/j.procs.2024.04.164.

[63] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, 2016, https://doi.org/10.1145/2939672.2939785.

[64] M. Alizamir *et al*., "An interpretable XGBoost-SHAP machine learning model for reliable prediction of mechanical properties in waste foundry sand-based eco-friendly concrete," *Results in Engineering*, vol. 25, p. 104307, 2025, https://doi.org/10.1016/j.rineng.2025.104307.

[65] R. Liang *et al*., "Interpretable machine learning assisted spectroscopy for fast characterization of biomass and waste," *Waste Management*, vol. 160, pp. 90-100, 2023, https://doi.org/10.1016/j.wasman.2023.02.012.

[66] L. Hssaini, "ML-driven olive oil quality prediction: Comparative evaluation of FTMIR data preprocessing techniques using RF and XGBoost models in multi-stage validation," *Measurement: Food*, vol. 19, p. 100249, 2025, https://doi.org/10.1016/j.meafoo.2025.100249.

[67] S. Chang *et al*., "Development of a Multiscale XGBoost-Based Model for Enhanced Detection of Potato Late Blight Using Sentinel-2, UAV, and Ground Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1-14, 2024, https://doi.org/10.1109/TGRS.2024.3466648.

[68] S. Jeong, Y. Kim, S. H. Hur, H. Bang, H. Kim, and H. Chung, "Explainable extreme gradient boosting as a machine learning tool for discrimination of the geographical origin of chili peppers using laser ablation-inductively coupled plasma mass spectrometry, X-ray fluorescence, and near-infrared spectroscopy," *Journal of Agriculture and Food Research*, vol. 18, p. 101446, 2024, https://doi.org/10.1016/j.jafr.2024.101446.

[69] D. Tarwidi, S. R. Pudjaprasetya, D. Adytia, and M. Apri, "An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach," *MethodsX*, vol. 10, p. 102119, 2023, https://doi.org/10.1016/j.mex.2023.102119.

[70] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295-316, 2020, https://doi.org/10.1016/j.neucom.2020.07.061.

[71] Q. A. Hidayaturrohman and E. Hanada, "A Comparative Analysis of Hyper-Parameter Optimization Methods for Predicting Heart Failure Outcomes," *Applied Sciences*, vol. 15, no. 6, p. 3393, 2025, https://doi.org/10.3390/app15063393.

[72] S. Ünalan, O. Günay, I. Akkurt, K. Gunoglu, and H. O. Tekin, "A comparative study on breast cancer classification with stratified shuffle split and K-fold cross validation via ensembled machine learning," *Journal of Radiation Research and Applied Sciences*, vol. 17, no. 4, p. 101080, 2024, https://doi.org/10.1016/j.jrras.2024.101080.

[73] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, p. e623, 2021, https://doi.org/10.7717/peerj-cs.623.

[74] S. Bhattacharya, K. Kalita, R. Čep, and S. Chakraborty, "A Comparative Analysis on Prediction Performance of Regression Models during Machining of Composite Materials," *Materials*, vol. 14, no. 21, p. 6689, 2021, https://doi.org/10.3390/ma14216689.

[75] A. Dey and U. Sarma, "Performance Evaluation of Machine Learning Regression Algorithms for Soil Nitrogen Estimation," *2023 IEEE Fifth International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, pp. 1-5, 2023, https://doi.org/10.1109/ICAECC59324.2023.10560157.

[76] A. Vatanshenas and T. T. Länsivaara, "Estimating maximum shear modulus (G) using adaptive neuro-fuzzy inference system (ANFIS)," *Soil Dynamics and Earthquake Engineering*, vol. 153, p. 107105, 2022, https://doi.org/10.1016/j.soildyn.2021.107105.

[77] J.-S. R. Jang and C.-T. Sun, E. Mizutani, "Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence," *USA: Prentice-Hall*, 1996, http://www.soukalfi.edu.sk/01_NeuroFuzzyApproach.pdf#.

[78] M. Dahria, S. Defit, and Y. Yuhandri, "Development of Euclidean Distance Algorithm for ANFIS Optimization in IoT-based Pond Water Quality Prediction," *Journal of Robotics and* Control, vol. 6, no. 4, pp. 1777-1789, 2025, https://doi.org/10.18196/jrc.v6i4.26497.

[79] P. Mangena, "Analysis of correlation between seed vigour, germination and multiple shoot induction in soybean (Glycine max L. Merr.)," *Heliyon*, vol. 7, no. 9, p. e07913, 2021, https://doi.org/10.1016/j.heliyon.2021.e07913.

[80] A. Koskosidis, E. M. Khah, O. I. Pavli, and D. N. Vlachostergios, "Effect of storage conditions on seed quality of soybean (Glycine max L.) germplasm," *AIMS Agriculture and Food*, vol. 7, no. 2, pp. 387-402, 2022, https://doi.org/10.3934/agrfood.2022025.

[81] C. Zhou *et al*., "Soybean plants enhance growth through metabolic regulation under heterogeneous drought stress," *Agricultural Water Management*, vol. 303, p. 109029, 2024, https://doi.org/10.1016/j.agwat.2024.109029.

[82] X. Li, K. Liu, S. Rideout, L. Rosso, B. Zhang, and G. E. Welbaum, "Seed physiological traits and environmental factors influence seedling establishment of vegetable soybean (Glycine max L.)," *Frontiers in Plant Science*, vol. 15, 2024, https://doi.org/10.3389/fpls.2024.1344895.

[83] R. E. Lima, P. C. Coradi, D. M. Rodrigues, P. E. Teodoro, L. P. R. Teodoro, and D. P. de Oliveira, "Monitoring and predicting the quality of soybeans for different drying and storage technologies on a real

scale using sensors and Machine Learning models," *Journal of Stored Products Research*, vol. 108, p. 102386, 2024, https://doi.org/10.1016/j.jspr.2024.102386.

[84] M. A. Hussain *et al*., "Comparative analysis of physiological variations and genetic architecture for cold stress response in soybean germplasm," *Frontiers in Plant Science*, vol. 13, 2023, https://doi.org/10.3389/fpls.2022.1095335.

[85] S. Günaydın, N. Çetin, C. Sağlam, K. Sacilik, and A. Jahanbakhshi, "Comparative analysis of visible and near-infrared (Vis-NIR) spectroscopy and prediction of moisture ratio using machine learning algorithms for jujube dried under different conditions," *Applied Food Research*, vol. 5, no. 1, p. 100699, 2025, https://doi.org/10.1016/j.afres.2025.100699.

[86] J. Szulc, G. Gozdecka, and W. Poćwiardowski, "The application of NIR spectroscopy in moisture determining of vegetable seeds," *Czech Journal of Food Sciences*, vol. 38, no. 2, pp. 131-136, 2020, https://doi.org/10.17221/57/2019-CJFS.

[87] X. He, X. Feng, D. Sun, F. Liu, Y. Bao, and Y. He, "Rapid and Nondestructive Measurement of Rice Seed Vitality of Different Years Using Near-Infrared Hyperspectral Imaging," *Molecules*, vol. 24, no. 12, p. 2227, 2019, https://doi.org/10.3390/molecules24122227.

[88] J. Yasmin *et al*., "Near-infrared hyperspectral imaging for online measurement of the viability detection of naturally aged watermelon seeds," *Frontiers in Plant Science*, vol. 13, 2022, https://doi.org/10.3389/fpls.2022.986754.

[89] P. Reddy, K. M. Guthridge, J. Panozzo, E. J. Ludlow, G. C. Spangenberg, and S. J. Rochfort, "Near-Infrared Hyperspectral Imaging Pipelines for Pasture Seed Quality Evaluation: An Overview," *Sensors*, vol. 22, no. 5, p. 1981, 2022, https://doi.org/10.3390/s22051981.

[90] Z. Chang, M. Chen, G. Cheng, C. Jin, and T. Yang, "Moisture Content Detection of Soybean Grains Based on Hyperspectral Imaging," *INMATEH - Agricultural Engineering*, vol. 74, no. 3, pp. 562-570, 2024, https://doi.org/10.35633/inmateh-74-50.

[91] Q. Hu, W. Lu, Y. Guo, W. He, H. Luo, and Y. Deng, "Vigor Detection for Naturally Aged Soybean Seeds Based on Polarized Hyperspectral Imaging Combined with Ensemble Learning Algorithm," *Agriculture*, vol. 13, no. 8, p. 1499, 2023, https://doi.org/10.3390/agriculture13081499.

[92] K. H. S. Peiris, S. R. Bean, and S. V. K. Jagadish, "Extended multiplicative signal correction to improve prediction accuracy of protein content in weathered sorghum grain samples," *Cereal Chemistry*, vol. 97, no. 5, pp. 1066-1074, 2020, https://doi.org/10.1002/cche.10329.

[93] M. Rajabi-Sarkhani *et al*., "Identifying Optimal Wavelengths from Near Infrared Spectroscopy Using Meta-heuristic Algorithms to Assess Peanut Seed Viability," *Agronomy*, vol. 13, p. 2939, 2023, https://doi.org/10.20944/preprints202311.0733.v1.

[94] H. Zhang, K. Kang, C. Wang, Q. Sun, and B. Luo, "Cross-variety seed vigor detection using new spectral analysis techniques and ensemble learning methods," *Journal of Food Composition and Analysis*, vol. 136, p. 106845, 2024, https://doi.org/10.1016/j.jfca.2024.106845.

[95] T. Zhu and J. Xing, "Near-infrared spectroscopy and ensemble learning modeling for moisture detection in forest floor leaf litter," *Vibrational Spectroscopy*, vol. 140, p. 103841, 2025, https://doi.org/10.1016/j.vibspec.2025.103841.

[96] P. Zhou, J. Chen, L. Du, and X. Li, "Balanced Spectral Feature Selection," *IEEE Transactions on Cybernetics*, vol. 53, no. 7, pp. 4232-4244, 2023, https://doi.org/10.1109/TCYB.2022.3160244.

[97] W. Cao, "A comparative study of hybrid adaptive neuro-fuzzy inference systems to predict the unconfined compressive strength of rocks," *Journal of Engineering and Applied Sciences*, vol. 72, no. 1, p. 3, 2025, https://doi.org/10.1186/s44147-024-00574-9.

[98] M. Al-Amery *et al*., "Near-infrared spectroscopy used to predict soybean seed germination and vigour," *Seed Science Research*, vol. 28, no. 3, pp. 245-252, 2018, https://doi.org/10.1017/S0960258518000119.

[99] M. F. da Silva *et al*., "Near infrared spectroscopy for the classification of vigor level of soybean seed," *Revista Ciência Agronômica*, vol. 55, 2024, https://doi.org/10.5935/1806-6690.20240005.

[100] Q. Hu, W. Lu, Y. Guo, W. He, H. Luo, and Y. Deng, "Vigor Detection for Naturally Aged Soybean Seeds Based on Polarized Hyperspectral Imaging Combined with Ensemble Learning Algorithm," *Agriculture*, vol. 13, no. 8, p. 1499, 2023, https://doi.org/10.3390/agriculture13081499.

[101] X. Han, C. Zhengguang, and L. Jinming, "Rapid detection of maize seed germination using near-infrared spectroscopy combined with Gaussian process regression," *Food Chemistry*, vol. 491, p. 145254, 2025, https://doi.org/10.1016/j.foodchem.2025.145254.