Contents lists available at ScienceDirect

Current Research in Food Science

journal homepage: www.sciencedirect.com/journal/current-research-in-food-science

Research paper

Unlocking chickpea flour potential: AI-powered prediction for quality assessment and compositional characterisation

Ali Zia [a,b,*], Muhammad Husnain [a,c], Sally Buck [a], Jonathan Richetti [a], Elizabeth Hulm [a], Jean-Philippe Ral [a], Vivien Rolland [a], Xavier Sirault [a]

[a] *Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia*
[b] *College of Science and School of Computing, Australian National University, Australia*
[c] *School of Information & Communication Technology, Griffith University, Australia*

ARTICLE INFO

Keywords:
Chickpea flour dataset
Deep learning
PLSR
NIR
Spectroscopy
Flour quality
AI

ABSTRACT

The growing demand for sustainable, nutritious, and environmentally friendly food sources has placed chickpea flour as a vital component in the global shift to plant-based diets. However, the inherent variability in the composition of chickpea flour, influenced by genetic diversity, environmental conditions, and processing techniques, poses significant challenges to standardisation and quality control. This study explores the integration of deep learning models with near-infrared (NIR) spectroscopy to improve the accuracy and efficiency of chickpea flour quality assessment. Using a dataset comprising 136 chickpea varieties, the research compares the performance of several state-of-the-art deep learning models, including Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and Graph Convolutional Networks (GCNs), and compares the most effective model, CNN, against the traditional Partial Least Squares Regression (PLSR) method. The results demonstrate that CNN-based models outperform PLSR, providing more accurate predictions for key quality attributes such as protein content, starch, soluble sugars, insoluble fibres, total lipids, and moisture levels. The study highlights the potential of AI-enhanced NIR spectroscopy to revolutionise quality assessment in the food industry by offering a non-destructive, rapid, and reliable method for analysing chickpea flour. Despite the challenges posed by the limited dataset, deep learning models exhibit capabilities that suggest that further advancements would allow their industrial applicability. This research paves the way for broader applications of AI-driven quality control in food production, contributing to the development of more consistent and high-quality plant-based food products.

1. Introduction

In the current era of global transformation of the food system, the search for sustainable, nutritious, and environmentally friendly food sources has intensified. Driven by growing concerns about health, environmental sustainability and food security (Willett et al., 2019), leguminous crops, such as chickpeas, are an attractive option in food and agricultural systems. Chickpeas, with their high levels of protein, fibre, and essential micronutrients (Wood and Grusak, 2007; Jukanti et al., 2012; Madurapperumage et al., 2021), are associated with a relatively low environmental footprint (Bar-El Dadon et al., 2017) and improved soil quality (Kirkegaard et al., 2008). The versatility of chickpeas, especially when processed into flour, offers great potential for incorporation into food products, catering to the increasing demand for plant-based alternatives (Bravo-Núñez and Gómez, 2021; Chandler and McSweeney, 2022; Mokni Ghribi et al., 2018). However, the inherent variability resulting from genetic diversity, environmental influences, and processing techniques presents a significant challenge in the production of standardised chickpea flour (Wang et al., 2017; Hall et al., 2017; De Santis et al., 2021). Understanding this variability requires an accurate assessment of the composition, including protein content, moisture levels, and fibre content. Traditional quantification approaches often involve destructive and time-consuming procedures (American Association of Cereal Chemists. Approved Methods Committee, 2000). High-throughput, non-destructive approaches are sought as an alternative to support the use of chickpea flour in new applications.

Near Infra-red (NIR) spectroscopy has emerged as a preferred solution, offering a rapid, non-destructive, and simultaneous assessment of multiple quality attributes (Porep et al., 2015; Huang et al., 2008; Ingle et al., 2016). The ease and speed of NIR applications align with

* Corresponding author at: Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia.
E-mail address: ali.zia@csiro.au (A. Zia).

https://doi.org/10.1016/j.crfs.2025.101030
Received 6 November 2024; Received in revised form 28 February 2025; Accepted 11 March 2025
Available online 21 March 2025

industry requirements for efficient quality control. NIR has been successfully deployed for compositional analysis in a variety of industrial applications, including its relatively long-term use to reliably measure protein content in wheat (Delwiche et al., 1998). However, while NIR spectra interpretation relies on precise calibration and sophisticated mathematical models, such as Partial Least Squares Regression (PLSR), these traditional approaches often struggle with complex, non-linear relationships in the data and are limited in their ability to capture subtle variations in grain traits or adapt to novel metrics. This is where AI-driven techniques offer a significant advantage. Unlike PLSR, AI models such as Convolutional Neural Networks (CNNs) excel at identifying intricate patterns and relationships within large, high-dimensional datasets, making them more effective in handling complex traits or new grain types. AI-driven models can overcome the limitations of PLSR by offering superior predictive accuracy and adaptability, particularly in contexts where traditional methods fall short (Nadimi and Paliwal, 2024).

The advent of Artificial Intelligence (AI), particularly machine learning and deep learning models, heralds a new era in the evolution of NIR spectroscopy. These computational models excel in managing high-dimensional datasets and identifying complex patterns, with the potential to significantly enhance the predictive accuracy of NIR-based quality assessments (Zhang et al., 2022a). Deep learning, with its capacity for feature extraction and pattern recognition in large datasets, offers a promising avenue for refining the interpretation of NIR spectra. This integration is poised to revolutionise quality assessment practices, making it a vital area of research and development within the food industry (Zhang et al., 2022a).

This paper hypothesises that the integration of deep learning models with NIR spectroscopy can significantly improve the accuracy and efficiency of chickpea flour quality assessment, outperforming traditional methods such as Partial Least Squares Regression (PLSR). By offering a rapid, scalable, and nondestructive approach to predict key chickpea flour quality metrics, this study aims to expand AI-enhanced NIR spectroscopy in agricultural and food sectors, promoting more sustainable and nutritionally consistent food systems globally.

The key contributions of this research are:

- Introduction of a unique dataset comprising NIR spectra for flours of 136 chickpea varieties along with the grain-composition profiles (ground truth) obtained through wet lab analysis.
- Examination of various preprocessing pipelines to enhance chickpea NIR data quality.
- Extensive evaluation and comparison of deep learning models, specifically CNN, ViT, and GCN, against the widely used PLSR approach. The findings show the potential and superiority of these models in predicting the composition of chickpea flour.
- Comprehensive discussion on potentials, current barriers, and future directions in wide-spread adoption of AI-driven NIR spectroscopy to serve as a rapid, non-destructive quality control method, with broader applications in assessing other pulse crops.

The rest of the paper is organised as follows: Section 2 describes the dataset, preprocessing techniques, the employed deep learning architectures, and the overall experimental setup. Section 3 presents the results of the experiments, comparing the performance of the deep learning models with PLSR. Section 4 discusses the implications of these findings for the food industry and outlines potential areas for further research. Finally, Section 5 summarises this work's key outcomes and potentials in enhancing quality assessment in food production.

2. Materials and methods

The traditional method used as a baseline for performance was the partial-least square regression (PLSR) (Höskuldsson, 1988). Several state-of-the-art deep learning algorithms were used to explore their effectiveness in chickpea flour profiling, and preprocessing steps preceded all these algorithms. The evaluation was performed using a 10-fold cross-validation, and two metrics were used: the coefficient of determination (R2) and the root square mean error (RMSE). For further deep learning in agriculture guidelines and recommendations, see Richetti et al. (2023).

2.1. Chickpea flour dataset

This research used 100 and thirty-six unique varieties of chickpeas of diverse geographical origin. Most lines were sourced from the ICRISAT genebank, with additional elite Australian germplasm sourced commercially. The grain was harvested from plants grown under common conditions in Perth, Western Australia. The plants were grown in glasshouse conditions temperatures between 18 °C and 32 °C, under natural illumination in 15 cm deep pots containing a 1:1 mix of potting mix and sand. The seeds were treated with inoculum, and slow-release fertiliser (ozmocote) was applied at sowing. The grains were decorticated by hand after being soaked in 4 °C to soften the seed coat. The decorated grains were dried at 40 °C and milled in a TissueLyzer II (Qiagen) with a 2 cm ball bearing at 25 Hz for 2 min. The flour was stored in airtight containers at room temperature until use.

Briefly, the wet laboratory analysis methods used for ground truth included the extraction of sugars from flour through ethanol extraction and quantification by anthrone assay followed by starch digestion and measurement by megazyme total starch assay kit. The remaining pellet was then digested by proteinase K and dried to give total insoluble fibre by mass (Pritchard et al., 2011). Protein was measured using the Bradford assay, and lipids were measured gravimetrically after chloroform-methanol extraction. The full description of the methods alongside the data is available at Buck and Ral (2024). The spectroscopy generating the reflectance between 680 and 2600 nm every 1 nm was conducted via Unity Scientific® SpectraStarTM 2600XT-R, where each sample consisted of 1g of milled flour. Thus, the dataset consisted of 136 sample points (observations), 1921 spectral features, and 06 regression targets.

The output values varied from 11.6 to 26.9% of protein content with an average and standard deviation of 16.9 ± 2.2%, 22.3 to 40.5% of starch content with an average and standard deviation of 34.4 ± 3.1%, 2.5 to 7.2% of soluble sugars with an average and standard deviation of 4.3 ± 1.1%, 11.3 to 25.3% of insoluble fibres with an average and standard deviation of 17.4 ± 1.9%, 6.3 to 10.6% of total lipids with an average and standard deviation of 8.7 ± 1.2%, and 6.8 to 9.2% of moisture by mass with an average and standard deviation of 8.3 ± 0.4%. All were normally distributed ($p_{\text{value}} < 0.05$), except insoluble fibre and protein.

2.2. Pre-processing

Two main strategies come into play in a limited dataset scenario: better preprocessing or generating synthetic data. Preprocessing is crucial because it refines the existing data, making it more suitable for model training by removing noise and irrelevant variations. On the other hand, generating synthetic data, also known as data augmentation, is often a valuable technique in machine learning that helps create additional data points and improve model generalisation. However, in our case, synthetic data generation was not a viable approach due to the structure of our dataset. Each observation was associated with a unique regression value, resulting in no redundancy in the ground truth values. This lack of variability was a barrier to accurately estimating the underlying statistical distribution, making data augmentation likely to introduce bias and increase the risk of overfitting. Consequently, we focused on developing robust preprocessing pipelines to optimise extracting informative features from the limited dataset.

Preprocessing NIR data helps eliminate physical phenomena, such as scattering and baseline shifts, which can often obscure the true

Table 1
Shortlisted preprocessing pipelines.

Pipeline ID	Pipeline sequence	Rationale
PP-01	Standard Scaler + Global MinMax Scaler + SG Smoothing	Applies dual scaling strategies (mean-centring and range scaling) combined with SG smoothing to normalise data and reduce spectral noise, which is suitable for datasets with widely varying ranges, thus preventing model bias towards higher magnitude features.
PP-02	Log Transform + SNV + Standard Scaler + SG Smoothing + Normalisation	This intensive preprocessing sequence is designed to thoroughly prepare spectral data by reducing skewness, normalising variance, and smoothing, thereby providing a robust basis for complex multivariate analyses, which is especially useful in datasets with severe anomalies.
PP-03	Standard Scaler + Normalisation	A basic preprocessing to Standardise and normalise data, ensuring all features have a mean of 0, a standard deviation of 1, and are rescaled to a $[0,1]$ range. The purpose is to eliminate bias from varying feature scales, facilitate efficient learning, and improve model generalisation by providing uniformly scaled input data. This preparation is crucial for optimising the performance of gradient-based methods (DL models) sensitive to feature distribution and scale.
PP-04	Standard Scaling + Translation $(value - \|min\| + 1)$ + Square	Aims to standardise the data, translate it to ensure all values are positive, and then apply a squaring transformation. The translation step, $(value - \|min\| + 1)$, avoids squaring values less than 1, which would otherwise become even smaller and risk smoothing out intricate details in the data. Squaring after translation amplifies differences between larger values, enhancing the model's ability to capture nonlinear relationships while preserving important details that might have been lost if values remained small.

chemical information within the spectra. These physical effects often dominate the variation in the NIR data, making preprocessing essential to improve the accuracy of subsequent multivariate models used in deep learning applications. Typically, a combination of scatter correction methods such as Multiplicative Scatter Correction (MSC) and Standard Normal Variate (SNV) and spectral derivatives such as Savitzky-Golay (SG) smoothing are used along with some scaling, translation, and normalisation schemes for the preprocessing of NIR data (Rinnan and Engelsen, 2009). These methods are designed to reduce unwanted variations due to scattering and to enhance the signal-to-noise ratio, thereby improving model performance.

To determine the most effective preprocessing for our specific dataset, we tested various preprocessing pipelines, each combining various preprocessing techniques to focus on different aspects of the data. These included combinations of log transformation, SNV, standard scaling, SG smoothing, normalisation, and translation techniques. The rationale behind testing these specific pipelines was to explore a range of preprocessing approaches that could potentially enhance the signal quality and robustness of the data before feeding it into our DL models.

By systematically evaluating these pipelines, we aimed to identify the most suitable preprocessing strategy that could maximise model performance despite the challenges posed by our limited dataset. Keeping distinct rationales in view, we shortlisted four pipelines, as listed in Table 1. The effects of different preprocessing pipelines on our dataset can be visualised in Appendix A.

2.3. Partial least square regression

Partial Least Squares Regression (PLSR) is an advanced statistical technique used to model complex relationships between multiple independent variables (predictors) and one (response). PLSR can be extended to handle multiple response variables; it does so separately rather than performing true multi-task learning, where a model simultaneously learns multiple related tasks. Unlike traditional linear regression, which focuses on maximising the variance explained in the dependent variable, PLSR simultaneously projects both predictors and responses onto a new set of latent variables. This dual projection reduces dimensionality while preserving as much information as possible, making PLSR particularly effective when dealing with highly collinear, noisy, or sparse datasets. Due to its robustness and versatility, PLSR has become an essential tool in food engineering for tasks such as quality control, product classification, and process optimisation (Aghdamifar et al., 2023; Schuster et al., 2023; Dal-Pastro et al., 2016; Cheng and Sun, 2005).

The implementation of PLSR in our work was facilitated by the Scikit-learn library (Sklearn), which provides comprehensive tools for model development and analysis. A key aspect of the optimisation of the PLSR model involves selecting the appropriate number of components (NC), which are the latent variables that represent the underlying structure in the data. The determination of the optimal NC is critical, as it directly influences the model's predictive accuracy. This selection process typically involves cross-validation, where the model's performance is assessed on a validation dataset to ensure it generalises well to new data, thereby achieving a balance between model complexity and predictive power. Table 3 enlists our strategy to optimise these PLSR-related hyperparameters using 10-fold cross-validation.

2.4. Deep learning

The rapid advancement of computing hardware, coupled with the evolution of sophisticated frameworks for artificial intelligence (AI) programming, has significantly accelerated the development and application of deep learning across a broad spectrum of research domains (Dean, 2022; Talaei Khoei et al., 2023). Deep learning has gained considerable traction in the field of food science, where it is being utilised to address challenges ranging from quality assessment to predictive modelling. In our study, the evaluated Deep Learning (DL) algorithms are convolutional neural networks (CNN), visual transformers (ViT), and Graph Convolutional Networks (GCNs).

2.4.1. Convolutional Neural Networks - CNN encoders

Convolutional Neural Networks (CNN) are regularised feed-forward neural networks (Nebauer, 1998), suitable for a range of applications, that take their name from a mathematical linear operation between matrices called convolution. CNNs have multiple layers, including nonlinearity, pooling, convolutional and fully connected layers, with the latter two being parameterised (Albawi et al., 2017).

Here, we use two variations of the CNN Encoders — CNN Encoder 01 and CNN Encoder 02 (described below). These CNN encoder architectures feature a rich combination of fully connected (FC), batch normalisation (BN), and pooling layers to regress against the chickpea flour profiling tasks.

CNN Encoder 01 (CNN1): This model variation uses an approach similar to Yang et al. (2023) to enhance feature selection robustness and boost model generalisation. The model's progression from 100 to 250 out-channels through multiple convolutional layers, coupled with max

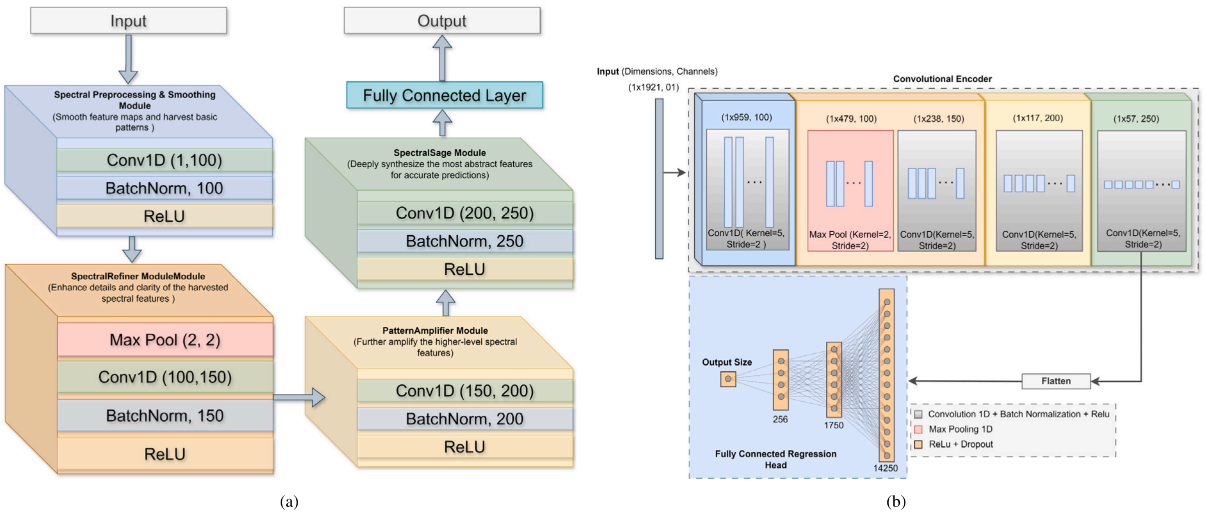
Input
Spectral Preprocessing & Smoothing Module
(Smooth feature maps and harvest basic patterns)
Conv1D (1,100)
BatchNorm, 100
ReLU
SpectralRefiner ModuleModule
(Enhance details and clarity of the harvested spectral features)
Max Pool (2, 2)
Conv1D (100,150)
BatchNorm, 150
ReLU
PatternAmplifier Module
(Further amplify the higher-level spectral features)
Conv1D (150, 200)
BatchNorm, 200
ReLU
SpectralSage Module
(Deeply synthesize the most abstract features for accurate predictions)
Conv1D (200, 250)
BatchNorm, 250
ReLU
Fully Connected Layer
Output
(a)
Input (Dimensions, Channels)
(1x1921, 01)
Convolutional Encoder
(1x959, 100)
Conv1D(Kernel=5, Stride=2)
(1x479, 100)
Max Pool (Kernel=2, Stride=2)
(1x238, 150)
Conv1D(Kernel=5, Stride=2)
(1x117, 200)
Conv1D(Kernel=5, Stride=2)
(1x57, 250)
Conv1D(Kernel=5, Stride=2)
Flatten
Output Size
256
1750
14250
Fully Connected Regression Head
Convolution 1D + Batch Normalization + Relu
Max Pooling 1D
ReLu + Dropout
(b)

Fig. 1. (a) Modular overview of the CNN Encoder Variation 01 and (b) Detailed end-to-end model architecture for CNN Encoder variation 01.

pooling, mirrors the approach of extracting detailed spectral features while maintaining model robustness. The Encoder type design captures intricate details in spectra (Jinadasa et al., 2021), ensuring the model's effectiveness in the regression analysis (like ours), where precision and adaptability are key.

Fig. 1(a) presents the overall modular architecture for this first variation. The model uses four key modules, named according to their functionalities in our spectral analysis context, and a fully connected layer to serve as a regression head. The description of each module is as follows:

1- *Spectral preprocessing and smoothing module:* The first convolutional layer is set to smooth the input spectrum and, therefore, reduce noise. Additionally, channel enhancement, in conjunction with the empirical configuration of the stride and kernel size, enables this module to harvest basic spectral shape patterns.

2- *SpectralRefiner module:* Following the initial feature extraction, the SpectralRefiner acts as both an amplifier and a detail enhancer. It starts with a max pooling operation to emphasise the most salient features, effectively "amplifying" the significance of the most pronounced elements by downsampling the data while retaining the maximum value within each pooling window. The subsequent convolutional layer further hones these features into a refined set, sharpening the feature maps and improving the clarity of the harvested spectral details.

3- *PatternAmplifier module:* Building upon refined features, the PatternAmplifier module is aptly named for its role in further enriching spectral features. The module increases the number of feature channels through its convolutional layer, allowing the network to develop a more complex and nuanced understanding of the spectral data. This convolution is paired with batch normalisation and ReLU activation to maintain healthy gradient flow and introduce the necessary non-linearity for capturing intricate spectral patterns.

4- *SpectralSage module:* As the final convolutional module before the transition to dense layers, the SpectralSage is tasked with the deep synthesis of the most abstract features necessary for accurate predictions. It acts as the wise consolidator of the network, taking the amplified and refined feature representations and distilling them into a form that is primed for the fully connected layers to interpret. This module embodies the culmination of the model's feature extraction prowess, setting the stage for precise and insightful regression or classification outcomes.

Fig. 1(b) illustrates the complete end-to-end forward pass of the model to map the input NIR signal with 1921 spectral values to the regression output.

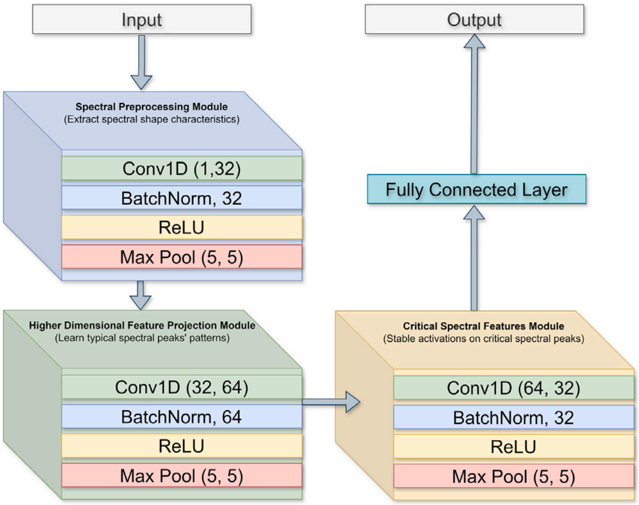
Input
Spectral Preprocessing Module
(Extract spectral shape characteristics)
Conv1D (1,32)
BatchNorm, 32
ReLU
Max Pool (5, 5)
Higher Dimensional Feature Projection Module
(Learn typical spectral peaks' patterns)
Conv1D (32, 64)
BatchNorm, 64
ReLU
Max Pool (5, 5)
Critical Spectral Features Module
(Stable activations on critical spectral peaks)
Conv1D (64, 32)
BatchNorm, 32
ReLU
Max Pool (5, 5)
Fully Connected Layer
Output

Fig. 2. Overview of the CNN Encoder Variation 02. The kernel size and padding pairs for the 1D convolutional layers are $(19, 9)$, $(25, 0)$, and $(21, 1)$, respectively. The stride is 1 for all these Convolutional layers.

CNN Encoder 02 (CNN2): This second variation of the CNN encoder is inspired by the customised CNN-based architecture used by Chadalavada et al. (2022). This variation aims to use fewer convolutional layers than the first variation, as recent studies have demonstrated that NIR spectral analysis only requires three convolutional layers (Zhang et al., 2020b; Cataltas and Tutuncu, 2023). Keeping this motivation in mind, we designed this minimalistic variation with only three modules (as depicted in Fig. 2), where convolution is the key for each module. The description of each module is as follows:

1. *Spectral processing module:* The first layer processes the spectra to enhance important features while reducing noise by acting akin to the Savitzky-Golay filter (Zhang et al., 2020b). This results in feature maps that retain the original spectrum's overall structure/ shape characteristics but are less cluttered and more distinguishable (due to max pooling), setting a robust foundation for subsequent, more intricate layers of analysis.

2. *Higher-dimensional feature projection module:* Employing a convolutional layer with higher-dimensional feature projection, this module captures and learns more intricate traits, such as typical spectral peaks and their patterns, which represent the more nuanced characteristics of the NIR spectral data. It refines the spectral data into a complex feature set, which, after stabilisation through batch normalisation and

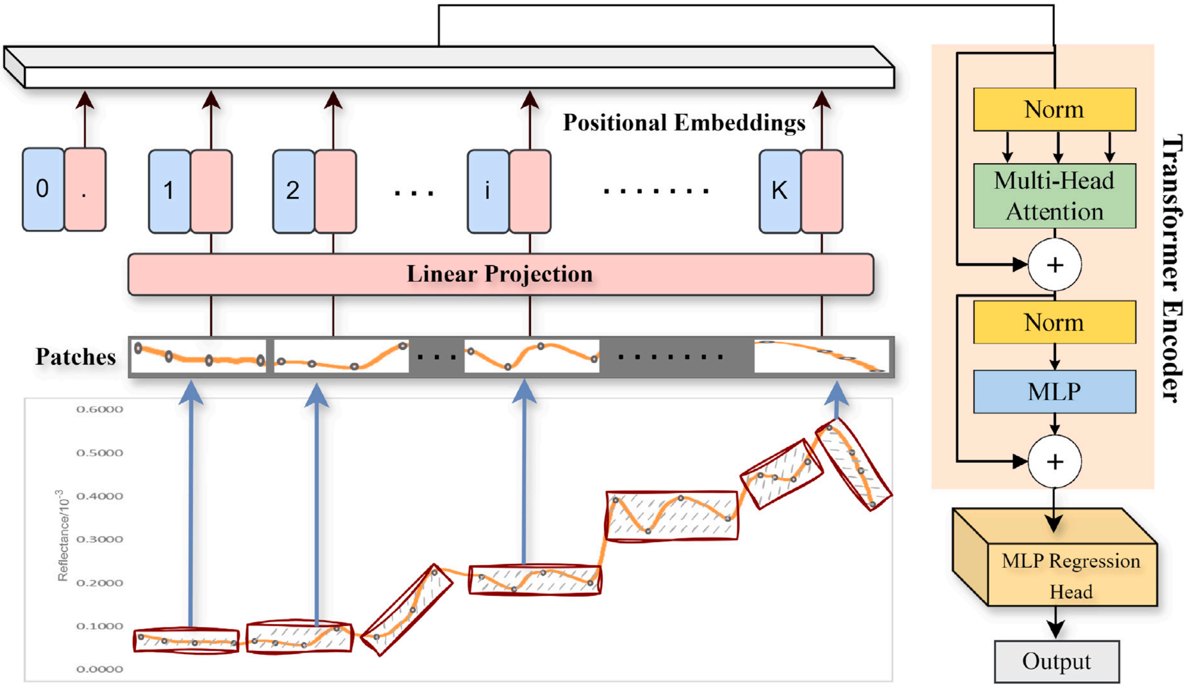

Fig. 3. Illustration on our formulation for feeding K patches of an NIR signal to Vision Transformer (ViT).

dimensionality reduction via max pooling, is primed for the third module's critical analysis of spectral peaks.

3. *Critical peak filtration module:* Interprets the refined feature set to isolate and improve key spectral peaks, employing max pooling for focused dimensionality reduction and feature prioritisation in preparation for the final predictive analysis. This encoder architecture uses a dense layer similar to CNN1 to serve as the regression head.

2.4.2. Vision Transformers - ViT

Vision Transformers (ViTs) are a class of deep learning models that have been adapted from the transformer architecture, originally developed for natural language processing (Vaswani et al., 2017), to address tasks in computer vision. By segmenting images in specialised applications like ours, NIR reflectance spectra—into a sequence of patches, ViTs can analyse complex visual data (Dosovitskiy et al., 2021). This approach allows them to capture both the local details and the broader context within the data, making them particularly effective for tasks that require an understanding of both fine-grained and holistic features and is particularly beneficial for analysing NIR spectra (Fu et al., 2022).

Local attention enables the model to detect subtle nuances within specific segments of the spectrum, which is critical for identifying the distinct spectral signatures (Liu et al., 2023). Meanwhile, its capacity to understand broader contexts integrates these local insights across the entire spectrum, allowing the model to recognise relationships from localised to global ones. This intricate balance between understanding local attention and retaining a satisfactory global context ensures a comprehensive analysis of the NIR spectra, suggesting the suitability of ViT for spectral analysis (Chen et al., 2024).

For this study, we shortlisted a set of hyperparameters to suit the unique characteristics of the NIR spectral data. The model's configuration includes an effective 'image size' of $1,921 \times 1$, representing each sample's NIR (Near-Infrared) spectrum, which is partitioned into patches of 10×1. These patches are then embedded into a 1024-dimensional space, enabling the model to capture an array of spectral information. The architecture of the transformer is defined by 6 layers, each with 8 attention heads, allowing for detailed attention mechanisms across different segments of the spectrum. An integral component of the model is its multi-layer perceptron (MLP) with 2048 nodes, which further processes the features extracted by the attention mechanism. Multiple dropout rates were tested to enhance generalisation and mitigate overfitting, where 0.1 was the best for both within the transformer layers and in the embedding stage. These parameter values were selected empirically based on the model-specific parameter tunning (see Model Selection Stage 2.5.1). Fig. 3 summarises our formulation for feeding an NIR signal to the Vision Transformer model.

2.4.3. Graph Convolutional Networks - GCN

Graph Convolutional Networks (GCNs) are a subclass of Graph Neural Networks (GNNs) that leverage the structural information inherent in graph data to perform deep learning tasks (Kipf and Welling, 2017). By extending the principles of convolutional operations from traditional data to graph-structured data, GCNs are capable of capturing the complex local and global interactions and dependencies between data points represented by nodes in a graph (Wu et al., 2023). Our formulation considers each spectral reading within a signal as a node in a graph, with nodes sequentially indexed from 1 to n, corresponding to the total number of spectral readings. To construct the graph (Fig. 4), each node is connected to its K-nearest neighbours on either side, establishing a local neighbourhood for every node. This connection is not binary but is quantified with weighted edges to reflect the relative position of each neighbour. For any given node X at index i, its connection to a neighbour Y at index k is determined by the edge weight $E(X,Y) = (k-i)^2$. This weighting scheme emphasises the proximity effect, where closer neighbours, in terms of spectral position, have a stronger influence on the node, represented by a lower edge weight, while distant neighbours exert a lesser influence, indicated by higher edge weights. Once a graph representation for the spectral signal is ready, it undergoes a series of graph convolutional operations through the model's layers. Our GCN comprises six convolutional layers, starting with an initial feature expansion to 256 dimensions in the first layer to capture a wide array of spectral features. Subsequent layers continue to process these features, with the final layers narrowing down to focus on the most relevant features for a specific target. The depth and design of the network, including the number of layers and their dimensions, were chosen based on empirical experimentation to best capture the complex relationships and patterns within the spectral data (see 2.5.1).

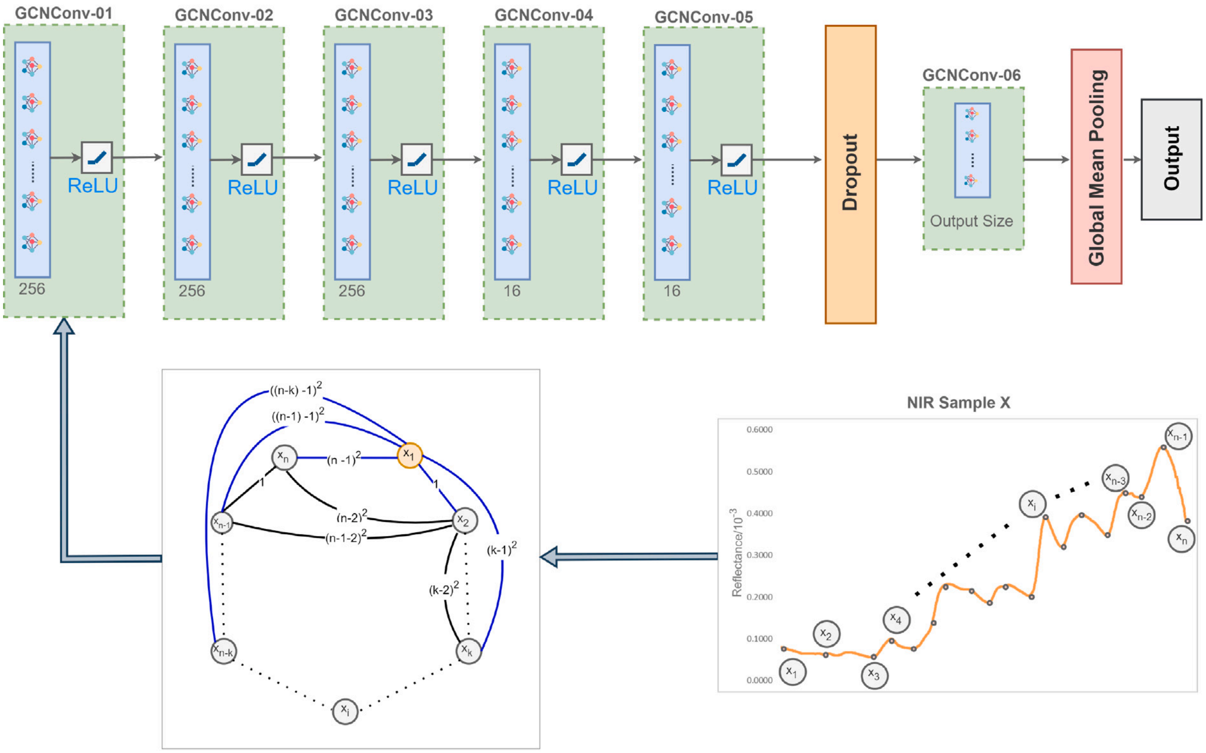

Fig. 4. Our formulation for mapping a spectral signal to a graph and then applying the Graph Convolutional Network (GCN) for regression.

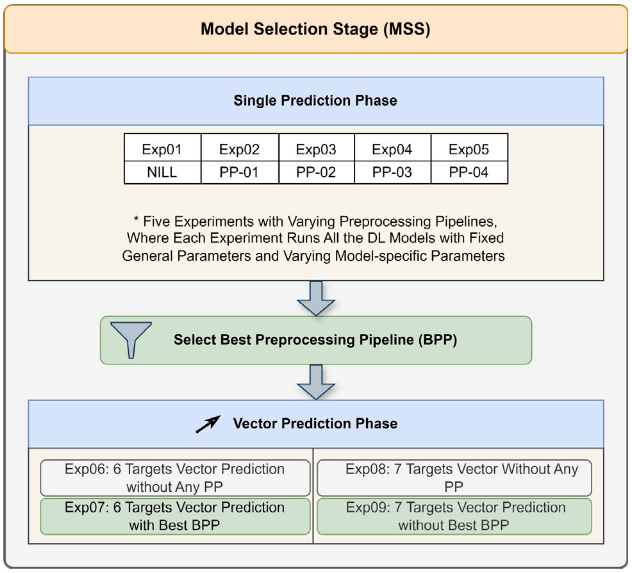
Model Selection Stage (MSS)
Single Prediction Phase
Exp01
NILL
Exp02
PP-01
Exp03
PP-02
Exp04
PP-03
Exp05
PP-04
* Five Experiments with Varying Preprocessing Pipelines, Where Each Experiment Runs All the DL Models with Fixed General Parameters and Varying Model-specific Parameters
Select Best Preprocessing Pipeline (BPP)
Vector Prediction Phase
Exp06: 6 Targets Vector Prediction without Any PP
Exp07: 6 Targets Vector Prediction with Best BPP
Exp08: 7 Targets Vector Without Any PP
Exp09: 7 Targets Vector Prediction without Best BPP

Fig. 5. Model Selection Stage (MSS): The MSS consists of two phases. In the Single Prediction Phase, five experiments (Exp01–Exp05) were conducted with different preprocessing pipelines (PP), except Exp01 which had no preprocessing (NILL). Models were trained for each target variable individually. After selecting the best preprocessing pipeline (BPP), the Vector Prediction Phase involved training models to predict multiple targets simultaneously, with and without the BPP.

2.5. Experiments and setup

Training a Deep Learning model involves optimising many general and model-specific hyperparameters. General hyperparameter optimisation includes (but is not limited to) optimising learning rate, epochs, step size, batch size, Adam optimiser-related parameters, random state of the split, activation, and loss functions. Moreover, model-specific hyperparameters like patch size and the number of attention heads in ViT, distance metrics, the number of convolutional layers and aggregation types in GCN, and the number of convolutional layers, kernel size, stride, dropout rate, and others specific to the CNN Encoder architecture further complicate the optimisation process. Additionally, preceding the training with different preprocessing pipelines adds another layer of complexity.

Due to the extensive number of hyperparameters, both general and model-specific, combined with multiple preprocessing pipelines, fully optimising these parameters across all models would require immense computational resources, making such exhaustive optimisation practically unfeasible. Instead of this exhaustive training, we divided our deep learning experiments into two stages: (1) the Model Selection Stage and (2) the Extensive Evaluation Stage.

2.5.1. Model Selection Stage (MSS)

This stage aimed to shortlist a model and a preprocessing pipeline for the subsequent exhaustive hyperparameter optimisation and extensive cross-validation stage (i.e., EES) by evaluating models using simple Train/Test Split training. Here, we kept general hyperparameters constant across the experiments and varied only the preprocessing pipelines and some of the model-specific parameters. Table 2 lists all the general hyperparameters and ranges of model-specific hyperparameters used in this stage.

To further simplify the selection stage, we split it into two phases: (1) Single Prediction and (2) Vector Prediction. The single prediction phase involved separately training all models for each output variable (i.e., protein content, starch content, soluble sugars, insoluble fibres, total lipids, and moisture). In vector prediction, a single model is trained to predict all six output targets simultaneously, producing a vector of regression values, each corresponding to one of the targets. This phase aimed to test whether we can predict all six targets using a single model as correctly as we can using a separate model for each target type.

Fig. 5 illustrates our entire experimental setup for MSS. The single prediction phase involved five experiments, each running all the

Table 2
Details of general and model-specific hyperparameters used in the MSS. General hyperparameters were frozen while different combinations of model-specific hyperparameters were tried.

General Hyperparameters	
Learning Rate	0.0001
Activation	ReLu
Optimiser	Adam (betas = (0.9, 0.999), eps = 1e−08, weight_decay = 0)
Split type	train_test_split
Epochs	4000
Split Size	(80%, 20%)
Split Rand State	42
Batch Size	4
Loss Function	MSELoss()
CNN Encoders (CNN Encoder 01 & CNN Encoder 02)	
Kernel Sizes	[5, 7, 9, …, 21, 23, 25]
Strides	[1, 2]
Layers in FC Head	[2, 3, 4, 5]
FC Dropout Rates	[0.1, 0.3, 0.5, 0.7]
Graph Convolutional Network (GCN)	
GCNConv Layers	[1, 3, 5, 7]
Node Neighbours	[1, 3, 5, 7, 11]
Dropout Rates	[0.1, 0.3, 0.5, 0.7]
Pooling	Global Mean Pooling
Vision Transformer (ViT)	
Dropouts	[0.1, 0.3, 0.5, 0.7]
Heads	[2, 4, 6, 8]
MLP Dims	[512, 1024, 2048]
Attention Layers	[2, 4, 6, 8]
Patch Sizes	[(5,1), (10, 1), (20, 1), (30, 1)]
Patch Embedding Dims	[512, 1024, 2048]

models. Each experiment had a different preprocessing pipeline, except Exp01, where no preprocessing was used. For each experiment in this phase, we kept the general hyperparameters constant and used multiple combinations of model-specific hyperparameters for each model (see Table 2). Once the Single Prediction Phase was complete, we analysed the results and selected the pipeline that performs better for most of the scenarios. Then, we used this best preprocessing pipeline (BPP) to enter the Vector Prediction Phase.

The vector prediction phase involved four experiments: Exp06, Exp07, Exp08, and Exp09. Exp06 and Exp07 run all the models with fixed general and varied model-specific hyperparameters to predict all the six outputs (i.e., protein content, starch content, soluble sugars, insoluble fibres, total lipids, and moisture) when no processing is done on the data and when BPP is applied, respectively. Exp08 and Exp09 are similar to Exp06 and Exp07 but have an added output generated by the remaining percentage of grain Remaining (%) $= 100 - \sum_{i=1}^{6} ov_i$. Here, ov_i is the output value of each output for the ith sample. The purpose of such formulation was to predict the “unaccounted for” proportion of the grain because, aside from potential error in the measurements, that gap should come from mineral components (also referred to as ash) and soluble fibre.

2.5.2. Extensive Evaluation Stage (EES)

This stage involved a 10-fold exhaustive cross-validation of the model shortlisted in the previous stage. Since we had already identified the optimal model-specific hyperparameter configuration and the best preprocessing pipeline (BPP) from the MSS results, we only varied the general hyperparameters during this stage (see Table 3). Additionally, we ran a 10-fold cross-validation of the partial least squared regression (PLSR) model with the same BPP to establish a fair comparison of this traditional regression method with the shortlisted deep learning model. Fig. 6 shows a complete experimental setup illustrating the distinct aims of the MSS and EES.

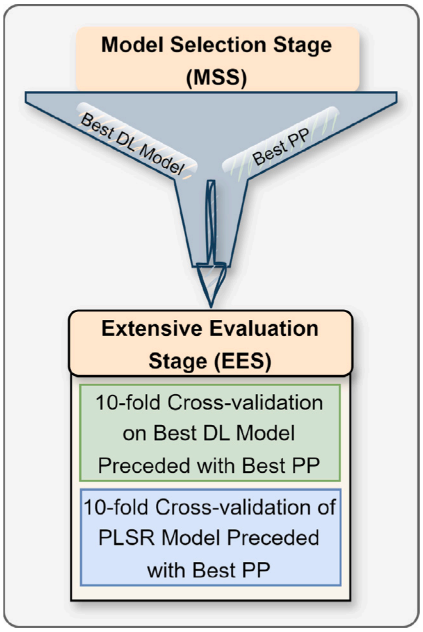
Model Selection Stage (MSS)
Best DL Model
Best PP
Extensive Evaluation Stage (EES)
10-fold Cross-validation on Best DL Model Preceded with Best PP
10-fold Cross-validation of PLSR Model Preceded with Best PP

Fig. 6. Extensive Evaluation Stage (EES): The best DL model identified from the MSS was extensively evaluated in this stage. It underwent general hyperparameter tuning and was compared with the PLSR model, which was tuned using PLSR-specific hyperparameters. Both models were subjected to 10-fold cross-validation. In this stage, data were processed through the BPP selected during the MSS before being fed into the models.

Table 3
Shows best hyperparameters for the models used in EES. A range of general hyperparameters was tried for CNN1 while keeping model-specific hyperparameters constant. However, PLSR (a non-DL method) does not have equivalent general hyperparameters. Therefore, we only tried a range of hyperparameters specific to PLSR. A bold-face value in a range is the one with the best results in the EES.

General Hyperparameters Ranges for CNN1	
Learning Rate	[0.1, 0.01, **0.001**, 0.0001]
Activation	ReLu

Optimiser	Adam (
	betas = [(0.9, 0.99), **(0.9, 0.999)**],
	eps = [1e−07, **1e−08**, 1e−09],
	weight_decay = [**0**, 1e−4, 5e−4, 1e−5])

Split type	KFolds(n_splits=10, random_state=42, shuffle=True)
Epochs	6000
Early Stop	[True, **False**]
Batch Size	[2, **4**, 8, 16]
Loss Function	MSELoss()
Best CNN1-specific Hyperparameters	
Kernel Size	5 for each Conv layer
Stride	2 for each Conv layer
Layers (FC Head)	3
FC Dropout	0.5
Ranges of PLSR-specific Hyperparameters	
N Components	[2, 3, **5**, 7, 11, …, 29, 31]
Max Iterations	[**500**]
Tolerence	[1e−2, 1e−3, 1e−4, 1e−5, **1e−6**]
Scaling	[**True**, False]

3. Results

3.1. Results of MSS

3.1.1. Best preprocessing scheme

Analysis of the results from the single prediction phase during the MSS (Table 4) revealed that no single preprocessing pipeline consistently outperformed the others across all cases. Each preprocessing

Table 4
Results of experiments for the MSS. BM: Best Model that yielded the best R2 score; CNN1: CNN Encoder variation 01; CNN2: CNN Encoder variation 02; ViT: Vision Transformer; GCN: Graph Convolutional Network.

		Protein (%)		Starch (%)		Soluble sugar (%)		Insoluble fibres (%)		Total lipids (%)		Moisture by mass (%)	
		BM	R^2	BM	R^2	BM	R^2	BM	R^2	BM	R^2	BM	R^2
Single Prediction	**Exp01**	CNN1,2	0.60	ViT	0.05	CNN1	0.13	CNN2	0.30	CNN1	0.47	CNN1	0.71
	Exp02	CNN2	0.04	ViT	0.04	CNN2	0.18	CNN1	0.18	CNN2	0.56	CNN1	0.54
	Exp03	ViT	0.60	ViT	0.00	GCN	0.27	CNN2	0.15	CNN2	0.42	CNN2	0.60
	Exp04	ViT	0.41	CNN2	0.00	CNN1	0.12	GCN	−0.03	CNN2	0.35	CNN2	0.59
	Exp05	CNN1,2	0.55	GCN	0.15	CNN1	0.18	CNN2	0.30	CNN1,2	0.51	CNN1, ViT	0.72
Vector Prediction	**Exp06**	CNN1	0.36	GCN	0.01	CNN1	0.02	CNN1	0.026	CNN1	0.21	ViT	0.00
	Exp07	CNN1	0.30	GCN	0.03	CNN2	0.08	ViT	−0.10	GCN	0.00	ViT	0.03
	Exp08	CNN1	0.24	GCN	0.03	CNN1	0.01	CNN1	0.01	GCN	0.01	ViT	−0.04
	Exp09	CNN1	0.34	GCN	0.02	CNN1	0.08	ViT	−0.01	GCN	0.03	GCN	−0.15

pipeline had distinct strengths and weaknesses depending on the specific target quality parameter. PP-01 (Exp02) showed minimal improvements across targets, struggling particularly with Starch and Protein. This is because the rigid global min–max scaling suppressed the overall variability, while the Savitzky-Golay filter in this pipeline further smoothened the local intricacies. PP-02 (Exp03) provided moderate improvement on targets like Soluble Sugar and Protein but often over-processed the data, leading to inconsistencies, particularly for Starch. PP-03 (Exp04) performed poorly in all targets, demonstrating that its minimal preprocessing steps were insufficient to prepare the data for model training, especially for complex features such as starch and insoluble fibres. Compared to no-preprocessing, PP-03 performed worse as it just excessively increased the variability in the data, leading to more complex patterns on which models struggled to converge.

However, PP-04 balanced the need for scaling and feature enhancement, leading to consistent improvements across most targets, as demonstrated by Exp05. The translation and squaring steps in PP-04 proved particularly beneficial in preserving important details and capturing nonlinear patterns that were missed by the other pipelines. Even when compared to the no-preprocessing scenario, which showed strong results for a few targets or quality parameters like Protein, PP-04 outperformed by providing more reliable and balanced improvements across all targets, making it the most robust and versatile preprocessing pipeline in this study. Based on these observations, PP-04 was shortlisted as the best-performing preprocessing (BPP) approach for the vector prediction phase of MSS. In this phase, we conducted vector prediction experiments using both PP-04 and the option of no preprocessing to ensure fair evaluation.

Vector prediction performed worst (Exp06 to 09), indicating that multi-task modelling is still challenging in this context. Comparing all four DL models across all nine experiments allowed the determination of the best-performing algorithm for each output. The results reveal that the CNN Encoder-based architectures generally outperformed other models, effectively capturing major variations even without preprocessing (Exp01). Specifically, CNN Encoder 01 (CNN1) demonstrated the most consistent performance across various targets, making it the preferred choice for further analysis.

3.1.2. Best model

The CNN encoder models' performance supersedes the rest of the architectures, as clearly evident by the results shown in Table 4. For this study, the superior performance of CNNs can be attributed to their inherent ability to extract localised features from one-dimensional signals. In the context of NIR spectroscopy data, the convolutional filters in CNNs effectively capture subtle local spectral variations and reduce noise, which is crucial for accurate compositional analysis. In contrast, ViTs rely on a patch-based self-attention mechanism that is more effective for data with strong spatial structure. Moreover, ViTs are infamous for requiring large datasets to converge. Given that this study uses a tiny dataset, it is no surprise that it lagged behind CNNs on this particular dataset. GCNs require an explicit graph representation and, if not formulated properly, may not fully capture the continuous nature of spectral information. Additionally, over-smoothing, an intrinsic issue with graph convolutional networks (GCNs), can significantly cause these architectures to diminish the subtle nuances of the data along with the noise, making these networks follow the very mean trend. Therefore, our GCN architecture might have struggled in generalisation, causing a low performance on unseen instances.

3.2. Results of EES

CNN Encoder 01 (CNN1) emerged as the best overall model from the MSS, with PP-04 identified as the most effective preprocessing pipeline. This best-performing DL method (CNN1) also outperformed the traditional method (PLSR) with an overall smaller average RMSE, larger R2, and less variation across all target variables (Figs. 7 and 8). This is a promising result as this work has limited data (a total of 136 observations only) compared to other DL developments and yet was able to outperform traditional predictive methods for NIR analysis on chickpea flour quality. Albeit the performance is not enough for industry application as is, it is expected that the DL techniques will further improve if more data is made available.

The improvement compared to PLSR varied depending on the target. The biggest improvement was observed in soluble sugars (%), where the variation of R2 is greatly diminished while its value increased from an average of 0.0 to 0.2 (Fig. 7) and a reduction of RMSE from an average of 1.0 to 0.8 with similar variation (Fig. 8). The smallest improvement was on moisture (%) with almost no change in RMSE, about 0.2 for both (Fig. 8), and a small increase in R2 from 0.7 to 0.8 (Fig. 7). The RMSE of moisture was smaller than the standard deviation of the data, which indicates that model performance is unlikely to improve further.

4. Discussion

This study provides a comprehensive overview and an approach to the application of AI to Near-Infrared Spectroscopy (NIR) to enhance the assessment of chickpea flour composition. This integration marks a key advancement in addressing the inherent variability and quality control challenges in chickpea flour production, which is crucial for meeting the burgeoning demand for plant-based food alternatives. Although chickpea serves as an important test case, the implications of this research extend beyond chickpeas and could be transformative for the evaluation of other pulse crops, including lentils, peas and beans, which share similar compositional attributes and are also important in the global food supply.

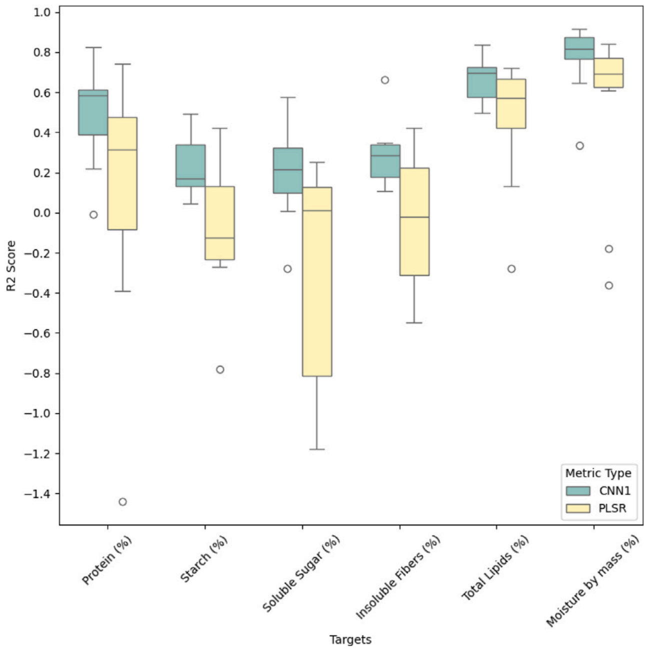

Fig. 7. Coefficient of Determination (R2 score) for each target comparing the CNN encoder variation 1 (CNN1) and the partial least square regression (PLSR) models' performance with 10-fold cross-validation. Table 3 details the best hyperparameters that yielded these results. For exact max, min, and average metrics, see Table B.5 in Appendix B.

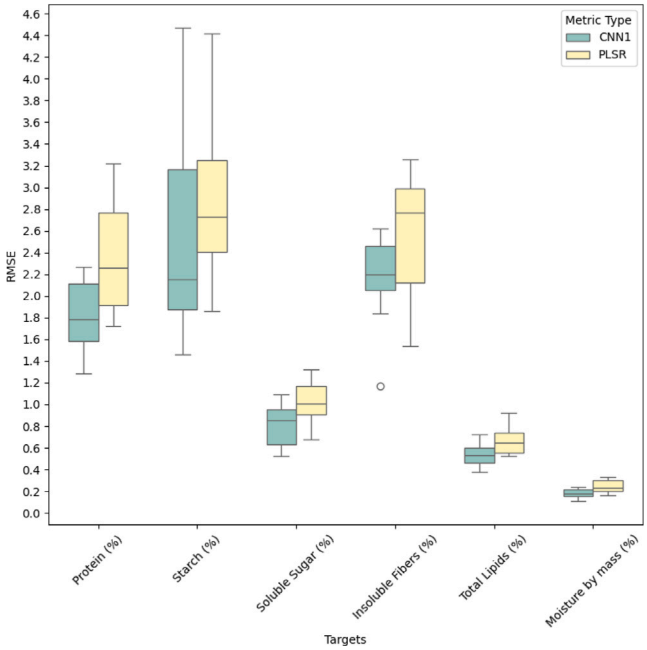

Fig. 8. Root Mean Square Error (RMSE) for each target comparing the CNN encoder variation 01 (CNN1) and the partial least square regression (PLSR) models' performance using a 10-fold cross-validation.

4.1. Dataset characteristics and challenges

The dataset used in this study comprises a diverse collection of chickpea samples, reflecting a broad spectrum of genetic variation. This diversity is crucial for developing a robust AI model capable of accurately predicting quality attributes across different chickpea varieties. Each sample in the dataset is associated with comprehensive NIR spectral data and proximal composition, including protein, fibre, starch soluble sugar, lipid content, and moisture levels. The diversity of the samples used in the generation of the dataset is instrumental in training AI models that can handle the inherent variability in chickpea flour's composition.

One of the significant challenges highlighted in the study is the limited size of the dataset. With AI, especially deep learning models, the quantity and quality of data are pivotal for model performance. A limited dataset can restrict the model's ability to learn complex patterns and generalise to new, unseen data. Moreover, the chickpea data used had issues related to noise within the NIR spectral data, which can further complicate model training and accuracy. Noise in the data can stem from various sources, including instrument calibration errors or inconsistencies in sample preparation.

To tackle this challenge, data augmentation techniques (e.g., generative adversarial networks or diffusion models) and transfer learning approaches can be employed to enrich and expand the effective training set. Such strategies are expected to enhance the robustness of the models against unseen data by providing a richer representation of the underlying variability.

4.2. AI-enhanced NIR spectroscopy versus traditional PLSR

The empirical results of our research demonstrate improvements in AI-enhanced NIR spectroscopy over traditional Partial Least Squares Regression (PLSR) methods. Specifically, our findings revealed that a Convolutional Neural Network (CNN) Encoder outperformed PLSR across all target variables (Figs. 7 and 8), offering a more accurate, non-destructive means of assessing key quality attributes of chickpea flour such as protein content, moisture levels, and fibre composition. This improvement in predictive accuracy and efficiency can be attributed to the deep learning model's adeptness at handling the high-dimensional and complex data associated with NIR spectroscopy. By extracting nuanced patterns and features from the spectral data, the AI models facilitate a more refined analysis, overcoming the limitations inherent in traditional chemometric approaches.

4.3. Implications for the plant-based food industry

Reliable, high throughput, non-destructive composition characterisation approaches, such as NIR, provide significant opportunities to the food sector. NIR is routinely used for the characterisation of wheat grain composition (Zhang et al., 2022b), allowing for quality metrics to be defined by composition traits. The application of composition-based quality metrics in chickpeas could assist growers with the reduced variability in valuing that often occurs as a result of the subjectivity of the assessors. Wheat also acts as a point of reference for further analyses that can be facilitated by NIR analysis, such as assessing safety parameters or contaminants (Zhang et al., 2022b; Badaró et al., 2022). Expanding the training of models such as those presented here to datasets beyond proximal composition in chickpeas could support the identification of more specific traits such as individual proteins or lipids.

In the food sector, NIR composition assessment has significant potential for recipe formulation, generating repeatable, reliable products (Wang et al., 2024; Silaghi et al., 2010). For chickpeas, this holds promise not only in recipe formulation but also in the development of protein products (Ingle et al., 2016; Neves et al., 2022). Bypassing the need for time-intensive, destructive composition measures should help to support the rapid uptake of chickpea protein ingredients. This is further supported by the ability to implement NIR online in production systems (Porep et al., 2015; Huang et al., 2008), which facilitates seamless integration with production systems for data collection and quality control.

Although NIR approaches are used routinely and have proven their value in crops such as wheat (Zhang et al., 2022b; Badaró et al., 2022;

Schuster et al., 2023) and soy (Neves et al., 2022), the expansion of techniques to new crops, such as chickpeas, will facilitate reliable quality assessment and recipe formulation in more settings. In addition, the scalability and cost-effectiveness of this method contribute to improving the overall efficiency of food production systems, aligning with the global shift towards sustainable food practices.

The fundamental concepts of the AI-enhanced NIR approach are not exclusive to chickpeas; they are equally relevant to other pulses, including lentils, peas and beans, which also show substantial differences in protein, fibre and moisture levels. The nutritional value of these pulses, particularly their protein content, is increasingly acknowledged, making them essential components of plant-based diets. Despite their nutritional importance, current market practices prioritise grain size and colour, largely due to the lack of reliable methods for rapid composition analysis. The AI-enhanced NIR spectroscopy method demonstrated in this study has the potential to revolutionise the way pulses are evaluated and priced, shifting the focus towards nutritional quality attributes. This shift could drive innovation in product formulation and significantly enhance the value proposition of pulse-based ingredients in the food industry.

4.4. Future research directions

The study's outcomes highlight several avenues for further research that could extend the applicability and effectiveness of AI-powered quality assessment methods:

Exploration of Advanced Data Augmentation and Noise Reduction Techniques: The study highlights the limitations posed by a relatively small dataset. Data augmentation techniques such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models could be investigated to generate synthetic data. These synthetic data could improve deep learning models training by providing a more extensive and diverse dataset, potentially leading to better model performance and generalisation. Improved predictive accuracy and robustness of deep learning models for chickpea flour quality assessment, enabling more reliable applications in industrial settings.

Though this study used a mix of standard preprocessing techniques (scaling, normalisation, log transformation, basic translation, and Savitzky-Golay filtering), incorporating advanced scatter correction methods such as wavelet-based denoising (Shin et al., 2010; Molavi and Dumont, 2012; Sahoo et al., 2024; Ye and Ma, 2024) and adaptive baseline correction methods (e.g., Zhang et al. (2010, 2020a), Li et al. (2023)) can further be evaluated to reduce noise and correct baseline shifts in NIR data.

Integration of Multi-Modal Data Sources: Future research could explore the integration of multi-modal data sources, such as combining Near-Infrared (NIR) spectroscopy with other spectroscopic techniques (e.g., Raman, FTIR) or imaging technologies (e.g., hyperspectral imaging). The incorporation of these complementary spectroscopy methods is a promising and feasible approach to further improve predictive accuracy. These modalities provide orthogonal chemical and structural information that can complement NIR spectra. Multi-modal data fusion — achieved through feature-level or decision-level integration — can capture a more comprehensive chemical profile (Ye et al., 2022). However, Integrating multi-modal data can significantly increase computational cost and model complexity, potentially hindering real-time applications and requiring specialised hardware or optimised algorithms. Balancing model complexity with computational efficiency is crucial for practical implementation, especially in industrial settings where rapid analysis is often required. This might involve exploring lightweight deep learning architectures or dimensionality reduction techniques to reduce computational burden without compromising predictive accuracy.

While incorporating Raman or FTIR spectroscopy alongside NIR shows promise for improving chickpea flour quality predictions, feasibility depends on addressing practical challenges. Raman and FTIR offer complementary information about molecular vibrations and functional groups, respectively, potentially enhancing model accuracy through data fusion techniques like multi-block analysis (Ye et al., 2022). However, this increases experimental costs and complexity, potentially hindering industrial applications. Ensuring accurate data alignment and registration across different techniques is crucial, given variations in spectral resolution. Furthermore, increased model complexity from multi-modal data may hinder interpretability, requiring careful model selection and potentially explainable AI techniques.

Deep Learning Model Interpretability and Explainability: As deep learning models become more complex, ensuring their interpretability and explainability becomes critical, especially for regulatory compliance and industry adoption. Future studies could focus on developing interpretable deep learning models or enhancing existing models with explainable AI (XAI) techniques. This would help stakeholders understand how specific predictions are made, which is essential for trust and transparency. Increased adoption of AI-based quality assessment in the food industry, driven by enhanced trust in the predictive models.

Longitudinal Studies on Environmental Influence: The current study mentions the variability in chickpea flour quality due to environmental influences. Future research could involve longitudinal studies that monitor how different environmental conditions (e.g., soil type, climate variations) affect chickpea flour composition over multiple growing seasons. This data could be used to refine deep learning models, making them more adaptive to varying environmental conditions. Development of deep learning models that can adapt to environmental variability, leading to more consistent quality in chickpea flour products.

Deep Learning Optimisation of Pulse Breeding Programmes: Deep learning models could be applied to optimise pulse breeding programmes by predicting which genetic variations are most likely to result in desirable properties (e.g., high protein content and low antinutrients). This research direction would involve collaboration between geneticists, agronomists, and AI specialists. Accelerated development of pulse varieties tailored for specific food applications, contributing to the efficiency of breeding programmes and the availability of superior pulse-based ingredients.

Scalability and Industrial Application of Deep Learning Models: Future studies could focus on the scalability of AI models for industrial applications. This includes the development of lightweight, real-time deep learning models that can be implemented directly on production lines for continuous quality monitoring. Research should also explore the integration of deep learning models into existing manufacturing processes and quality control systems. Increased efficiency and cost-effectiveness in chickpea flour production, with real-time quality control allowing rapid adjustments and reduced waste.

Although our CNN models show promising improvements over traditional PLSR methods, their current accuracy — particularly for parameters such as soluble sugars and starch — remains below the thresholds required for industrial application. For deployment in real-world settings, we envision the need to achieve prediction accuracies corresponding to R^2 values consistently above 0.9 for key quality parameters. Moreover, operational scalability will require the integration of real-time, online processing capabilities, enhanced computational hardware for rapid inference, and the use of larger, more diverse datasets to ensure robust performance under varying production conditions.

Investigating AI-Assisted Food Safety and Contaminant Detection: In addition to quality assessment, deep learning models could be further developed to detect food safety issues, such as contamination by pathogens or the presence of harmful chemicals. Research in this area could involve training AI models to recognise spectral signatures associated with various contaminants in chickpea flour. Enhanced food safety, with AI providing a non-destructive, rapid method for detecting contaminants during production.

Consumer Acceptance and Market Implications: As AI technologies are increasingly integrated into food production, understanding consumer acceptance and market implications becomes essential. Future research could explore how consumers perceive AI-driven quality assessments and how these perceptions influence their purchasing decisions. Better alignment of AI technologies with consumer expectations, leading to greater market acceptance and successful commercialisation of AI-enhanced food products.

5. Conclusion

This study demonstrates the potential for integrating artificial intelligence, particularly deep learning models, with Near-Infrared (NIR) spectroscopy to assess the quality of chickpea flour. The findings reveal that deep learning models, especially convolutional neural networks (CNNs), when compared to Partial Least Squares Regression (PLSR), provided more accurate and reliable predictions of chickpea's key quality attributes such as protein content, starch, soluble sugars, insoluble fibres, total lipids, and moisture levels. Furthermore, these methods apply to other pulses, such as lentils, peas, and beans. The study also highlights the need for further research to enhance the scalability and industrial applicability of such deep learning models.

CRediT authorship contribution statement

Ali Zia: Conceptualization, Methodology, Project Administration, Software, Writing – Original Draft, Writing – review & editing. **Muhammad Husnain:** Formal analysis, Investigation, Writing – review & editing. **Sally Buck:** Data curation, Resources, Validation. **Jonathan Richetti:** Visualization, Writing – review & editing. **Elizabeth Hulm:** Data curation, Resources, Validation. **Jean-Philippe Ral:** Supervision, Funding acquisition, Writing – review & editing. **Vivien Rolland:** Supervision, Funding acquisition, Writing – review & editing. **Xavier Sirault:** Supervision, Funding acquisition.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, no generative AI or AI-assisted technologies were used to create the content of this manuscript. The content is entirely original and was authored by the individuals listed. To improve the readability and language quality of certain parts of the paper, co-authors may have used AI-assisted tools like Grammarly.

Declaration of competing interest

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

The research presented in this manuscript received support from the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia, specifically for data collection efforts. The funders did not have any role in the study design, analysis, or interpretation of data, the writing of the manuscript, or the decision to submit the paper for publication.

Appendix A. Preprocessing pipelines

Each pipeline was designed to address specific challenges inherent in our dataset, such as noise, variance normalisation, and scaling. This appendix presents the visualisations of our four preprocessing pipelines (PP-01 to PP-04) to better understand how each transforms the NIR data, addressing a specific challenge.

PP-01: The first preprocessing pipeline (Fig. A.9) combines Standard Scaling, Global MinMax Scaling, and Savitzky-Golay (SG) Smoothing. This approach is particularly effective for datasets with significant variations in feature magnitudes, as it normalises the data by reducing the impact of extreme range differences, followed by SG smoothing to reduce spectral noise. This ensures that the model does not become biased towards higher magnitude features and enhances the interpretability of the spectra.

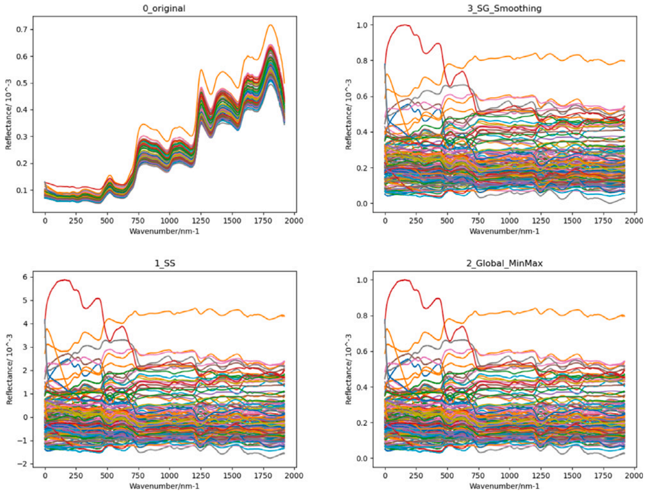

Fig. A.9. Preprocessing Pipeline PP-01.

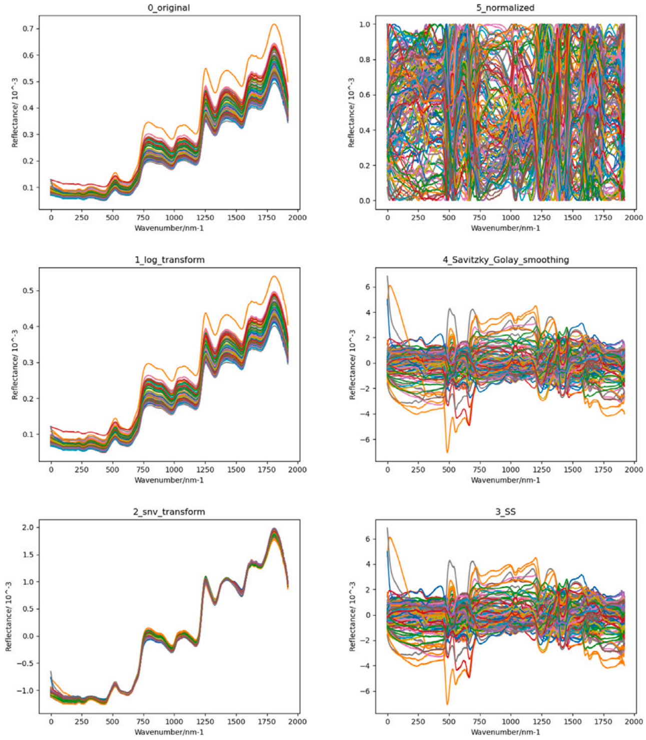

Fig. A.10. Preprocessing Pipeline PP-02.

PP-02: The second preprocessing pipeline (Fig. A.10) involves a Log Transformation, Standard Normal Variate (SNV) correction, Standard Scaling, SG Smoothing, and Normalisation. This intensive sequence is designed to thoroughly prepare spectral data by reducing skewness, normalising variance, and enhancing signal quality, making it particularly suited for datasets with severe anomalies or outliers. The processed spectra show consistent patterns, indicating effective handling of irregularities.

PP-03: The third preprocessing pipeline (Fig. A.11) is a simpler approach that includes Standard Scaling followed by Normalisation. This basic preprocessing ensures that all features have a mean of 0 and a

Table B.5
CNN1 and PLSR 10-fold cross-validation summary.

		R2			RMSE		
		Best	Worst	Average	Best	Worst	Average
CNN1	**Protein (%)**	0.822	−0.007	0.494	1.284	2.264	1.823
	Starch (%)	0.493	0.046	0.219	1.463	4.468	2.595
	Soluble Sugar (%)	0.576	−0.278	0.197	0.527	1.089	0.815
	Insoluble Fibres (%)	0.663	0.107	0.291	1.171	2.617	2.150
	Total Lipids (%)	0.836	0.497	0.667	0.381	0.721	0.534
	Moisture by mass (%)	0.914	0.337	0.771	0.108	0.236	0.179
PLSR	**Protein (%)**	0.741	−1.438	0.098	1.718	3.217	2.341
	Starch (%)	0.417	−0.777	−0.090	1.860	4.417	2.923
	Soluble Sugar (%)	0.253	−1.180	−0.316	0.678	1.321	1.008
	Insoluble Fibres (%)	0.419	−0.549	−0.043	1.540	3.261	2.599
	Total Lipids (%)	0.716	−0.278	0.458	0.518	0.916	0.662
	Moisture by mass (%)	0.839	−0.359	0.529	0.159	0.326	0.245

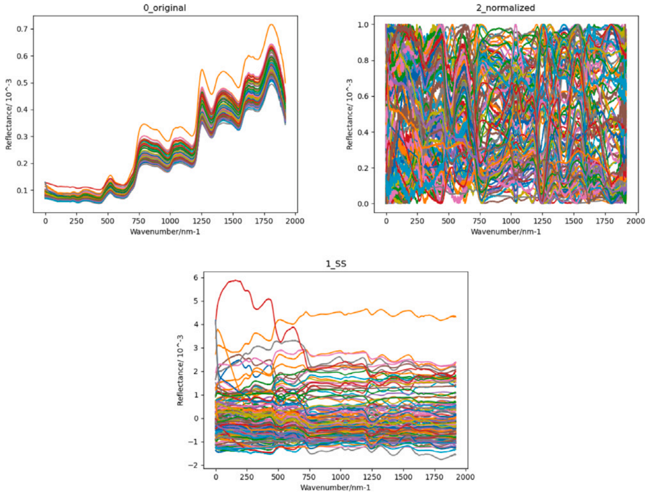

Fig. A.11. Preprocessing Pipeline PP-03.

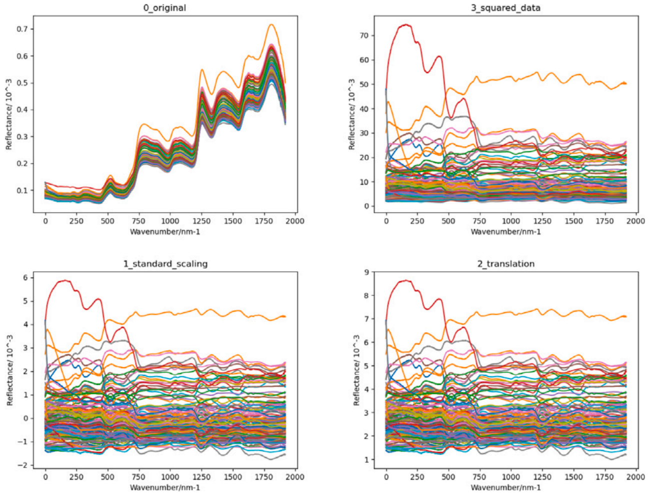

Fig. A.12. Preprocessing Pipeline PP-04.

standard deviation of 1, with the normalisation rescaling the features to a [0, 1] range. This method effectively reduces biases in gradient-based models due to varying feature scales, facilitating efficient learning and improving model generalisation.

PP-04: The fourth preprocessing pipeline (Fig. A.12) involves Standard Scaling, Translation ($value - |min| + 1$), and Squaring the resultant. This pipeline standardises the data and translates it to ensure all values are positive and greater than 1. The reason behind this unconventional translation is that if values were less than 1, their square would have been smaller and would result in actually smoothing out the local intricate distances. The squares transformation amplifies differences between larger values while preserving details in smaller values. This preprocessing biases the models to capture nonlinear relationships, which is crucial for the accuracy of deep learning models.

These preprocessing pipelines were systematically evaluated to determine their effectiveness in improving the quality of our NIR data. Though there was no clear winner in performing better over all the six targets, PP-04 still appeared to be a more effective strategy, leading to superior performance (on average) on all the targets. While it was not always the top performer for every target, its consistent ability to deliver reasonable R2 scores across the board made it the most reliable choice for this study (see 3.1 for more details).

Appendix B. Detailed metrics for CNN01 and PLSR

Detailed 10-cross validation results for CNN1 and PLSR corresponding to Figs. 7 & 8.

Data availability

Data will be made available on request.

References

Aghdamifar, E., Rasooli Sharabiani, V., Taghinezhad, E., Rezvanivand Fanaei, A., Szymanek, M., 2023. Non-destructive method for identification and classification of varieties and quality of coffee beans based on soft computing models using VIS/NIR spectroscopy. Eur. Food Res. Technol. 249 (6), 1599–1612.

Albawi, S., Mohammed, T.A., Al-Zawi, S., 2017. Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology. ICET, IEEE, pp. 1–6.

American Association of Cereal Chemists. Approved Methods Committee, 2000. Approved Methods of the American Association of Cereal Chemists v. 1-2, AACC, URL https://books.google.com.au/books?id=xJwQAQAAMAAJ.

Badaró, A.T., et al., 2022. Near infrared techniques applied to analysis of wheat-based products: Recent advances and future trends. Food Control 140, 109115. http://dx.doi.org/10.1016/j.foodcont.2022.109115.

Bar-El Dadon, S., Abbo, S., Reifen, R., 2017. Leveraging traditional crops for better nutrition and health - The case of chickpea. Trends Food Sci. Technol. 64, 39–47.

Bravo-Núñez, A., Gómez, M., 2021. Enrichment of cakes and cookies with pulse flours. A review. Food Rev. Int. 39 (5), 1–19. http://dx.doi.org/10.1080/87559129.2021.1983591.

Buck, S., Ral, J.-P., 2024. Proximal Grain Composition of 240 Globally Diverse Chickpea Lines. v1. Data Collection, CSIRO.

Cataltas, O., Tutuncu, K., 2023. Detection of protein, starch, oil, and moisture content of corn kernels using one-dimensional convolutional autoencoder and near-infrared spectroscopy. PeerJ Comput. Sci. 9, e1266.

Chadalavada, K., Anbazhagan, K., Ndour, A., Choudhary, S., Palmer, W., Flynn, J.R., Mallayee, S., Pothu, S., Prasad, K.V.S.V., Varijakshapanikar, P., Jones, C.S., Kholová, J., 2022. NIR instruments and prediction methods for rapid access to grain protein content in multiple cereals. Sensors 22 (10), 3710.

Chandler, S.L., McSweeney, M.B., 2022. Characterizing the properties of hybrid meat burgers made with pulses and chicken. Int. J. Gastron. Food Sci. 27, 100492. http://dx.doi.org/10.1016/j.ijgfs.2022.100492.

Chen, Z., Xie, Y., Wu, Y., Lin, Y., Tomiya, S., Lin, J., 2024. An interpretable and transferrable vision transformer model for rapid materials spectra classification. Digit. Discov. 3 (2), 369–380.

Cheng, Q., Sun, D.W., 2005. Application of PLSR in correlating physical and chemical properties of pork ham with different cooling methods. Meat Sci. 70 (4), 691–698.

Dal-Pastro, F., Facco, P., Bezzo, F., Zamprogna, E., Barolo, M., 2016. Using PLS and NIR spectra to model the first-breakage step of a grain milling process. In: Kravanja, Z., Bogataj, M. (Eds.), 26th European Symposium on Computer Aided Process Engineering. In: Computer Aided Chemical Engineering, vol. 38, Elsevier, pp. 1171–1176. http://dx.doi.org/10.1016/B978-0-444-63428-3.50200-9.

De Santis, M.A., et al., 2021. Influence of organic and conventional farming on grain yield and protein composition of chickpea genotypes. Agronomy 11 (2), 1–14. http://dx.doi.org/10.3390/agronomy11020191.

Dean, J., 2022. A golden decade of deep learning: Computing systems and applications. Daedalus 151 (2), 58–74.

Delwiche, S.R., Graybosch, R.A., Peterson, C.J., 1998. Predicting protein composition, biochemical properties, and dough-handling properties of hard red winter wheat flour by near-infrared reflectance. Cereal Chem. 75 (4), 412–416. http://dx.doi.org/10.1094/cchem.1998.75.4.412.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16 × 16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations.

Fu, P., Wen, Y., Zhang, Y., Li, L., Feng, Y., Yin, L., Yang, H., 2022. SpectraTr: A novel deep learning model for qualitative analysis of drug spectroscopy based on transformer structure. J. Innov. Opt. Heal. Sci. 15 (03).

Hall, C., Hillen, C., Robinson, J.G., 2017. Composition, nutritional value, and health benefits of pulses. Cereal Chem. 94 (1), 11–31. http://dx.doi.org/10.1016/j.jneb.2020.09.002.

Höskuldsson, A., 1988. PLS regression methods. J. Chemom. 2 (3), 211–228. http://dx.doi.org/10.1002/cem.1180020306.

Huang, H., Yu, H., Xu, H., Ying, Y., 2008. Near infrared spectroscopy for on/in-line monitoring of quality in foods and beverages: A review. J. Food Eng. 87 (3), 303–313. http://dx.doi.org/10.1016/j.jfoodeng.2007.12.022.

Ingle, P.D., et al., 2016. Determination of protein content by NIR spectroscopy in protein powder mix products. J. AOAC Int. 99 (2), 360–363. http://dx.doi.org/10.5740/jaoacint.15-0115.

Jinadasa, M.W.N., Kahawalage, A.C., Halstensen, M., Skeie, N.-O., Jens, K.J., 2021. Deep learning approach for Raman spectroscopy. In: Pathak, C.S., Kumar, S. (Eds.), Recent Developments in Atomic Force Microscopy and Raman Spectroscopy for Materials Characterization. IntechOpen, Rijeka, http://dx.doi.org/10.5772/intechopen.99770.

Jukanti, A.K., et al., 2012. Nutritional quality and health benefits of chickpea (*Cicer arietinum L.*): A review. Br. J. Nutr. 108 (S1), S11–S26. http://dx.doi.org/10.1017/s0007114512000797.

Kipf, T.N., Welling, M., 2017. Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations.

Kirkegaard, J., et al., 2008. Break crop benefits in temperate wheat production. Field Crop. Res. 107 (3), 185–195. http://dx.doi.org/10.1016/j.fcr.2008.02.010.

Li, Y., Wang, X., Yu, H., Du, W., 2023. Pattern-coupled baseline correction method for near-infrared spectroscopy multivariate modeling. IEEE Trans. Instrum. Meas. 72, 1–9.

Liu, H., An, Q., Huan, Z., Bürmen, M., Deng, Q., Marques, T., 2023. ISRToken: Learning similarities tokens for precise infrared spectrum recognition model via transformer. Infrared Phys. Technol. 133, 104700.

Madurapperumage, A., et al., 2021. Chickpea (*Cicer arietinum L.*) as a source of essential fatty acids–a biofortification approach. Front. Plant Sci. 12, 734980. http://dx.doi.org/10.3389/fpls.2021.734980.

Mokni Ghribi, A., et al., 2018. Toward the enhancement of sensory profile of sausage "Merguez" with chickpea protein concentrate. Meat Sci. 143, 74–80. http://dx.doi.org/10.1016/j.meatsci.2018.04.025.

Molavi, B., Dumont, G.A., 2012. Wavelet-based motion artifact removal for functional near-infrared spectroscopy. Physiol. Meas. 33 (2), 259.

Nadimi, M., Paliwal, J., 2024. Recent applications of near-infrared spectroscopy in food quality analysis. Foods 13 (16), 2633.

Nebauer, C., 1998. Evaluation of convolutional neural networks for visual recognition. IEEE Trans. Neural Netw. 9 (4), 685–696.

Neves, M.D.G., Poppi, R.J., Breitkreitz, M.C., 2022. Authentication of plant-based protein powders and classification of adulterants as whey, soy protein, and wheat using FT-NIR in tandem with OC-PLS and PLS-DA models. Food Control 132, 108489. http://dx.doi.org/10.1016/j.foodcont.2021.108489.

Porep, J., Kammerer, D., Carle, R., 2015. On-line application of near infrared (NIR) spectroscopy in food production. Trends Food Sci. Technol. 46 (2), 211–230. http://dx.doi.org/10.1016/j.tifs.2015.10.002.

Pritchard, J.R., et al., 2011. A survey of β-glucan and arabinoxylan content in wheat. J. Sci. Food Agric. 91 (7), 1298–1303. http://dx.doi.org/10.1002/jsfa.4316.

Richetti, J., Diakogianis, F.I., Bender, A., Colaço, A.F., Lawes, R.A., 2023. A methods guideline for deep learning for tabular data in agriculture with a case study to forecast cereal yield. Comput. Electron. Agric. 205, 107642. http://dx.doi.org/10.1016/j.compag.2023.107642.

Rinnan, F.v.d., Engelsen, S.B., 2009. Review of the most common pre-processing techniques for near-infrared spectra. TRAC Trends Anal. Chem. 28 (10), 1201–1222. http://dx.doi.org/10.1016/j.trac.2009.07.007.

Sahoo, G.R., Freed, J.H., Srivastava, M., 2024. Optimal wavelet selection for signal denoising. IEEE Access.

Schuster, C., Huen, J., Scherf, K.A., 2023. Prediction of wheat gluten composition via near-infrared spectroscopy. Curr. Res. Food Sci. 6, 100471.

Shin, H., Sampat, M.P., Koomen, J.M., Markey, M.K., 2010. Wavelet-based adaptive denoising and baseline correction for MALDI TOF MS. Omics: A J. Integr. Biology 14 (3), 283–295.

Silaghi, F.A., et al., 2010. Estimation of rheological properties of gelato by FT-NIR spectroscopy. Food Res. Int. 43 (6), 1624–1628. http://dx.doi.org/10.1016/j.foodres.2010.05.007.

Talaei Khoei, T., Ould Slimane, H., Kaabouch, N., 2023. Deep learning: systematic review, models, challenges, and research directions. Neural Comput. Appl. 35 (31), 23103–23124.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Vol. 30. Curran Associates, Inc.

Wang, Z., et al., 2017. Isolation, identification and characterization of a new type of lectin with alpha-amylase inhibitory activity in chickpea. Protein Pept. Lett. 24 (11), 1008–1020. http://dx.doi.org/10.2174/0929866524666170711120501.

Wang, Z., et al., 2024. Infrared guided smart food formulation: an innovative spectral reconstruction strategy to develop anticipated and constant apple puree products. Food Innov. Adv. 3 (1), 20–30. http://dx.doi.org/10.48130/fia-0024-0003.

Willett, W., et al., 2019. Food in the Anthropocene: the EAT–Lancet commission on healthy diets from sustainable food systems. Lancet 393 (10170), 447–492. http://dx.doi.org/10.1016/s0140-6736(19)31101-8.

Wood, J., Grusak, M., 2007. Nutritional value of chickpea. Chickpea Breed. Manag. 101–142. http://dx.doi.org/10.1079/9781845932138.005.

Wu, Y., Zhu, X., Huang, Q., Zhang, Y., Evans, J., He, S., 2023. Predicting the quality of tangerines using the GCNN-LSTM-AT network based on vis–NIR spectroscopy. Appl. Sci. 13 (14), 8221.

Yang, W., Liu, L., Deng, W., Huang, W., Ye, J., Hu, S., 2023. Deep retrieval architecture of temperature and humidity profiles from ground-based infrared hyperspectral spectrometer. Remote. Sens. 15 (9), 2320.

Ye, W., Ma, L., 2024. Denoising reconstruction of Raman spectra based on wavelet transform and weighted Wiener estimation. In: 2024 AI Photonics Technology Symposium, Vol. 13227. SPIE, pp. 50–56.

Ye, N., Zhong, S., Fang, Z., Gao, H., Du, Z., Chen, H., Yuan, L., Pan, T., 2022. Performance improvement of NIR spectral pattern recognition from three compensation models' voting and multi-modal fusion. Molecules 27 (14), 4485.

Zhang, Z.-M., Chen, S., Liang, Y.Z., 2010. Baseline correction using adaptive iteratively reweighted penalized least squares. Analyst 135 (5), 1138–1146.

Zhang, W., Kasun, L.C., Wang, Q.J., Zheng, Y., Lin, Z., 2022a. A review of machine learning for near-infrared spectroscopy. Sensors 22 (24), 9764.

Zhang, F., Tang, X., Tong, A., Wang, B., Wang, J., Lv, Y., Tang, C., Wang, J., 2020a. Baseline correction for infrared spectra using adaptive smoothness parameter penalized least squares method. Spectrosc. Lett. 53 (3), 222–233.

Zhang, X., Xu, J., Yang, J., Chen, L., Zhou, H., Liu, X., Li, H., Lin, T., Ying, Y., 2020b. Understanding the learning mechanism of convolutional neural networks in spectral analysis. Anal. Chim. Acta 1119, 41–51.

Zhang, S., et al., 2022b. Application of near-infrared spectroscopy for the nondestructive analysis of wheat flour: A review. Curr. Res. Food Sci. 5, 1305–1312. http://dx.doi.org/10.1016/j.crfs.2022.08.006.