Indian Journal of Agricultural Sciences **95** (7): 833–839, July 2025/Article
https://doi.org/10.56093/ijas.v95i7.160183

Near-infrared spectroscopy integrated machine learning techniques for viable seed identification

MONIKA SINGH[1, 2], ANU SHARMA[2]*, K K CHATURVEDI[2], SANJEEV KUMAR[2], DWIJESH CHANDRA MISHRA[2], ALKA ARORA[2], RAKESH BHARDWAJ[3], MRINMOY RAY[2], MAMATHA Y S[2] and SAMARTH GODARA[2]

ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110 012, India

Received: 19 November 2024; Accepted: 4 June 2025

ABSTRACT

The evaluation of seed viability is pivotal in agriculture, biodiversity conservation, and ecological research. Traditional methods used for testing the seed viability are often destructive and pose challenges regarding labour intensity and seed wastage. The study was carried out during 2022–23 at ICAR-Indian Agricultural Statistics Research Institute, New Delhi with the aim of collecting the seed genotypes and NIR spectroscopic instrument and computational approaches and appropriate hardware and software resources. A diverse dataset of NIR spectral data from various seed species was used and analysed using three sophisticated ML models, namely Linear Discriminant Analysis (LDA), Random Forest (RF), and Artificial Neural Networks (ANN). The performance of the developed models was evaluated based on accuracy, precision, recall, and F1 score metrics. Furthermore, the experimental results demonstrated that NIR spectroscopy and ML could effectively classify viable seed. The integration of artificial neural networks (ANNs) has demonstrated significant potential in capturing intricate patterns within spectral data, achieving an approximate accuracy of 95%. This highlights their effectiveness in precise classification tasks. Additionally, machine learning (ML)-based approaches have shown promise in conserving valuable seed resources by offering scalable solutions adaptable to large-scale agricultural and conservation applications. To enhance model transparency, Local Interpretable Model-Agnostic Explanations (LIME) has been employed, providing deeper insights into the ANN's decision-making process by identifying key spectral features that influence classification outcomes. It was observed that ML-based approaches have the potential to enable continuous monitoring, contributing to the conservation of valuable seed resources. Additionally, these methods may offer a scalable solution that can be adapted for large-scale agricultural and conservation applications.

Keywords: Classification, Machine learning, Near-infrared (NIR) spectroscopy, Sustainable conservation, Viable seed

Ensuring the quality of seeds in production processes is crucial, as the viability of seeds directly influence the germination rate, affecting both research and breeding programmes. Accurate identification of viable seeds enhances seed batches' overall quality and germination rates. Traditional methods for assessing seed viability and damage include measurements of seedling growth characteristics, adversity resistance, and various physiological and biochemical tests, as well as physical and chemical methods. While these methods are precise, they often do not meet the requirements for non-destructive, rapid assessments in large-scale seed processing due to their complexity, time consumption, cost and potential seed damage. In contrast, NIRS offers a non-destructive and rapid alternative for seed evaluation (Al-Amery *et al.* 2018). NIRS technology coupled with ML methods used for capturing unique spectral signatures from seeds, reflecting the molecular vibrations related to various functional groups (Workman and Weyer 2012). Daneshvar *et al.* (2015) findings supported the efficacy of ML models in enhancing the accuracy to distinguish viable and non-viable seeds using single-seed NIR spectral data. Kosmowski and Worku (2018) have utilised predictive multiclass ML models for grain cultivar identification and classification. Al-Amery *et al.* (2018) used NIRS and Partial Least Squares regression (PLSR) to predict soybean seed germination and vigor. Priyadarshi *et al.* (2022) evaluated the advantage of ANN model on limited NIR spectral sample size various machine learning based regression techniques for predicting protein components and determination of starch in chickpea germplasm. Tigabu

[1]ICAR-Indian Agricultural Research Institute, New Delhi; [2]ICAR-Indian Agricultural Statistics Research Institute, New Delhi; [3]ICAR-National Bureau of Plant Genetic Resources, New Delhi. *Corresponding author email: anusharma.iasri@icar.org.in

et al. (2019) utilized multivariate discriminant analysis of single seed NIR spectra to distinguish between viable and dead-filled seeds of various pine species. Explainable Artificial Intelligence (XAI) has gained attention, but its use in agriculture remains limited. Wei *et al.* (2022) explored deep learning interpretability in agricultural classification using a fruit leaves dataset, analyzing whether models focus on leaf appearance or lesion texture features during feature extraction. The present study was undertaken with the aim of pre-processing and dimensionality reduction for handling multi-collinearity using the spectrum of different aged individual kernels of soybean and to develop and train ML models by capturing the variability on their spectral signatures due to seed ageing during analysis for rapid and non-destructive identification of viable seeds using performance measures of ML techniques with standard parameters and making model explainable using Explainable AI techniques irrespective to the black box computation analysis.

MATERIALS AND METHODS

The present study was carried out during 2022–23 at ICAR-Indian Agricultural Statistics Research Institute, New Delhi. Under the development of the methodology, different steps have been approached such as, pre-processing and dimensionality reduction for handling multi-collinearity form spectral data; employing ML techniques for rapid and non-destructive identification of viable seeds based on spectral signatures; to compare and evaluate various ML techniques with standard parameters of evaluation of NIR spectral data of soybean crop. The work flow of the proposed study is shown in Fig. 1.

Sample collection and spectral data acquisition: The soybean seeds were harvested during 2022, 2023; and seed genotypes stored under Medium-Term Storage (MTGS) condition were collected from the ICAR-National Bureau of Plant Genetic Resources, New Delhi. Soybean seed samples harvested over the past two years were collected to capture variations in spectral patterns associated with seed aging. By selecting seeds from different harvesting periods, the study aimed to account for the natural variability in seed composition and quality over time. This approach ensured a comprehensive analysis of spectral differences, enabling a better understanding of how aging influences NIR spectral measurements. A total of 1,304 seed genotypes samples were taken for spectral measurement and analysis. Spectral data of each genotype was collected using the FOSS-6500 model of NIR Spectroscopy, an advanced instrument designed to precisely measure the NIR light spectrum across wavelengths from 1100–2500 nm at 0.5 nm intervals. The setup utilized portable NIR sensors mounted on a device, which was connected to a computer via a Universal Serial Bus (USB) interface. The USB connection played a crucial role in ensuring seamless data transfer between the NIR sensor and the computer, allowing real-time acquisition, processing, and storage of spectral data. Additionally, USB connectivity facilitated power supply to the sensor, eliminating the need for external power sources and enhancing the portability of the setup. This system was instrumental in collecting spectral data from two categories of seeds, viable and non-viable. After the spectral measurements, a standard germination test was conducted in experimental environment to detect seed status as viable and non-viable. Standard germination testing was conducted on each individual seed sample following the protocols and guidelines established by the International Seed Testing Association (ISTA). ISTA is a globally recognized organization that develops and maintains standardized procedures for seed testing to ensure accuracy, consistency, and reliability in seed quality assessment. Adhering to ISTA protocols is essential in agriculture, as it guarantees uniformity in germination testing across different laboratories and regions, enabling fair trade and regulatory compliance. The spectral and referenced data was acquired having both the reference value against all their spectral intensities captured through spectroscopy. The practice of splitting data into training and testing sets using an 80:20 ratio is widely adopted in machine learning, including the training of Artificial Neural Networks (ANNs). This approach allocates 80% of the data for training the model and reserves the remaining 20% for evaluating its performance. Empirical studies have demonstrated that such a division often yields optimal results, balancing the need for sufficient training data to build robust models and adequate testing data to assess their generalization capabilities (Gholamy *et al.* 2018). The acquired dataset was randomly split into two parts for model training and development 80% was used for training, while the

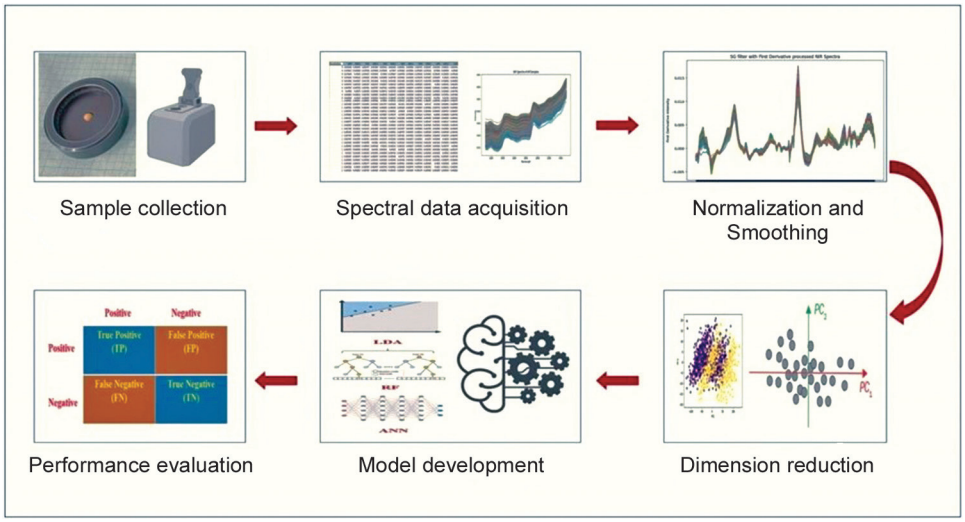

Fig. 1 Workflow of the methodology for viable seed classification.

remaining 20% testing set to evaluate the model's performance.

Normalization and smoothing: Spectral data represents the absorption of near-infrared light by various molecules, offering valuable insights into their composition and properties. Raw spectral data often contains noise and scatter effects resulting from spectral measurement processes. To ensure the integrity and enhancing quality of spectral data for subsequent analysis, various pre-processing methods were employed. Pre-processing methods are helpful in eliminating noise and correcting for scatter effects from spectral data (Rinnan *et al.* 2009). Standard Normal Variate (SNV) is used to correct scatter effects and normalize spectra, reducing variability caused by differences in particle size, path length or other scattering effects (Barnes *et al.* 1989). Savitzk Golay (SG) derivative and filtering method applies a polynomial smoothing function to the spectrum that helps to reduce noise while preserving important features like peaks and fine structures. Hence, this method applies a polynomial smoothing filter to compute the derivative efficiently (Savitzky and Golay 1964). Spectral data of Soyabean seed was processed using a combination of scatter correction and smoothing using SNV and SG derivative smoothing. The pre-processed data was further scaled to standardize for its suitability for comparative analysis. This process adjusted the spectra with a mean of zero and a standard deviation of one.

Dimension reduction: Variable selection in NIR spectroscopy poses challenges due to the multitude of variables involved and their high correlation (Roger *et al.* 2011). To handle multi-collinearity in spectroscopic data, Principal Component Analysis (PCA) is particularly useful for handling and simplifying high-dimensional data into uncorrelated variables, known as principal components (PCs) (Wold *et al.* 1987). By emphasizing the principal components that encompass the greatest variance, PCA removes less significant features that add minimal value to the overall dataset. This process of dimensionality reduction enhances storage efficiency and speeds up computation (Zhao *et al.* 2021, Machine Learning Models). After pre-processing the raw spectra data the PCA has been applied simplifying for handling of extensive datasets.

ML model development: NIR-based machine learning has proven to be a promising tool for predicting seed viability. Olesen *et al.* (2011) successfully applied this approach to assess the viability of spinach seeds, while Shrestha *et al.* (2017) conducted single-seed analysis to evaluate the viability of tomato seeds. Ozturk *et al.* (2023) highlighted the use of NIRS and ML to classify different types of powdered food materials with high accuracy. Esteve Agelet *et al.*(2012) developed ANN model for classifying conventional and genetically modified soybean seeds using spectrographic imaging data. The models were trained at the IASRI Computer Laboratory on an HP desktop computer featuring an Intel® Core i5 processor. Utilizing Python 3.8.8, The implementation utilized the Keras Application Programming Interface (API) version 2.4.3 using Python, 3.8.8, which provides a high-level framework for building and training deep learning models. TensorFlow 2.3 was used as the backend, enabling efficient computation and model execution. To enhance processing speed, Graphics Processing Unit (GPU) acceleration was employed, significantly improving the performance of complex neural network operations. This was achieved through the installation of Compute Unified Device Architecture (CUDA), a parallel computing platform and programming model developed by NVIDIA, which allows GPUs to perform general-purpose computing tasks, optimizing deep learning workflows. The Linear Discriminant Analysis LDA model was chosen for its effectiveness in handling small-sized datasets with a well-defined number of classes. LDA helps in finding a linear combination of features that characterises or separates two or more classes of objects or events. LDA works by increasing the difference between different classes while reducing the spread within each class. This helps in better distinguishing one class from another in the dataset (Grassi *et al.* 2018). RF is an ensemble learning method known for its high accuracy and robustness, which comes from its ability to operate on large datasets with high-dimensional features and manage over-fitting. It operates by constructing a multitude of decision trees at training time and outputting the class, which is the mode of the classes of the individual trees (Breiman 2001). Spectral data typically involves complex information spread across different wavelengths or frequencies. ANNs are flexible computing frameworks and can model a wide variety of complex, non-linear relationships within the spectral data, capturing intricate patterns and details that might be present in spectra. By integrating the Adam optimizer, binary cross-entropy loss, and Rectified Linear Unit (ReLU) activation functions, the ANN model can effectively learn and classify NIR spectra data. The ReLU enables the network to model complex, non-linear patterns within the spectral data, enhancing the overall predictive performance (Hornik *et al.* 1989). Adam's adaptive learning rates and fast convergence facilitates efficient training, while binary cross-entropy loss ensures accurate probabilistic outputs for binary classification tasks. The neural network model discerned and processed these fine-grained patterns and variations effectively. This capability is crucial for making precise and reliable predictions or classifications based on the spectral inputs. All models in this study underwent hyper-parameter tuning via random search. Random search for hyper-parameter tuning involves randomly sampling combinations of hyper-parameters within specified ranges (Smith 2018). On taking consideration of limiting data size, three discrimination analysis based models using machine learning algorithms were developed involving systematic normalization and smoothing of the NIR spectra data. The dimensionality reduction via PCA and subsequent analysis for classification using three distinct machine learning algorithms those are LDA, RF and ANN have been done. ANN model was developed by using the neural networks that have been initiated with an input layer designed to

handle input vectors of PCs as input dimensions. This was succeeded by a dense layer containing neurons appropriate to optimal training and learning of the model, which employed the ReLU activation function to introduce non-linearity and manage complex data relationships in all the dense layers except the input and output layer. Another hidden dense layers followed consisting neurons utilized the ReLU activation function to further refine features extracted by preceding layers and representing data in more compact form. The network concluded with a dense output layer featuring a single neuron and a sigmoid activation function which produces a probabilistic output appropriate for binary classification tasks. This architectural design enabled the model to effectively capture and process the intrinsic patterns in the input data while maintaining an efficient and streamlined structure conducive to both computation and training. Early stopping with a patience of 40 helped to prevent overfitting and improved computational efficiency by monitoring validation accuracy and halting training if it does not improve for a specified number of epochs.

Model explanation interpretability: To enhance the transparency and interpretability of the model, feature importance analysis and visualization techniques were employed. Local Interpretable Model-Agnostic Explanations (LIME) was utilized to assess the contribution of individual features to the model's predictions. This technique provided insights into the decision-making process of the model, helping to identify key spectral features influencing classification outcomes (Saranya and Subhashini 2023).

Performance evaluation: In the study, a confusion matrix was employed to derive the following key metrics to evaluate the models' performance. Accuracy metrics evaluate the performance of a model by measuring the proportion of correct predictions over the total predictions made. The formula for accuracy is given by eq. (i):

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of prediction}} \quad \text{eq. (i)}$$

Precision metric measures the proportion of true positive (TP) predictions out of all positive that are the total true positive and false positive (FP) predictions made by the model. The formula for precision is given in eq. (ii).

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad \text{eq. (ii)}$$

The recall metric assesses a model's ability to identify all relevant instances within a dataset correctly. The formula for recall is given in eq. (iii).

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad \text{eq. (iii)}$$

The F1 score depicted in eq. (iv) is a harmonic mean of precision and recall, providing a balanced measure of a model's performance.

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{eq. (iv)}$$

Where the terms used in calculating performance measures were inferred as True Positives (TP) indicated seeds of viable class classified correctly. True Negatives (TN) represented numbers of seeds belong to non-viable class classified correctly. False Positives (FP) calculated the number of non-viable seeds classified incorrectly as viable seeds. False Negatives (FN) denoted the number of viable seeds incorrectly classified as non-viable seeds.

RESULTS AND DISCUSSIONS

The graph (Supplementary Fig. 1) demonstrated the plot of spectra pre-processed using the methods namely SNV normalization and SG smoothing. The pre-processing on spectra has standardized the spectra and enhanced the signal-to-noise ratio. This pre-processing step has made the spectra more suitable for quantitative and qualitative analysis by ensuring that the remaining variability reflects genuine chemical differences rather than noise or scatter effects. The processed data now exhibiting consistent and distinguishable spectral features is well-prepared for subsequent analysis such as Principal Component Analysis (PCA) and classification tasks aiding in the effective differentiation between viable and non-viable samples. Despite the pre-processing the plot showed notable peaks and troughs at wavelengths around 1400 nm, 1900 nm and 2400 nm which represented key absorptions or features in the NIR spectra. The factor loadings derived from PCA analysis presented in Supplementary Table 1 revealed the contribution of each original variable to the principal components. High factor loadings for certain variables indicate their strong influence in capturing the most of the variance in data. The variables with high loadings on the first principal component (PC1) were most significant in explaining the primary variation within the dataset. Each PC represents a linear combination of the original spectral variables, and its factor loadings indicate how much each wavelength contributes to that particular principal component. Higher absolute values in a specific PC column suggest that the corresponding wavelength strongly influences that component. In this study, PCA effectively reduced data dimensionality that captured the maximum variance (97%) under the five principal components. The scree plot (Supplementary Fig. 2) was used for choosing the optimal number of principal components in PCA analysis. The plot illustrated the cumulative explained variance against the number of principal components. The x-axis represented the number of PCs ranging from 0–9 while the cumulative explained variance on the y-axis indicates the proportion of total variance retained by the selected principal components (PCs). It starts at approximately 0.970, signifying that the initial PCs already capture 97% of the variance in the dataset. As additional PCs are included, the cumulative variance gradually approaches 1.000, indicating that nearly all variance is accounted for making further components less significant in contributing new information. By identifying the point where the explained variance begins to plateau indicated that the first few principal components captured

Table 1 Performance Evaluation of models on testing and training dataset

	Training Dataset				Validation Dataset			
Models	Acc.	Pr.	Re.	F1	Acc.	Pr.	Re.	F1
LDA	70.85	70.00	71.00	70.00	73.94	73.00	74.00	73.00
RF	98.75	99.00	99.00	99.00	88.50	88.00	89.00	88.00
ANN	96.64	95.61	95.37	95.49	95.78	94.68	93.68	94.17

Acc., Accuracy (%); Pr., Precision (%); Re., Recall (%); F1, F1 score (%).

the majority of the data variance making them the most significant for dimensionality reduction from the feature dimension of 2800 to the latent lower dimensional space under five PCs only. The most important wavelength band has been highlighted in the range between1150–1500 nm. This band linked to vibrations from O–H, C–O, N–H, and C–H bonds found in proteins, lipids, and carbohydrates enabled the creation of stable and understandable models in NIR spectroscopy (Ambrose *et al.* 2016, Baek *at al.* 2019). The five principal components (PCs) were used as the feature set for further model development. The performance differences among the models can be attributed to their respective architectures and ability to handle the dataset's complexity. The neural network architecture consists of multiple layers, beginning with an input layer designed to receive data with five features, corresponding to the selected five PCs. The first dense layer, followed with an output shape of transformed the data into a 50-dimensional space. Next, a flatten layer is employed to reshape the data preserving the same dimension count. The data is then processed through a second dense layer which reduced the dimensions to five again. This is followed by a third dense layer further reduced the output having the shape 3-dimensional space, and finally, a fourth dense layer produced output the data with a single unit, resulting in an output shape of 1-dimensional space. ANN utilised the defined various hidden layers to fit model inputs to nonlinear outputs, demonstrating strong performance in capturing underlying patterns in the data. The neural network architecture employed in this study was trained for 400 epochs using binary cross-entropy loss and the Adam optimizer. Initially, the model started with a loss of 0.6028 and an accuracy of 0.6702. As training progressed, the loss steadily decreased to 0.0948, while accuracy improved to 0.9578. Fig. 2 represented the sample of model progress trend about the learning of model and accuracy of classification till epochs 60. This trend reflects a typical machine learning training pattern, where the model gradually learns and refines its ability to classify data more accurately over successive epochs. By using early stopping with a patience of 40, the training process was halted if the validation loss did not improve after 40 consecutive epochs. The performance of the machine learning models undertaken in the present study was assessed using accuracy, precision, recall and F1-score. The performance of the models was analysed on both the training and testing datasets as presented in Table 1. The table showed that the model LDA was simpler and more interpretable, adequately achieved accuracy around 70.85% in training and improved to 73.94% attained accuracy on testing dataset. RF demonstrated superior pattern recognition in the training phase with accuracy of 98.75% and 88.00% F1-scores. However, RF decreased to 88.50% in testing accuracy, indicated possible over-fitting. ANN showed consistent robustness and maintained high performance with about 95% accuracy across both phases of training and testing. High precision (94.68%) and recall (93.68%) indicated accurate positive predictions done by ANN on testing dataset. The F1-score on testing dataset stands at 94.17% reflected overall robust performance of the neural network machine learning model.

Typical machine learning training patterns by plotting the accuracy graph against the number of epochs displayed in Fig. 2. Despite fluctuations, overall improvement was observed indicating effective learning and convergence. The graph illustrated the accuracy of a machine learning model over a series of 70 epochs for both the training and validation datasets. The x-axis represented the epochs ranging from 0–70 while the y-axis showed accuracy ranging from 0.6–1.0. The training accuracy represented by the blue dashed line started at around 0.6 and gradually increased showing a steady improvement as the number of epochs progresses ultimately approaching 0.95. Similarly, the testing accuracy depicted by the orange solid line follows a comparable trend with some fluctuations and also reaches approximately 0.95 by the end of the epochs. Key observations from the graph

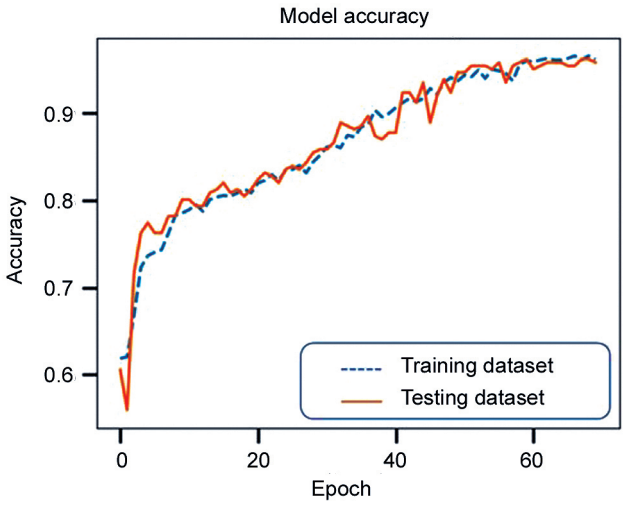

Fig. 2 ANN-model accuracy graph plot.

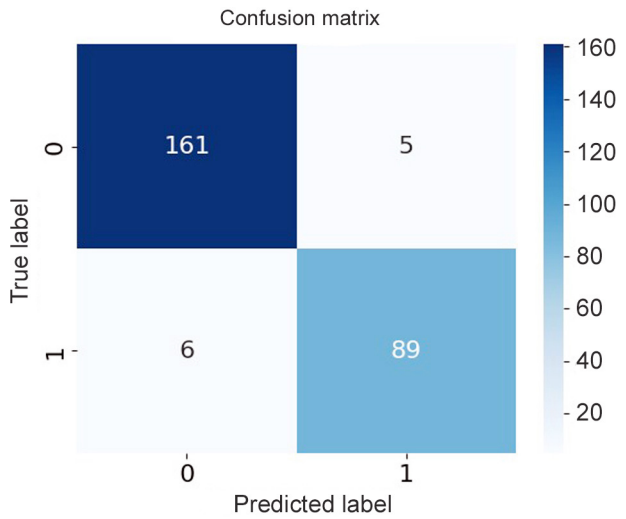

Fig. 3 Confusion matrix of PC-ANN model.

include the consistent improvement of both training and validation accuracies indicated effective learning by the model. The close alignment between the two accuracy lines suggested that the model generalizes well to unseen data without over-fitting. The Receiver Operating Characteristic (ROC) curve attained an Area Under Curve (AUC) of 0.99%, demonstrating highly accurate identification of viable seeds and a low False Positive Rate, which enhances the efficiency of classification.

The confusion matrix results demonstrated the strong performance of PC-ANN model in distinguishing samples lying between the two classes (Fig. 3). It correctly identified 161 negative instances and 89 positive instances, with only 5 false positives and 6 false negatives. These results reflected high accuracy, precision, recall, and specificity. The model's ability to minimize both false positives and false negatives highlights its reliability and robustness, making it well-suited for the classification task at hand. Ozturk *et al.* (2023) highlighted the use of NIRS and ML to classify different types of powdered food materials with high accuracy. The results underscored the efficacy of the proposed approach with the PCA analysis based ANN model emerging as the most promising model achieved an impressive accuracy of 95.78% on the testing dataset signified the capability of ANN architecture combined with the supervised method of dimension reduction and feature selection to extract intricate spectral patterns which is essential for assess the accurate class belonging the viable status. The model has the ability of robust classification of seeds as compare to the traditional discriminating analysis models. The Supplementary Fig. 3 presented the [Local Interpretable Model-Agnostic Explanations (LIME)] results for four different instances (36, 106, 142, and 206), illustrating the contribution of principal components (PCs) to the classification decision. Each subplot highlights the impact of different PCs, with positive contributions increasing the likelihood of classification into a certain category and negative contributions decreasing it. Across the instances, PC1 and PC2 emerge as the most influential features, although their impact varies. For instance, in Instance 36, PC1 (≤ -41.21) had the strongest negative contribution, whereas PC2 (>4.90) played a significant positive role. Similarly, in Instance 142, PC1 (>36.54) has the highest positive contribution, while PC2 (≤ -4.41) exerts a strong negative influence. In contrast, Instance 206 exhibits much smaller feature contributions, indicated that classification in this case might be less sensitive to individual PCs. These variations in feature importance suggested that different spectral characteristics influenced classification decisions uniquely across instances. The LIME explanations provided valuable interpretability, allowing to understand which spectral components play a crucial role in model predictions. This understanding helps refine feature selection, ensuring that the model's decisions are based on meaningful spectral variations rather than arbitrary factors.

This study highlighted the importance of pre-processing in standardizing spectral data, enhancing the signal-to-noise ratio and improving its suitability for both quantitative and qualitative analysis. As a result, the processed data was optimally prepared for feature extraction and modeling. The selection of principal components was guided by the distribution of variance, ensuring that the most informative features were retained for analysis. The plot facilitated the selection of the five most impactful components for analysis. This contributed to improved data processing and interpretation, essential for robust and reliable analytical outcomes. By harnessing the power of NIR spectroscopy coupled with ML, the neural network based methodology offers a scalable and precise solution for viable seed classification which is crucial for agricultural and ecological applications. The effectiveness of the Artificial Neural Network (ANN) model in handling the complex and non-linear nature of the dataset requires a comprehensive approach. This includes integrating pre-processing steps, feature extraction and exploring a range of advanced Machine Learning (ML) and Deep Learning (DL) models incorporating broader spectral data and validating the methodology through field trials. This multi-faceted strategy ensures that the analytical needs are met and the model performs robustly in real-world applications.

Overall, our findings highlighted the potential of NIR spectroscopy integrated with ML framework to assess the seed status based on its viability offer rapid, non-destructive and accurate classification and monitoring system. This methodology is promising for optimising seed processing workflows, conserving valuable seed resources and supporting sustainable agricultural practices and conservation efforts. Akkem *et al.* (2024) provided a comprehensive review of the techniques in smart farming, emphasizing their potential to enhance model performance and reliability. Leveraging such advanced data augmentation approaches can significantly improve the robustness and accuracy of predictive models in agriculture. While this study demonstrated the effectiveness of NIR spectroscopy integrated with ML models, further advancements

can enhance its practical applications. Incorporating synthetic data generation techniques, such as Variational AutoEncoders (VAEs) and Generative Adversarial Networks (GANs) can help to address data limitations and improve model generalization. Additionally, the implementation of a recommendation system for seed health status may be deployed using the Streamlit framework as an intuitive interface for end-users to interact with ML predictions based on the seed data. Expanding such interactive applications can enhance decision-making for farmers and agricultural stakeholders. Future research can focus on integrating real-time data streams, optimizing user experience and expanding the framework to support a wider range of crops.

REFERENCES

Agelet L E, Gowen A A, Hurburgh Jr C R and ODonell C P. 2012. Feasibility of conventional and Roundup Ready soybeans discrimination by different near infrared reflectance technologies. *Food Chemistry* **134**: 1165–72.

Akkem Y, Biswas S K and Varanasi A. 2024. A comprehensive review of synthetic data generation in smart farming by using variational autoencoder and generative adversarial network. *Engineering Applications of Artificial Intelligence* **131**: 10788.

Al-Amery M, Geneve R L, Sanches M F, Armstrong, P R, Maghirang E B, Lee C, Vieira R D and Hildebrand D F. 2018. Near-infrared spectroscopy used to predict soybean seed germination and vigour. *Seed Science Research* **28**(3): 245–52.

Ambrose A, Lohumi S, Lee W H and Cho B K. 2016. Comparative non-destructive measurement of corn seed viability using Fourier transform near-infrared (FT-NIR) and Raman spectroscopy. *Sensors and Actuators B: Chemical* **224**: 500–06,

Baek I, Kusumaningrum D, Kandpal LM, Lohumi S, Mo C, Kim MS and Cho B K. 2019. Rapid measurement of soybean seed viability using kernel-based multispectral image analysis. *Sensors* **19**(2): 271. https://doi.org/10.3390/s19020271

Barnes R J, Dhanoa M S and Lister S J. 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy* **43**(5): 772–77.

Breiman L. 2001. Random Forests. *Machine Learning* **45**(1): 5–32

Daneshvar A, Tigabu M, Karimidoost A and Oden P. 2015. Single seed near infrared spectroscopy discriminates viable and non-viable seeds of *Juniperus polycarpos*. *Silva Fennica* **49**(5). https://doi.org/10.14214/sf.1334

Gholamy Afshin, Kreinovich Vladik and Kosheleva Olga. 2018. Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation.

Grassi S, Casiraghi E and Alamprese C. 2018. Handheld NIR device: A non-target approach to assess authenticity of fish fillets and patties. *Food Chemistry* **243**: 382–88.

Hornik K, Stinchcombe M and White H. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks* **2**: 359–66.

Kosmowski F and Worku T. 2018. Evaluation of a miniaturized NIR spectrometer for cultivar identification: The case of barley, chickpea and sorghum in Ethiopia. *PLOS One* **13**(3): e0193620.

Machine Learning Models (n.d.). PCA: An Unsupervised Dimensionality Reduction Technique. https://machinelearningmodels.org/pca-an-unsupervised-dimensionality-reduction-technique/

Nicolai B M, Beullens K and Bobelyn E. 2007. Non-destructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biology Technology* **46**(2): 99–118.

Olesen M H, Shetty N, Gislum R and Boelt B. 2011. Classification of viable and non-viable spinach (*Spinacia oleracea* L.) seeds by single seed near infrared spectroscopy and extended canonical variates analysis. *Journal of Near Infrared Spectroscopy* **19**(4): 285–86.

Ozturk S, Bowler A, Rady A and Watson N J. 2023. Near-infrared spectroscopy and machine learning for classification of food powders during a continuous process. *Journal of Food Engineering* **341**: 111339. https://doi.org/10.1016/j.jfoodeng.2022.111339

Priyadarshi M B, Sharma A, Chaturvedi K, Bhardwaj R, Lal S, Kumar S, Mishra D and Singh M. 2022. Machine learning algorithms for protein physicochemical component prediction using near infrared spectroscopy in chickpea germplasm. *Indian Journal of Plant Genetic Resources* **35**(1): 44–48.

Priyadarshi M B, Sharma A, Chaturvedi K K, Bhardwaj R and Singh M. 2022. Development and comparison of regression models for determination of starch in chickpea using NIR spectroscopy. *International Journal of Agriculture Environment and Biotechnology* 15(3): 683–91.

Rinnan A, Berg F V D and Engelsen S B. 2009. Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry* **28**(10): 1201–22.

Roger JM, Palagos B, Bertrand D and Fernandez-Ahumada E. 2011. CovSel: Variable selection for highly multivariate and multi-response calibration. *Chemometrics and Intelligent Laboratory Systems* **106**(2): 216–23.

Saranya A and Subhashini R. 2023. A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision Analytics Journal* **7**: 100230.

Savitzky A and Golay M J E. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* **36**(8): 1627–39. doi:10.1021/ac60214a047

Shrestha S, Deleuran L C and Gislum R. 2017. Separation of viable and non-viable tomato (*Solanum lycopersicum* L.) seeds using single seed near-infrared spectroscopy. *Computers and Electronics in Agriculture* **142**: 348–55.

Smith L N. 2018. A Disciplined Approach to Neural Network Hyper-Parameters: Part 1-Learning Rate, Batch Size, Momentum, and Weight Decay. *ArXiv*abs/1803: 09820.

Tigabu M, Daneshvar A, Jingjing R, Wu P, Ma X and Oden P C. 2019. Multivariate discriminant analysis of single seed near infrared spectra for sorting dead-filled and viable seeds of three pine species: Does one model fit all species? *Forests* **10**(6): 469.

Wei K, Chen B, Zhang J, Fan S, Wu K, Liu G and Chen D. 2022. Explainable deep learning study for leaf disease classification. *Agronomy* **12**: 1035.

Wold S, Esbensen K and Geladi P. 1987. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **2**(1–3): 37–52.

Workman J and Weyer L. 2012. *Practical Guide and Spectral Atlas for Interpretive Near-Infrared Spectroscopy*, 2nd edn. CRC Press, Taylor & Francis Group.

Zhao B, Dong X, Guo Y, Jia X and Huang Y. 2021. PCA dimensionality reduction method for image classification. *Neural Processing Letters* **54**: 347–68.