Journal of Chemometrics

WILEY

Journal of CHEMOMETRICS

REVIEW ARTICLE OPEN ACCESS

How Are Chemometric Models Validated? A Systematic Review of Linear Regression Models for NIRS Data in Food Analysis

Jokin Ezenarro[1] | Daniel Schorn-García[2]

[1]Universitat Rovira i Virgili, ChemoSens Group, Department of Analytical Chemistry and Organic Chemistry, Campus Sescelades, Tarragona, Spain | [2]Department of Viticulture and Oenology, South African Grape and Wine Research Institute, Stellenbosch University, Stellenbosch, South Africa

Correspondence: Jokin Ezenarro (jokin.ezenarro@urv.cat) | Daniel Schorn-García (dschorng@sun.ac.za)

Received: 30 January 2025 | **Revised:** 24 March 2025 | **Accepted:** 25 April 2025

Funding: This study was supported by the Chemometrics and Sensorics for Analytical Solutions (CHEMOSENS, ref.2021 SGR 00705, Departament de Recerca i Universitats, Generalitat de Catalunya).

Keywords: analytical methods | figures of merit | prediction | quality control | spectroscopy | validation

ABSTRACT

Chemometric models play a critical role in the spectroscopic analysis of food, particularly with near-infrared spectroscopy (NIRS), enabling the accurate prediction and monitoring of physicochemical properties. Although chemometric methods have proven to be useful tools in NIRS analysis, their reliability depends on rigorous validation to ensure the rigour of their predictions and their applicability. This systematic review examines validation strategies applied to regression models in NIRS-based food analysis, emphasising the use of cross-validation, external validation and figures of merit (FoM) as key evaluation tools. This comprehensive literature search identified trends in validation methodologies, highlighting frequent reliance on partial least squares (PLS) regression and common flaws in validation methodologies and their reporting. While external validation is considered the best approach, many studies lack it and employ cross-validation methods solely, which may lead to overoptimistic model performance estimates. Furthermore, inconsistencies in the selection and definition of FoM hinder direct comparison across studies. This review underscores the need for increased methodological transparency and rigour in the validation of chemometric models to enhance their reliability.

1 | Introduction

Chemometric models have become indispensable tools in modern analytical chemistry, enabling the extraction and transformation of complex datasets into meaningful information generated by advanced spectroscopic techniques [1]. In the field of food analysis, these models are widely used to predict key physicochemical properties, monitor quality and ensure compliance with safety regulations [2]. Among the various chemometric approaches, regression models are particularly prominent for their ability to establish quantitative relationships between spectral data and reference measurements [3]. These models have found extensive application in conjunction with near-infrared (NIR) spectroscopy, a technique commonly used in food science for its rapid, nondestructive nature and its ability to simultaneously analyse multiple components. The integration of near-infrared spectroscopy (NIRS) with chemometric models has led to significant advancements in quality control and process monitoring. This development has firmly established NIRS as a fundamental component of food analysis [4, 5].

NIRS relies on the absorption of light in the NIR region, capturing overtones and combinations of fundamental vibrational

modes of molecular bonds. This spectral data is inherently multivariate, containing a wealth of information about the molecular composition of food samples. However, the complexity and high dimensionality of NIR spectra require the use of chemometric techniques to extract relevant information. Regression models are particularly valuable in this context, as they enable the quantitative determination of important food attributes, such as moisture, protein, fat and contaminant concentrations [6].

The rigour of chemometric regression models, however, is subject to rigorous validation. Validation is the process of assessing the ability of a model to perform robustly on new, unseen data, thereby ensuring its generalisability and predictive ability beyond the calibration dataset [7]. This is highly important in food analysis, where model predictions directly influence decisions related to production quality, regulatory compliance and consumer safety. Insufficient validation of models can lead to overfitting the calibration data, resulting in performance estimates that are excessively optimistic when applied to real-world conditions. Such failures can have significant implications, including economic losses, product recalls and compromised food safety [8].

In the context of NIRS and regression modelling, validation must account for the challenges posed by spectral data. NIR spectra are often characterised by high collinearity among variables, variability due to sample heterogeneity and noise introduced by instrumentation or environmental conditions [6]. Consequently, robust validation strategies are essential to ensure the reliability and applicability of chemometric models. External validation, in which an independent dataset is used to assess model performance and generalisability, is considered the gold standard. However, due to practical constraints, such as limited data availability, internal validation techniques like cross-validation (CV), which partition the calibration set into training and testing subsets, are often necessary for model hyperparameter optimisation, like the data preprocessing method or model dimensionality [9]. Moreover, the assessment of these analytical methods is based on the use of quantitative figures of merit (FoM), key performance indicators that provide insights into the predictive accuracy, robustness and practical applicability of the model. These metrics allow researchers to systematically evaluate different validation approaches, refine their models and enhance their reliability for real-world applications in food analysis [10].

Overall, this systematic review is focused on studying the validation of chemometric regression models applied to NIRS in food analysis that can be found in the literature. The growing reliance on these models underscores the need for rigorous and scientifically sound validation practices. By examining current methodologies, highlighting best practices and underscoring less effective ones, this review is aimed at contributing to the advancement of rigorous chemometric validation methodologies. Robust validation is an essential methodological step in ensuring that analytical methods based on spectroscopy and chemometrics achieve their full potential in improving the quality, safety and efficiency of food production and analysis.

2 | Theoretical Background

2.1 | Multivariate Regression Models

Multivariate regression models are a class of statistical techniques used to analyse and predict relationships between multiple independent variables ($\mathbf{X}$) and one or more dependent variables ($\mathbf{y}$) [3]. The most common methods to do so are the linear models; these models extend simple linear regression by considering multiple predictors, enabling the simultaneous evaluation of their effects on the response variable. In their basic form, linear models assume that the relationship between the predictors and the response is linear and can be expressed mathematically with Equation (1):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients and $\boldsymbol{\varepsilon}$ accounts for residuals. In the extensions of the univariate least squares linear regression such as multivariate or multiple linear regression (MLR), principal component regression (PCR) and partial least squares (PLS), the coefficients are estimated by minimising the residual sum of squares. This provides the best linear unbiased estimates, meaning that under the assumptions of normality, homoscedasticity and independence of errors, the estimated coefficients have the lowest variance among all linear estimators and, on average, equal the true parameter values without systematic over- or underestimation. These models are commonly used across various fields due to their flexibility in handling complex data and their interpretability in identifying the contribution of each predictor [5, 11].

In addition to linear models, more complex but flexible techniques have been developed to address limitations such as nonlinearity. These approaches can be broadly categorised based on their methodology. Kernel-based methods, such as support vector machines (SVMs) [12] and kernel partial least squares (KPLS) [13], map data into higher dimensional spaces using kernel functions to potentially capture nonlinear relationships. Other nonlinear approximation methods, including artificial neural networks (ANNs) [14], model complex dependencies through the use of one or more layers of nodes, also called neurons. In particular, extreme learning machines (ELMs) use one randomly weighted layer of neurons and calibrate the weights of the output nodes [15]. In somewhat a similar way, nonparametric ensemble learning techniques, such as random forests (RFs) and gradient boosting, capture nonlinearities and variable interactions by aggregating predictions from multiple decision trees [16]. In contrast, compositional data analysis (CoDa) is tailored for datasets with variables as proportions, common in food and environmental sciences, using transformations to handle their constrained nature [17]. These methods form a comprehensive chemometric toolkit, enabling the modelling of more complex systems beyond the scope of purely linear approaches.

Unlike univariate approaches, multivariate regression models can capture the combined effects of multiple predictors, making them particularly useful for systems with complex interactions such as in spectroscopy. However, the effectiveness of these models depends on their ability to account for multicollinearity,

reduce dimensionality where necessary and ensure predictive robustness through rigorous validation [18].

2.2 | External Validation

External validation is the process of evaluating a predictive model using an independent dataset that has not been involved in model development or calibration. External validation sets are a fundamental component of rigorous model evaluation in chemometrics, providing an independent and unbiased measure of the generalisation ability and performance of a model. By utilising a dataset entirely separate from the calibration process, external validation ensures that the assessment of model performance reflects its capacity to predict unseen data under realistic conditions. In domains such as spectroscopy and food analysis, where models often encounter complex and variable sample matrices, external validation serves as a gold standard for evaluating predictive performance, identifying overfitting and overall validating the reliability of chemometric models [9]. The selection of the external validation strategy could be performed by different methodologies, discussed in the following subsections, taking into account that they differ in how the validation set is selected, as illustrated in Figure 1.

2.2.1 | Independent Set

Using an independent dataset for validation represents the most rigorous approach to validating a model. Unlike data splitting or internal validation methods, an independent validation dataset is created separately from the data used for model training and optimisation, with truly unseen samples covering the intended domain for the proposed model. This ensures that the evaluation process reflects the ability of the model to adapt to the domain of work, free from any biases or overfitting introduced during calibration, including the data preprocessing and selection of hyperparameters. This approach provides a more objective measure of its predictive capability and practical utility, making it a recommended practice in chemometric model development [9].

2.2.2 | Random Splitting

When the availability of an independent validation set is limited or infeasible, splitting the calibration set into training and validation subsets becomes a practical alternative for assessing model performance. In this approach, the calibration data is divided into two parts: one subset is used to train the model, while the other is reserved for validating the model, in a proportion selected by the analyst. While not as rigorous as using a fully independent validation set, splitting the calibration set provides a mechanism to approximate external validation, especially in scenarios with limited data. However, it must be ensured that the used validation set is representative of the intended domain of work for the model in terms of variability [7, 9].

Random splitting is a straightforward method for partitioning a dataset into calibration and validation subsets by assigning samples randomly to each group. This approach is widely used due to its simplicity and computational efficiency, as it requires no

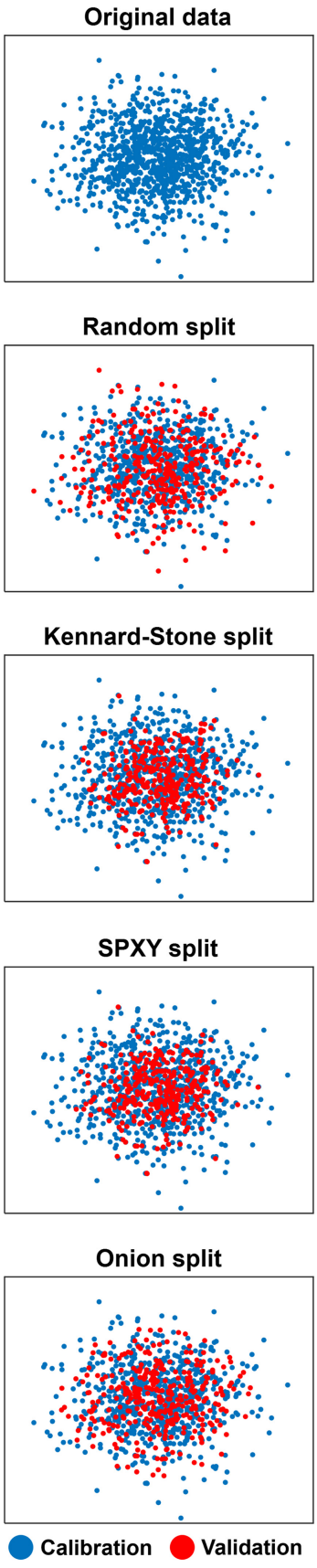

FIGURE 1 | Illustrative example of how a dataset would be split into calibration and validation subsets using different algorithms. Axes represent randomly generated Variable 1 versus Variable 2, which are normally distributed. For SPXY, a linear combination of variables in **X** and normally distributed random noise was used as the reference value (**y**).

complex algorithms or additional assumptions. However, it has notable limitations, particularly when applied to small or imbalanced datasets. Random splitting does not ensure that the subsets are representative of the overall data distribution, which may result in calibration or validation sets that lack critical variability or fail to cover the entire feature space. Consequently, this can lead to biased or unreliable estimates of model performance, for instance when replicates or somehow related samples are present in the dataset [9, 19]. While random splitting may be suitable for large datasets with balanced, independent and diverse samples, care must be taken to ensure that the subsets accurately reflect the variability present in the full dataset, often needing additional strategies such as stratification to address potential biases [20, 21].

2.2.3 | Kennard–Stone Algorithm

The Kennard–Stone algorithm is a systematic method for selecting representative subsets from a dataset to ensure uniform coverage of the input variable space. Initially, it identifies two samples that are the farthest apart in the variable space to maximise diversity. Subsequently, it iteratively selects samples that are farthest from those already chosen, ensuring each new sample adds maximum information about unexplored regions, which maximises the variability of the dataset [22]. This splitting can be done using Euclidean distances or using Mahalanobis distances, this is, variance-scaled distances, for a better representation of the space [23].

2.2.4 | SPXY Algorithm

The SPXY (sample set partitioning based on joint **X**–**y** distances) algorithm is an enhancement of the Kennard–Stone method. Unlike Kennard–Stone, SPXY incorporates variability in both the predictor (**X**) and response (**y**) spaces. By normalising and combining the distances in **X** and **y**, SPXY ensures that calibration and validation subsets are not only representative of the predictor space but also account for the diversity in the response values. This joint consideration improves the robustness and predictive ability of models by preventing extrapolation and better aligning the sample selection with the variability of the predicted feature. As a result, SPXY is valuable in applications where the feature to be predicted is not related to the main sources of variance in the analytical signal, which often occurs when food is analysed by NIRS [24].

2.2.5 | Onion Algorithm

The Onion algorithm is a method for splitting a dataset into calibration and validation sets based on their distance from the centre of the dataset in the feature space. The algorithm arranges samples in concentric layers of an 'onion', with each layer being included either in the calibration or validation set, alternately. Typically, the outermost samples, which represent the extremes of the data distribution, are selected for the calibration set, ensuring that the calibration set includes the most diverse and extreme samples, thus capturing the variability in the feature space. Onion splitting minimises the risk of extrapolation during validation by ensuring that the validation set lies within the space covered by the calibration set. This method is especially relevant in multivariate analysis, where uniform coverage of the data space is important for building robust and generalisable models [25].

2.2.6 | Structured or Custom Splitting

When the dataset is ruled by sample groups based on hierarchical structures, such as replicates, batch effects or sampling sites, structured splitting can ensure that the subsets used for calibration and validation maintain the inherent relationships within the data. Unlike random splitting, which can arbitrarily divide related samples, structured splitting ensures that entire groups of related samples are either included or excluded together in each subset. This approach preserves the integrity of the variability of the dataset and prevents information leakage, which could lead to overoptimistic model performance estimates [9].

Even when such structures are not present, the analyst can decide to purposely include or exclude certain samples from the validation set using other criteria, such as preserving the range or distribution of reference values and using particularly interesting samples or measurements. This creates a custom validation subset with the desired properties.

2.2.7 | Other Splitting Methods

More dataset splitting methods have also been proposed in the literature, for instance, variations and extensions of the SPXY algorithm, such as the weighted SPXY [26] or the kernel-based SPXY [27]. Also, other methods based on quantifying the relevance of each sample for building the model included one based on D-optimal [28] design of experiments, another one based on the combined analytical signal [29] or another based on multivariate leverage (LVG) [30]. Other methods propose data splitting based on sample dissimilarity, such as the OptiSim method [31], and, for instance, the SPlit algorithm [32], which tries to find the worst-case scenario for validation based on nearest neighbours. Each of them offers its advantages and tries to offer a new focus on the external validation of chemometric models.

2.3 | CV or Resampling Strategies

Validation strategies based on partitioning the calibration dataset are often used for optimising the hyperparameters of the model and estimating its robustness, even though a separate dataset could also be used. CV strategies involve systematically partitioning the calibration dataset into complementary subsets: all but one subset are used to train the model, and the remaining subset is used to test it. This process is repeated until all samples have been predicted to obtain estimates of predictive performance. Resampling methods, on the other hand, generate multiple new datasets by randomly sampling (with or without replacement) from the original data. These new datasets are used to repeatedly refit the model and assess variability in model parameters and performance metrics. While resampling methods do not always directly provide predictive performance estimates in the same way CV does, they are valuable for estimating uncertainty and stability of model parameters, sometimes

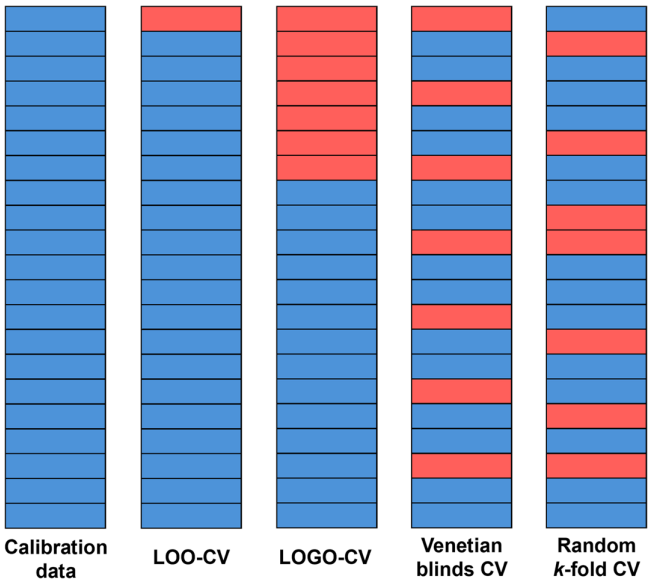

FIGURE 2 | Illustrative example of how a dataset would be split into subsets for cross-validation (CV) using different algorithms. Each rectangle represents a sample, blues samples represent the calibration subset in the first iteration and the red ones represent the validation subset in that iteration. The process is repeated until all samples have been used for validation once.

establishing confidence intervals. The choice of which method to use depends on dataset size, complexity and the presence of grouped or related samples [9]. Each methodology is presented in the following subsections and illustrated in Figure 2.

2.3.1 | Leave-One-Out Cross-Validation (LOO-CV)

In LOO-CV or full CV, the dataset is split such that each sample is used once as a validation set while the remaining samples are used to train the model. This process is repeated for every sample in the dataset, resulting in as many iterations as there are samples, which makes this algorithm computationally intensive. This CV method is useful for evaluating the effect of individual samples; however, it tends to be overly optimistic for evaluating model performance, particularly in datasets where samples with high correlation among them are present, such as those found in food NIRS [9].

2.3.2 | Leave-One-Group-Out Cross-Validation (LOGO-CV)

LOGO-CV, leave-one-patient-out cross-validation (LOPO-CV), custom subsets or contiguous blocks CV is a variation of CV designed for datasets with grouped or hierarchical structures, where samples are organised into distinct groups based on shared characteristics, such as replicates, batches or sampling sites. In LOGO-CV, entire groups are left out as the validation set during each iteration, while the remaining groups are used to train the model. This approach ensures that the validation set is independent of the training data and preserves the inherent relationships within groups, making it particularly suitable for evaluating intersample variability in datasets with nested or hierarchical variability [9].

2.3.3 | Venetian Blinds CV

Venetian blinds CV is a method of partitioning data for model validation by splitting the dataset into evenly spaced, non-overlapping subsets, resembling the alternating pattern of a venetian blind. Each subset or fold is used as the validation set once, while the remaining subsets are used for training the model. This approach ensures that samples from across the entire dataset are included in both training and validation, which is particularly useful for datasets with structured or time-ordered variability. This method is computationally efficient and is well-suited for datasets with evenly distributed data points, but it may not perform as effectively in datasets with correlations or groupings [9].

2.3.4 | Random Subsets or Random *k*-Fold CV

Random subsets or random *k*-fold CV is a method where the dataset is randomly divided into *k* equal-sized subsets or folds, and the model is trained and validated *k* times. In each iteration, one of the subsets is used as the validation set, while the remaining *k*-1 subsets form the training set. The process is repeated until each subset has served as the validation set once. When the predictions for all individual samples are available, the performance metrics are calculated, providing a robust estimate of model performance. This approach is widely used due to its simplicity and flexibility, especially for datasets with no inherent structure or groupings [9].

2.3.5 | Resampling Methods

Resampling strategies differ from other CV techniques by focusing on repeated random sampling of the dataset to calculate validation metrics in each iteration and combine them, providing metrics that reflect the robustness and stability of the model [9].

Bootstrap validation is a resampling method that creates multiple subsets of data by randomly sampling with replacement from the original dataset. Each bootstrap sample or iteration typically has the same size as the original dataset but may include repeated samples while leaving some samples out. A model is trained on each bootstrap sample and tested on the samples not included (the 'out-of-bag' samples), providing an estimate of model performance [33, 34].

Monte Carlo validation involves generating multiple random splits of the dataset into calibration and validation subsets. In each iteration, a model is trained on the calibration subset and tested on the validation subset, and the process is repeated many times with different random splits. This approach provides a comprehensive assessment of model performance across a range of potential data partitions, capturing variability in performance metrics due to different dataset compositions [35].

The jackknife method is a systematic resampling technique where one sample is removed from the dataset at a time, and the model is trained on the remaining data. The process is repeated until each sample has been excluded from the calibration process

once. By aggregating the results across all iterations, jackknifing may provide estimates of variability, uncertainty or stability in model parameters such as regression coefficients or dimensionality, particularly useful for detecting influential samples or outliers that may disproportionately affect the model [34].

2.3.6 | Permutation-Based Methods

Permutation-based methods can be used to assess the statistical significance of chemometric models by comparing their performance to that of models built on randomly permuted response variables. Although these methods do not directly provide a performance evaluation in terms of predictive ability, they can help in the model hyperparameter optimisation process by distinguishing between parameter settings that lead to models capturing true structure in the data and those that simply fit random noise. By repeatedly permuting the response variable and recalculating model metrics (such as explained variance), it is possible to build a null distribution against which the observed can be compared. A hyperparameter set that consistently yields a better model than the permuted counterparts can thus be selected with greater confidence. These methods become especially useful when dealing with small datasets or models with high complexity and risk of overfitting [20, 36].

2.3.7 | Other CV Methods

Procrustes CV is a novel method that bridges random *k*-fold CV and independent test set validation in chemometric modelling. It generates a 'pseudovalidation set' by incorporating sampling uncertainties from CV into the calibration data, simulating an independent test set. Unlike conventional CV, Procrustes CV enables the assessment of global model parameters, such as explained variance and residuals, and is particularly effective in detecting overfitting. While not a substitute for independent validation, this CV method offers a robust alternative for model optimisation and exploration [37].

Repeated double cross-validation (RDCV) is a technique that combines internal CV for model selection with external validation for unbiased performance estimation. The process consists of two nested CV loops: the outer loop, which partitions the dataset into training and validation subsets to evaluate final model performance, and the inner loop, which is used for model selection and hyperparameter tuning. This structure ensures that hyperparameter optimisation does not influence the final performance estimate, reducing overfitting risks, particularly in small datasets [38].

Probabilistic CV [39] and Bayesian CV [40] are other variations of the CV method, designed to address the limitations of deterministic model evaluations by introducing probabilistic elements into the validation process. Unlike standard CV, which evaluates model performance through fixed data partitions, these methods incorporate uncertainty into the validation procedure, modelling the prediction error as a probabilistic function. This approach splits the dataset into subsets and uses a probabilistic framework, such as Gaussian distributions, to estimate the model performance.

2.4 | Figures of Merit

FoMs are essential tools for evaluating the performance of analytical methods. These metrics provide quantitative assessment of model performance and robustness, including the quality of predictions, the reliability of measurements and the ability to generalise across different datasets [10]. Error metrics such as root mean square error (RMSE), mean absolute error (MAE), standard error (SE), relative error (RE) and mean square error (MSE) are widely used to quantify prediction errors. RMSE penalises larger errors more heavily, while MAE provides a simpler measure of average error, less sensitive to outliers. SE captures variability around the mean, and MSE evaluates the squared differences between observed and predicted values, emphasising larger deviations [41].

Correlation and fit metrics provide insights into the relationship between observed and predicted values. The correlation coefficient (R) and the coefficient of determination (R^2) measure the strength and direction of the linear relationship; the latter is also equivalent to the variance explained by the model (EV) [42]. The concordance correlation coefficient (CCC) evaluates agreement by combining measures of precision and accuracy [43].

Other regression characteristics between the predicted versus observed data, including slope, offset and bias, provide additional information about systematic deviations in predictions. A slope close to unity and an offset near zero indicate an unbiased model. In addition, statistical tests like the permutation test assess the significance of the observed relationships, ensuring model results are meaningful and not due to chance [8].

Performance ratios evaluate the predictive capability of models in a standardised way, so the values such as the ratio of performance to deviation (RPD) and the range error ratio (RER) can be compared between models. RPD standardises the RMSE or other error metrics, comparing them with the standard deviation of observations, with higher values indicating better performance, while RER considers the range of the data relative to RMSE [42]. Other metrics, like the ratio of performance to interquartile distance (RPIQ), enhance robustness by using the interquartile range instead of the standard deviation, making them less sensitive to outliers [44].

In addition to these, other properties of the model such as the limit of detection (LOD) and the limit of quantification (LOQ) are critical in analytical applications, determining the smallest amount of analyte that can be reliably detected or quantified, respectively [45]. Other less known metrics, like capacity detection (CCβ), identify the smallest quantity of analyte detectable with minimal false negatives [17, 46]. Altogether, these FoMs provide a comprehensive framework for evaluating and comparing chemometric models, ensuring their reliability and applicability in diverse scenarios.

However, it should be noted that more than one definition exists for some of the FoM, such as the RMSE, LOD or bias. Therefore, instead of offering mathematical definitions here, we refer the reader to other publications of reference in the topic [10, 45, 47–49].

3 | Methodology

3.1 | Systematic Search

The research method for the systematic review followed the PRISMA guidelines (Figure 3), which ensure transparency and reliability in systematic reviews by adhering to a structured checklist [50]. The eligibility criteria focused on studies addressing the use of NIRS for food analysis, specifically targeting validation strategies in predictive modelling.

The primary information source for this review was SCOPUS, where the search was conducted on 4 December 2024. The search strategy employed the following query:

(TITLE-ABS-KEY ("near-infrared" OR "near-infrared spectroscopy" OR "NIR" OR "NIR analysis") AND TITLE-ABS-KEY ("food" OR "food quality" OR "food analysis" OR "food safety" OR "food composition" OR "food authentication" OR "food science") AND TITLE-ABS-KEY ("PLS" OR "partial least squares" OR "PLS regression" OR "multivariate analysis" OR "chemometrics" OR "regression" OR "predictive modeling" OR "machine learning")) AND PUBYEAR > 2014 AND PUBYEAR < 2026.

This strategy was designed to include various terminologies and synonyms relevant to the field while excluding studies beyond the scope of the review. The choice to use broad search terms such as 'food' was deliberate to maintain inclusivity across various food matrices and to provide a general overview of validation strategies in the field. The search was limited to publications from 2015 to 2025, chosen to represent the most recent validation strategies in NIR applications for food analysis. This time frame was defined by the authors, as there is no established consensus on the appropriate frequency for systematic reviews [51].

A total of 1848 records were initially retrieved. These records were screened for relevance and quality, leading to the exclusion

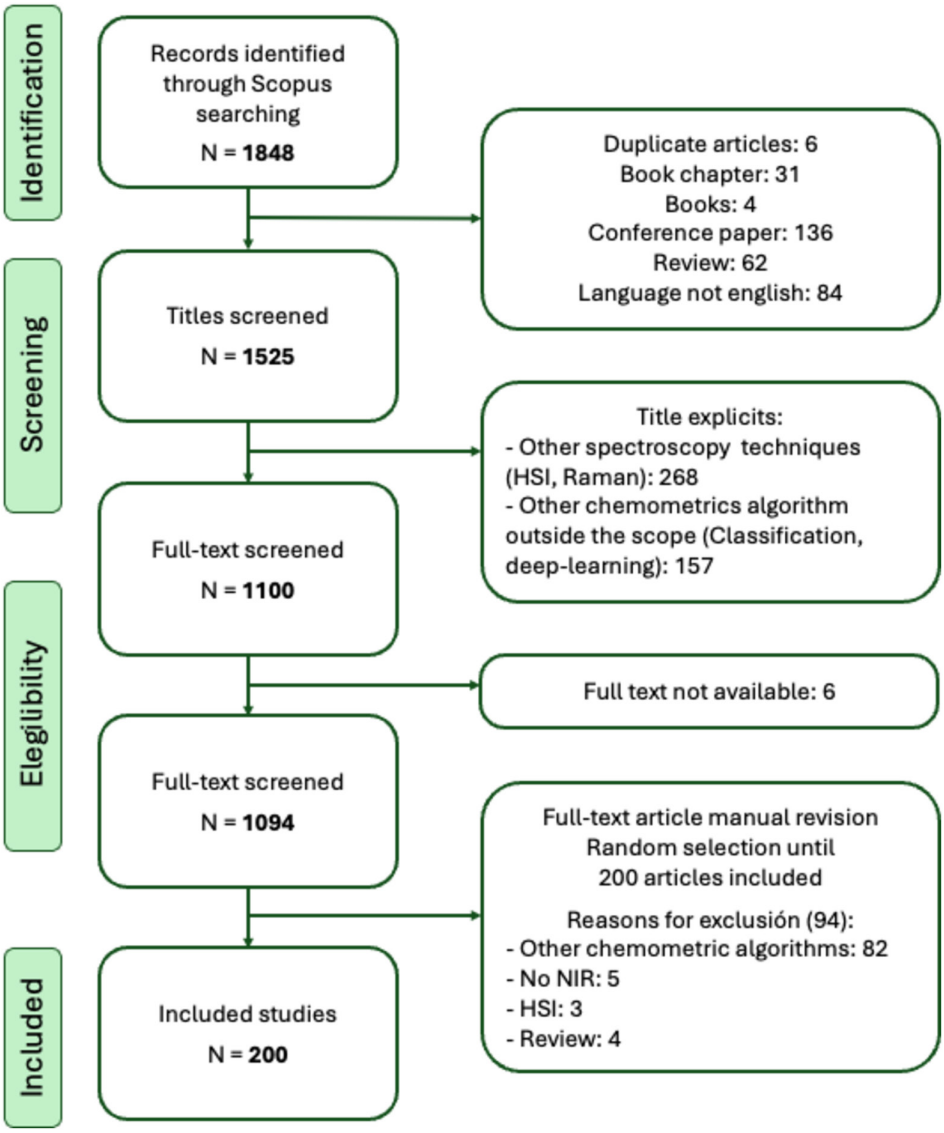

FIGURE 3 | PRISMA flow diagram for the literature search conducted on the validation strategies used in food properties prediction chemometric models.

of studies that did not focus on the application of NIR in food analysis. Articles outside the scope, such as those analysing nonprediction or unrelated methodologies, were removed. After screening, 1094 records were included in the review. Due to the large number of articles that met the inclusion criteria after screening, a random selection of 200 articles was performed to ensure manageability.

Data was systematically extracted into a Microsoft Excel file, including information on the journal, year, CV strategy, external validation strategy, FoM used, if the equation of those FoM were display, the algorithm used and if a graph showing the models was included.

3.2 | Validation Rigorousness Mark

To evaluate the rigour of validation strategies used in published studies, a validation rigorousness mark framework was developed. This framework systematically scores studies based on the use and description of CV and external validation methods. The validation rigorousness mark includes nine categories, ranked from 1 (*least rigorous*) to 9 (*most rigorous*), with criteria derived from the combination of these four elements: whether the study applies any form of CV to validate its findings, whether the study provides detailed information about the CV methodology (e.g., the number of folds, randomisation strategy or data splitting), whether the study employs external validation by testing its model or results on an independent dataset and whether the methodology for external validation is described in sufficient detail (e.g., dataset origin, preprocessing steps and evaluation metrics). The highest score (9) is awarded to studies that use both CV and external validation and provide detailed descriptions of their methodologies. Each article included in the review was systematically evaluated, and information regarding its validation approaches was extracted. Each study was marked as 'YES' or 'NO' for CV or external validation usage, and if any or both were used, the description of the strategy was evaluated. Then, a score was assigned based on the results of the scoring matrix shown in Table 1. Certain combinations of criteria were logically excluded. For instance, studies that do not employ CV cannot define a CV strategy, and such scenarios were omitted from the factorial scoring design.

All data were independently extracted and evaluated by two reviewers to ensure objectivity. Any discrepancies in scoring were resolved through consensus discussions. The distribution of validation rigorousness mark across studies was analysed to identify common validation practices and trends in methodological rigour. Descriptive statistics, including frequency distributions and mean validation rigorousness scores, were calculated. Correlations between validation rigorousness mark scores and study characteristics were visually assessed.

4 | Results and Discussion

The selected reviewed articles were analysed to understand the year distribution and its equivalence to the whole article set. The temporal distribution of the reviewed articles from 2015 to 2025, which has a similar distribution to the whole (data not shown), is shown in Figure 4. The number of

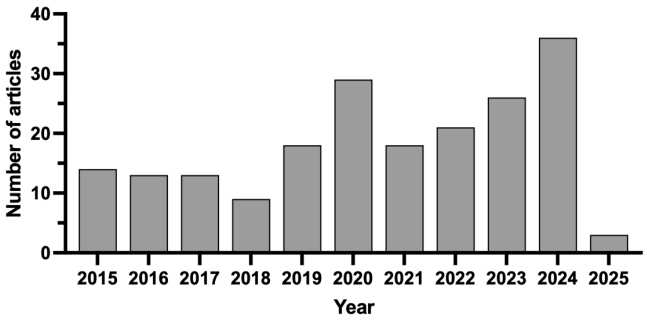

FIGURE 4 | Distribution of reviewed articles by year included in the review.

TABLE 1 | Explanation of the validation rigorousness mark based on the inclusion of cross-validation and external validation strategies, as well as the level of detail provided in the article regarding their implementation. Nine categories were considered in a full factorial design, excluding combinations where cross-validation or external validation were not used, as their strategies could not be defined.

Validation rigorousness mark	Use of CV	Is the CV strategy declared?	Use of external validation	Is the external validation strategy declared?
1	NO	NO	NO	NO
2	YES	NO	NO	NO
3	YES	YES	NO	NO
4	NO	NO	YES	NO
5	YES	NO	YES	NO
6	YES	YES	YES	NO
7	NO	NO	YES	YES
8	YES	NO	YES	YES
9	YES	YES	YES	YES

publications remained relatively stable from 2015 to 2019. A notable increase is observed in 2020, reaching its first peak, possibly influenced by the COVID-19 pandemic lockdown [52], followed by a fluctuating yet consistently higher year-to-year number of publications from 2021 to 2024. The highest number of reviewed articles was published in 2024, indicating a growing research interest in the topic. In contrast, the small number of included articles in 2025 likely reflects the incomplete data collection for that year at the time of this review. This trend suggests an increasing focus on the application of NIR-based regression models for food property prediction, particularly in recent years.

The reviewed journals span a diverse range of *Journal Citation Reports* (*JCR*) categories, reflecting the interdisciplinary nature of research on NIR-based regression models and food property prediction. The highest number of journals falls under the 'Food Science & Technology' category (29 journals), underscoring the strong relevance of this field to food analysis. Other prominent categories include 'Chemistry, Analytical' (11 journals) and 'Chemistry, Applied' (nine journals), highlighting the importance of chemical methodologies in food research. Notably, categories related to spectroscopy, instrumentation and imaging, such as 'Spectroscopy' (five journals), 'Instruments & Instrumentation' (four journals) and 'Imaging Science & Photographic Technology' (two journals), emphasise the key role of advanced analytical techniques in the field. Other categories, including 'Engineering', 'Remote Sensing', 'Artificial Intelligence' and 'Statistics', demonstrate the integration of computational, technological and mathematical approaches in food-related research. Overall, this distribution highlights the multidisciplinary nature of NIR-based food analysis, encompassing food science, chemistry, engineering, agriculture and data-driven methodologies.

Given the wide range of disciplines involved in NIR prediction model validation, also a variety of prediction algorithms have been used to develop the models. The distribution of regression algorithms employed in the reviewed articles is illustrated in the pie chart in Figure 5. For the 200 reviewed articles, PLS is the most commonly used regression algorithm, occupying the largest portion of the chart. This predominance underscores PLS as the preferred method for NIR-based prediction of food properties, likely due to its robustness in handling collinear spectral data and its well-established performance in NIRS [53]. Other regression algorithms appear much less frequently. Among these, SVM and MLR have a relatively larger share compared to RF, PCR, ELM and CoDa, which account for only minor portions of the total algorithms used.

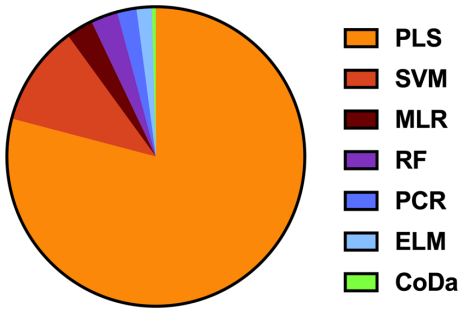

FIGURE 5 | Representation of the proportion of regression algorithms used in the reviewed articles. $N = 200$ articles.

This trend highlights the reliance on PLS within the field, while the presence of alternative algorithms, though in smaller proportions, suggests an ongoing exploration of machine learning and statistical methods to enhance predictive performance and model generalisation. Notably, 19% of the reviewed studies employed multiple regression algorithms, reflecting a growing interest in hybrid or comparative modelling approaches.

4.1 | External Validation

External validation is a critical step in chemometric model validation, ensuring that predictive models generalise beyond the calibration dataset and perform reliably on unseen data. However, 15% of the reviewed articles do not perform any external validation at all on their models. Not having a proper external validation and relying on CV strategies only may lead to overly optimistic performance estimates that do not hold in real-world applications.

Regarding the 85% that do perform some sort of external validation and six articles used more than one method, Figure 6 presents the distribution of the methods employed. A considerable proportion of the authors that used external validation (29%) did not disclose their external validation approach, highlighting a gap in methodological transparency. Among those that specified a validation strategy, random splitting and custom validation methods were the most frequently used, followed by more structured approaches such as the Kennard–Stone and the SPXY algorithms. It is also noteworthy the low use of totally independent datasets for validation, even if this is, in general, the most recommended approach, as it is the most realistic approach to model performance estimation [7, 9].

The use of custom validation strategies suggests an effort to tailor validation approaches to specific dataset characteristics. As shown in the breakdown of custom validation methods (Figure 6), a significant portion of these approaches relies on criteria regarding other properties of the samples not considered in the **X** block, such as the range or standard deviation of the reference values (**y**) or the sample classes. While these methods enhance representativeness, their implementation varies widely, and insufficient standardisation can lead to inconsistencies in model evaluation across studies. For instance, while randomly splitting a dataset into calibration and validation subsets is the most commonly used approach, it is essentially equivalent to performing random *k*-fold CV but with only a single partition. This means that although it offers the characteristics of a random sampling, it relies on just one division of the data, making it highly sensitive to chance. As a result, it may produce unreliable performance metrics, especially when working with small or unevenly distributed datasets. In such cases, it would be more appropriate to use alternative data-splitting strategies that account for dataset structure and variability, alongside random *k*-fold CV to obtain more robust and generalizable model evaluations.

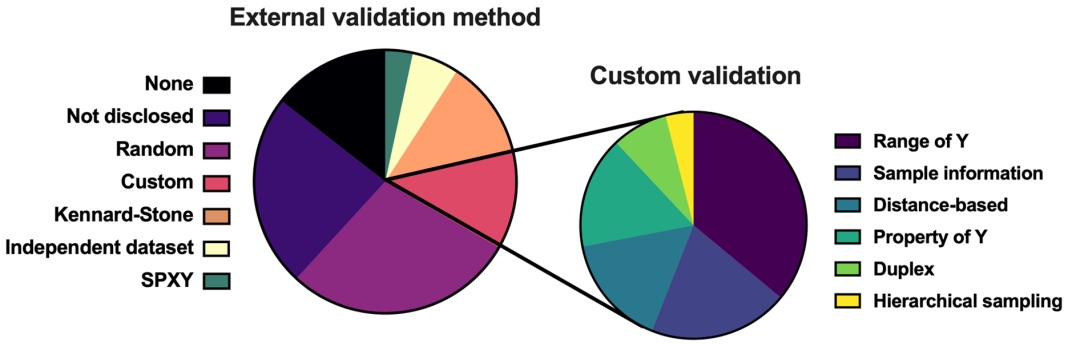

FIGURE 6 | Representation of the proportion of external validation methods used in the reviewed articles. $N = 170$ articles used external validation. $N = 25$ articles used a custom external validation strategy.

To improve the reliability of external validation, several well-designed strategies can be employed to ensure that the test set is representative of real-world conditions. One effective approach is partitioning data based on an independent variable such as time or batch, where models are tested on a completely separate time period, such as a different year or a new batch. Another useful strategy is maintaining the same proportion of sample types (e.g., different classes or categories) in both the calibration and test sets, which prevents biases introduced by imbalanced partitions. Additionally, ensuring that the range and distribution of the reference values ($\mathbf{y}$) is well represented in both subsets helps maintain a fair evaluation, reducing the risk of models being overly optimised for a specific domain. As a general rule of thumb, it is required to define the work domain of the model and to offer an external validation set that covers that domain. In this sense, by incorporating these structured validation strategies, models can be assessed more rigorously, ultimately leading to more reliable and generalisable results.

In parallel to the strategies used for external validation, Figure 7 illustrates the distribution of calibration-to-validation sample ratios reported in the studies that used external validation. The most commonly used ratios cluster between 2:1 and 4:1, with 3:1 being the most frequent, indicating a tendency to allocate approximately 75% of the data for calibration and 25% for validation. Ratios below 5:1 were much less common, suggesting that most studies use a reasonably sized validation set rather than a minimal proportion of data for performance assessment. The presence of ratios of 1:1 and above indicates that a small number of studies used validation sets as large as or larger than their calibration sets, which may occur when there is a high availability of samples.

The distinction between exact and approximate proportions, represented by different shades in the histogram in Figure 7, describes a difference in how calibration and validation sets are defined, with some studies adhering strictly to predetermined ratios and others using exact sample numbers, probably due to outlier removal or other data curation processes. In general, the relatively low variability in these ratios underscores that even if there is no standardised guideline for this choice, practitioners reach a certain degree of consensus on this topic.

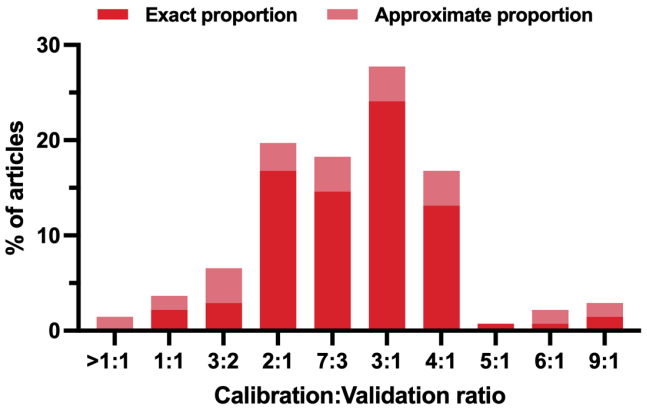

FIGURE 7 | Histogram of the ratio of samples used for the calibration and external validation subsets in the reviewed articles. $N = 170$ articles used external validation. Noncommon proportions were approximated to simple ones due to most of them being caused by the removal of some samples (like outliers) from one or both of the groups.

4.2 | Cross-Validation

Among the 200 reviewed articles, 83.5% employed some form of CV, while the remaining 16.5% either did not use CV or did not include it (Figure 8a). Three of the articles used more than one CV method. The most commonly reported method was random subsets CV, followed by LOO-CV and venetian blinds CV. However, approximately 25% of studies did not disclose the CV method used, further emphasising the lack of methodological transparency in validation practices. The use of resampling strategies was limited, only six articles performing Monte Carlo validation, despite their usefulness for the evaluation of model robustness and hyperparameter tuning.

Figure 8b details the k-fold strategies used by 51 studies employing random k-fold CV and eight studies using venetian blinds CV. The most frequently used k-values were 10-fold and fivefold, which may be explained by practicality reasons. Lower k-values (e.g., 3, 4 and 5) may help avoid overfitting by ensuring a more substantial set in each iteration. Higher k-values may lead to overoptimistic model performance estimates, as the validation subsets become too small to provide a reliable assessment of model generalisability.

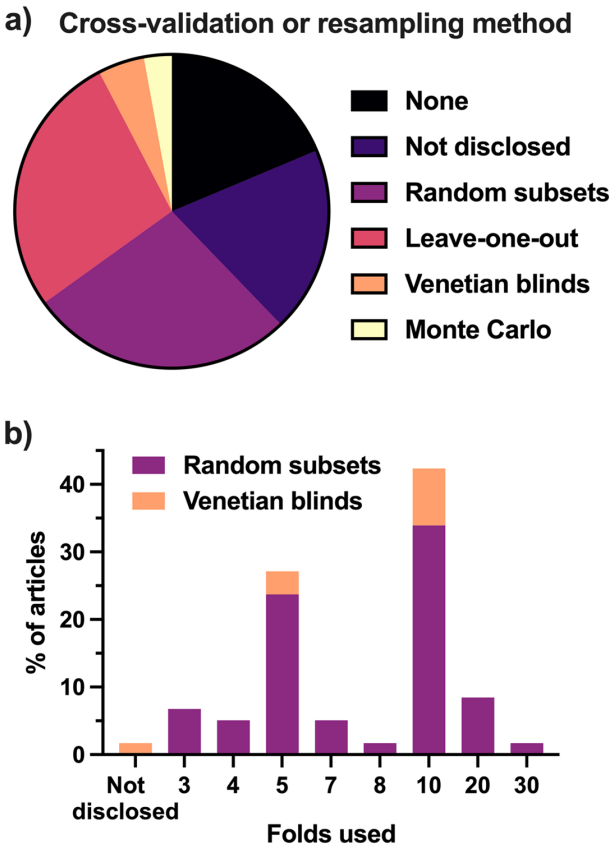

FIGURE 8 | (a) Proportion of cross-validation strategies used in the reviewed articles. $N = 167$ articles used cross-validation. (b) Histogram of k-fold strategies used in the reviewed articles. $N = 51$ articles used k-fold random cross-validation. $N = 8$ articles used venetian blinds cross-validation.

Interestingly, only four of the studies performed more than one iteration in the CV; this is, they performed random k-fold splitting more than once. This practice offers more robust CV metrics as they are calculated over several random data splittings and then averaged. However, the number of iterations used also varied widely (10, 25 and 100 iterations were used).

These findings highlight the need for more rigorous and transparent CV reporting. The variability in CV methodologies, combined with a lack of justification for k-fold choices in many studies, complicates direct comparisons between reported model performances. Standardising CV practices and encouraging authors to justify their validation strategy would enhance reproducibility and the reliability of chemometric models.

4.3 | General Considerations

To systematically and comprehensively assess the methodological robustness of validation practices in the reviewed literature, a validation rigorousness mark framework was implemented. This framework categorises studies into nine categoric levels, ranging from 1 (*least rigorous*) to 9 (*most rigorous*), based on the presence and description of CV and external validation methods, as described in Table 1. However, it should be noted that this does not ensure a proper validation, as the best methodology to be used in each study may change based on the characteristics of the dataset. Figure 9 shows the distribution of validation rigorousness marks across the reviewed studies.

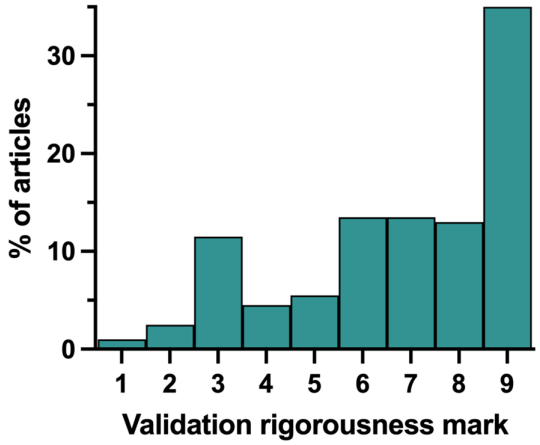

FIGURE 9 | Histogram of validation rigorousness mark for the reviewed articles. $N = 200$ articles.

The distribution reveals a clear trend toward higher validation scores, with the largest proportion of studies, over 30% of the articles, receiving the maximum mark (9). This means that a substantial number of studies implement both CV and external validation strategies while providing methodological descriptions for them, reflecting a growing commitment to rigorous validation practices among practitioners in NIR-based regression research for food applications. In comparison, relatively few studies received the lowest marks (< 6), indicating that only a small proportion of articles failed to apply or sufficiently describe validation strategies. However, the articles with mark 3 are more frequent among them, representing studies that used and described CV methods but did not use any external validation, relying on the metrics provided by the CV for assessing the generalisability of the model, often overoptimistic [54]. Also, it was found that some articles that do not use CV but do external validation use the validation set for optimising the models [55]; this also may result in overfitting and overoptimistic performance metrics. These findings underscore the need for continued emphasis on best practices in model validation to enhance reproducibility and reliability of chemometric models.

When studying the validation rigorousness mark in relation to sample size, in relation to the corresponding journal impact factors (IFs), or the average IF of editorials, no clear trends emerge (Figures S1, S2 and S3). This suggests that validation rigour is not inherently linked to journal prestige or editorial influence. Furthermore, given that only a few journals achieve the highest validation scores, these findings highlight the need for all journals and publishers to enforce stricter validation reporting requirements to ensure methodological transparency and reliability.

However, once the validation strategy is set, quantitative metrics are needed for evaluating the chemometric models, and this is where FoM plays an important role. These metrics are important for assessing model performance, comparing different algorithms and ensuring reproducibility across studies. But, despite their importance, the selection and reporting of FoM vary widely among studies, which can hinder direct comparisons and the reproducibility of results. A histogram illustrating the distribution of FoM reported in the 200 reviewed studies is shown in Figure 10. The most frequently used FoM are R and R^2, collectively appearing in over 90% of the articles, underscoring their widespread acceptance as key indicators of model fit. For evaluating model performance, RMSE is, by far, the most used metric, appearing in more than 80% of the articles. Another commonly used metric is the RPD, appearing in almost 50% of the articles, which is useful for standardising prediction error estimation. However, many articles use the RPD to 'objectively' categorise the proposed models as good or bad based on an arbitrary scale, instead of interpreting the results in depth, which may lead to misleading conclusions [56].

Notably, while all of the studies report FoM, only 33% of them explicitly provide the mathematical formula for any of the reported metrics. This also highlights a lack of methodological transparency, especially in FoM, where the definition is not clear or there is more than one available definition in the literature, such as bias or LOD and LOQ [45]. In addition, different software may calculate FoM in slightly different ways, the RMSE, for instance, and explicitly providing the formula ensures clarity and facilitates meaningful comparisons between studies. Bias, in particular, is a FoM that can be defined and calculated in multiple ways, leading to confusion when not explicitly described. It may refer to the mean difference between predicted and reference values, the systematic error in a model or even the slope deviation from unity in a regression analysis.

Additionally, the use of multiple error-based metrics (e.g., RMSE, SE, MSE and RE) in a single study, even interchangeably as some authors do [57], may be confusing and does not add significant value, as these metrics are mathematically related and convey similar information [41]. Other authors, however, base the discussion of results solely on the use of RMSE or another error-based metric, instead of offering a comprehensive evaluation of the proposed regression model, and can often be a biased or incomplete interpretation [44].

Furthermore, related to FoM, only 73.5% of the reviewed articles included a figure describing model prediction performance (predicted versus reference or similar), allowing for visual inspection of model characteristics. These plots are important for assessing key aspects such as linearity, linear range, bias and dispersion of residuals, providing a more intuitive evaluation of model quality beyond numerical FoM. These figures take on even more relevance depending on the choice of reported FoM, which may fail to capture certain model deficiencies [58]. Notably, while these visualisations are crucial, it is also important to consider whether calculating the p-value of the predicted versus measured values is truly necessary. A statistically significant correlation does not always imply that the model is robust or useful, as significance can be influenced by sample size and other factors unrelated to actual predictive performance.

In general, a more standardised and comprehensive approach to reporting FoM could improve clarity and prevent redundancies and missing information in model evaluation. Moreover, as explained, the inconsistency in their definitions or the lack of them affects the dissemination and practical applicability of results, particularly in regulatory or industrial settings where precise definitions are essential. Using proper validation strategies is essential; however, if the use of FoM is inconsistent or nonrigorous and nontransparent, the interpretation of results may not be accurate or comparable between studies, potentially offering misleading conclusions.

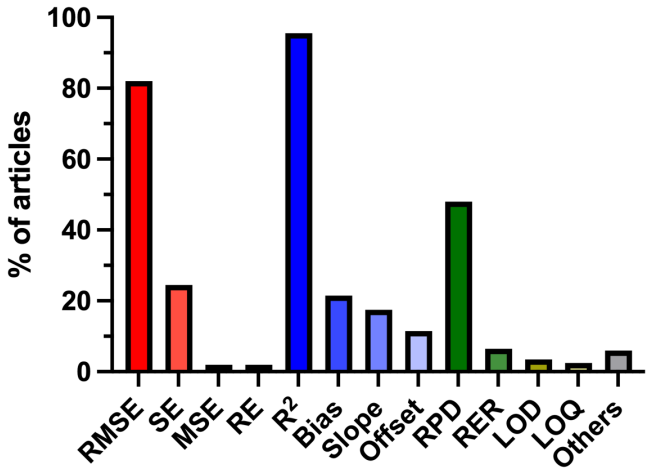

FIGURE 10 | Histogram of figures of merit used in the reviewed articles. $N = 200$ articles. In red: FoM related to prediction error. In blue: FoM related to predicted versus measured fit. In green: FoM that relativise prediction error. In yellow: FoM related to the limits of the model. In grey: other FoM.

5 | Conclusions

This systematic review highlights the important role of validation in chemometric regression models applied to NIRS for food analysis. The results revealed that PLS regression is the most commonly used approach, reflecting its suitability for handling collinear spectral data and its well-established performance in NIRS. While alternative machine learning approaches have been explored to a lesser extent, their presence in some studies suggests a growing interest in expanding the range of available methodologies.

Furthermore, the findings emphasise the importance of robust validation strategies, including both CV and external validation, to ensure the reliability and generalisability of predictive models. Despite the widespread adoption of somewhat standard validation methods, significant gaps remain regarding the reporting of these validation strategies and FoM. To enhance the rigour and reproducibility of these analytical methods, it is essential to adopt a more a rigorous approach. Future studies should ensure comprehensive validation strategies that incorporate both CV and external validation, with clear justification for the selected methods, to improve comparability between studies.

While most studies implemented some form of external validation, many did not provide explicit details on their validation methodology, potentially affecting the reproducibility of their results. Random data splitting remains the most commonly used external validation approach, yet its limitations in providing a truly independent and robust assessment suggest a need for more structured validation techniques, such as custom or fully independent test sets. Similarly, while CV was employed in the majority of studies, its implementation varied widely, with inconsistent reporting of *k*-fold parameters, iteration numbers and resampling methods.

The assessment of FoM further underscored the need for more rigorous reporting practices. While correlation-based metrics such as R^2 and error metrics like RMSE were widely used, the inconsistent definition and application of other FoM, especially those that have different possible definitions and have not been standardised, such as bias or LOD, complicate direct comparisons across studies. The frequent omission of mathematical formulas for these ambiguous metrics further exacerbates these issues. Additionally, despite their importance in assessing model performance, visual representations of model performance were absent in a considerable number of studies, limiting the ability to critically evaluate the proposed chemometric models. The standardisation of FoM, including explicit definitions, would help mitigate inconsistencies in performance evaluation, and the inclusion of visual performance assessments, such as predicted versus observed plots, would facilitate a more intuitive evaluation of model quality.

The growing reliance on NIR-based chemometric models in food analysis underscores the necessity of rigorous validation to maintain scientific integrity and practical applicability. By implementing standardised and transparent validation practices, the field can move toward greater applicability, fostering the development of more reliable and generalisable predictive models for food quality and safety applications.

Acknowledgements

During the preparation of this work, the authors used ChatGPT 4o in order to edit text and improve readability. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

References

1. What Is Chemometrics?" in *Chemometrics*, vol. 2016 (Wiley, 2016): 1–13, https://doi.org/10.1002/9783527699377.ch1.

2. J. U. Porep, D. R. Kammerer, and R. Carle, "On-Line Application of Near Infrared (NIR) Spectroscopy in Food Production," *Trends in Food Science and Technology* 46, no. 2 (2015): 211–230, https://doi.org/10.1016/j.tifs.2015.10.002.

3. E. Frank lldiko and J. H. Friedman, "A Statistical View of Some Chemometrics Regression Tools," *Technometrics* 35, no. 2 (1993): 109–135, https://doi.org/10.1080/00401706.1993.10485033.

4. H.-P. Wang, P. Chen, J.-W. Dai, et al., "Recent Advances of Chemometric Calibration Methods in Modern Spectroscopy: Algorithms, Strategy, and Related Issues," *TrAC Trends in Analytical Chemistry* 153 (2022): 116648, https://doi.org/10.1016/j.trac.2022.116648.

5. P. Dardenne, G. Sinnaeve, and V. Baeten, "Multivariate Calibration and Chemometrics for Near Infrared Spectroscopy: Which Method?," *Journal of Near Infrared Spectroscopy* 8, no. 4 (2000): 229–237, https://doi.org/10.1255/jnirs.283.

6. L. E. Agelet and C. R. Hurburgh, "A Tutorial on Near Infrared Spectroscopy and Its Calibration," *Critical Reviews in Analytical Chemistry* 40, no. 4 (2010): 246–260, https://doi.org/10.1080/10408347.2010.515468.

7. F. Westad and F. Marini, "Validation of Chemometric Models – A Tutorial," *Analytica Chimica Acta* 893 (2015): 14–24, https://doi.org/10.1016/j.aca.2015.06.056.

8. J. J. Roberts and D. Cozzolino, "An Overview on the Application of Chemometrics in Food Science and Technology—An Approach to Quantitative Data Analysis," *Food Analytical Methods* 9, no. 12 (2016): 3258–3267, https://doi.org/10.1007/s12161-016-0574-7.

9. E. Lopez, J. Etxebarria-Elezgarai, J. M. Amigo, and A. Seifert, "The Importance of Choosing a Proper Validation Strategy in Predictive Models. A Tutorial With Real Examples," *Analytica Chimica Acta* 1275 (2023): 341532, https://doi.org/10.1016/j.aca.2023.341532.

10. F. Allegrini and A. C. Olivieri, "Figures of Merit," in *Comprehensive Chemometrics*, vol. 2 (Elsevier, 2020): 441–463, https://doi.org/10.1016/B978-0-12-409547-2.14612-8.

11. J. M. Andrade-Garda, R. Boqué-Martí, J. Ferré-Baldrich, and A. Carlosena-Zubieta, "Partial Least-Squares Regression," in *Basic Chemometric Techniques in Atomic Spectroscopy*, ed. J. M. Andrade-Garda (Cambridge: Royal Society of Chemistry, 2009): 181–243, https://doi.org/10.1039/9781847559661-00181.

12. R. Trinchero and F. Canavero, "Machine Learning Regression Techniques for the Modeling of Complex Systems: An Overview," *IEEE Electromagnetic Compatibility Magazine* 10, no. 4 (2021): 71–79, https://doi.org/10.1109/MEMC.2021.9705310.

13. R. Rosipal and L. J. Trejo, "Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space," *Journal of Machine Learning Research* 2 (2001): 97–123.

14. P. J. Drew and J. R. T. Monson, "Artificial Neural Networks," *Surgery* 127, no. 1 (2000): 3–11, https://doi.org/10.1067/MSY.2000.102173.

15. G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme Learning Machine for Regression and Multiclass Classification," *IEEE Transactions on Systems, man, and Cybernetics, Part B (Cybernetics).* 42, no. 2 (2012): 513–529, https://doi.org/10.1109/TSMCB.2011.2168604.

16. P. F. Smith, S. Ganesh, and P. Liu, "A Comparison of Random Forest Regression and Multiple Linear Regression for Prediction in Neuroscience," *Journal of Neuroscience Methods* 220, no. 1 (2013): 85–91, https://doi.org/10.1016/j.jneumeth.2013.08.024.

17. J. A. Cayuela-Sánchez, A. El Ouaddari, A. El Amrani, and J. Jamal-Eddine, "Rapid Determination of Essential Oils Functional Groups Using Compositional Methods and VisNIR Spectroscopy," *Journal of Pharmaceutical and Biomedical Analysis* 227 (2023): 115278, https://doi.org/10.1016/j.jpba.2023.115278.

18. T. Fearn, "Chemometrics: An Enabling Tool for NIR," *NIR News* 16, no. 7 (2005): 17–19, https://doi.org/10.1255/nirn.856.

19. C. Beleites and R. Salzer, "Assessing and Improving the Stability of Chemometric Models in Small Sample Size Situations," *Analytical and Bioanalytical Chemistry* 390, no. 5 (2008): 1261–1271, https://doi.org/10.1007/s00216-007-1818-6.

20. C. L. Ramspek, K. J. Jager, F. W. Dekker, C. Zoccali, and M. van Diepen, "External Validation of Prognostic Models: What, Why, How, When and Where?," *Clinical Kidney Journal* 14, no. 1 (2021): 49–58, https://doi.org/10.1093/ckj/sfaa188.

21. E. W. Steyerberg, F. E. Harrell, G. J. J. M. Borsboom, M. J. C. Eijkemans, Y. Vergouwe, and J. D. F. Habbema, "Internal Validation of Predictive Models," *Journal of Clinical Epidemiology* 54, no. 8 (2001): 774–781, https://doi.org/10.1016/S0895-4356(01)00341-9.

22. R. W. Kennard and L. A. Stone, "Computer Aided Design of Experiments," *Technometrics* 11, no. 1 (1969): 137–148, https://doi.org/10.1080/00401706.1969.10490666.

23. A. Saptoro, M. O. Tadé, and H. Vuthaluru, "A Modified Kennard-Stone Algorithm for Optimal Division of Data for Developing Artificial Neural Network Models," *Chemical Product and Process Modeling* 7, no. 1 (2012), https://doi.org/10.1515/1934-2659.1645.

24. R. Galvao, M. Araujo, G. José, M. Pontes, E. Silva, and T. Saldanha, "A Method for Calibration and Validation Subset Partitioning," *Talanta* 67, no. 4 (2005): 736–740, https://doi.org/10.1016/j.talanta.2005.03.025.

25. N. B. Gallagher and D. O'Sullivan, *Selection of Representative Learning and Test Sets Using the Onion Method* (Eigenvector Inc., 2020).

26. H. Tian, L. Zhang, M. Li, et al., "Weighted SPXY Method for Calibration Set Selection for Composition Analysis Based on Near-Infrared Spectroscopy," *Infrared Physics & Technology* 95 (2018): 88–92, https://doi.org/10.1016/j.infrared.2018.10.030.

27. W. Gani and M. Limam, "A Kernel Distance-Based Representative Subset Selection Method," *Journal of Statistical Computation and Simulation* 86, no. 1 (2016): 135–148, https://doi.org/10.1080/00949655.2014.996758.

28. V. Fonseca Diaz, B. De Ketelaere, B. Aernouts, and W. Saeys, "Cost-Efficient Unsupervised Sample Selection for Multivariate Calibration," *Chemometrics and Intelligent Laboratory Systems* 215 (2021): 104352, https://doi.org/10.1016/J.CHEMOLAB.2021.104352.

29. A. L. Pomerantsev and O. Rodionova, "Subset Selection Using Combined Analytical Signal," *Microchemical Journal* 190 (2023): 108654, https://doi.org/10.1016/j.microc.2023.108654.

30. G. G. Siano and H. C. Goicoechea, "Representative Subset Selection and Standardization Techniques. A Comparative Study Using NIR and a Simulated Fermentative Process UV Data," *Chemometrics and Intelligent Laboratory Systems* 88, no. 2 (2007): 204–212, https://doi.org/10.1016/j.chemolab.2007.05.002.

31. R. D. Clark, "OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets," *Journal of Chemical Information and Computer Sciences* 37, no. 6 (1997): 1181–1188, https://doi.org/10.1021/CI970282V.

32. V. R. Joseph and A. Vakayil, "SPlit: An Optimal Method for Data Splitting," *Technometrics* 64, no. 2 (2022): 166–176, https://doi.org/10.1080/00401706.2021.1921037.

33. R. Wehrens, H. Putter, and L. M. C. Buydens, "The Bootstrap: A Tutorial," *Chemometrics and Intelligent Laboratory Systems* 54, no. 1 (2000): 35–52, https://doi.org/10.1016/S0169-7439(00)00102-7.

34. N. L. Afanador, T. N. Tran, and L. M. C. Buydens, "An Assessment of the Jackknife and Bootstrap Procedures on Uncertainty Estimation in the Variable Importance in the Projection Metric," *Chemometrics and Intelligent Laboratory Systems* 137 (2014): 162–172, https://doi.org/10.1016/j.chemolab.2014.05.016.

35. O. M. Kvalheim, B. Grung, and T. Rajalahti, "Number of Components and Prediction Error in Partial Least Squares Regression Determined by Monte Carlo Resampling Strategies," *Chemometrics and Intelligent Laboratory Systems* 188 (2019): 79–86, https://doi.org/10.1016/j.chemolab.2019.03.006.

36. F. Lindgren, B. Hansen, W. Karcher, M. Sjöström, and L. Eriksson, "Model Validation by Permutation Tests: Applications to Variable Selection," *Journal of Chemometrics* 10, no. 5–6 (1996): 521–532, https://doi.org/10.1002/(SICI)1099-128X(199609)10:5/6<521::AID-CEM448>3.0.CO;2-J.

37. S. Kucheryavskiy, S. Zhilin, O. Rodionova, and A. Pomerantsev, "Procrustes Cross-Validation—A Bridge Between Cross-Validation and Independent Validation Sets," *Analytical Chemistry* 92, no. 17 (2020): 11842–11850, https://doi.org/10.1021/acs.analchem.0c02175.

38. P. Filzmoser, B. Liebmann, and K. Varmuza, "Repeated Double Cross Validation," *Journal of Chemometrics* 23, no. 4 (2009): 160–171, https://doi.org/10.1002/cem.1225.

39. L. Martino, V. Laparra, and G. Camps-Valls, "Probabilistic Cross-Validation Estimators for Gaussian Process Regression," in *2017 25th European Signal Processing Conference (EUSIPCO)* (IEEE, 2017): 823–827, https://doi.org/10.23919/EUSIPCO.2017.8081322.

40. Q. F. Gronau and E.-J. Wagenmakers, "Limitations of Bayesian Leave-One-Out Cross-Validation for Model Selection," *Computational Brain & Behavior* 2, no. 1 (2019): 1–11, https://doi.org/10.1007/s42113-018-0011-7.

41. C. J. Willmott, K. Matsuura, and S. M. Robeson, "Ambiguities Inherent in Sums-of-Squares-Based Error Statistics," *Atmospheric Environment* 43, no. 3 (2009): 749–752, https://doi.org/10.1016/j.atmosenv.2008.10.005.

42. T. Fearn, "Assessing Calibrations: SEP, RPD, RER and R2," *NIR News* 13, no. 6 (2002): 12–13, https://doi.org/10.1255/nirn.689.

43. A. G. Asuero, A. Sayago, and A. G. González, "The Correlation Coefficient: An Overview," *Critical Reviews in Analytical Chemistry* 36, no. 1 (2006): 41–59, https://doi.org/10.1080/10408340500526766.

44. V. Bellon-Maurel, E. Fernandez-Ahumada, B. Palagos, J.-M. Roger, and A. McBratney, "Critical Review of Chemometric Indicators Commonly Used for Assessing the Quality of the Prediction of Soil Attributes by NIR Spectroscopy," *TrAC Trends in Analytical Chemistry* 29, no. 9 (2010): 1073–1081, https://doi.org/10.1016/j.trac.2010.05.006.

45. B. Giussani, G. Gorla, J. Ezenarro, J. Riu, and R. Boqué, "Navigating the Complexity: Managing Multivariate Error and Uncertainties in Spectroscopic Data Modelling," *TrAC Trends in Analytical Chemistry* 181 (2024): 118051, https://doi.org/10.1016/j.trac.2024.118051.

46. M. C. Ortiz, L. A. Sarabia, A. Herrero, et al., "Capability of Detection of an Analytical Method Evaluating False Positive and False Negative (ISO 11843) With Partial Least Squares," *Chemometrics and Intelligent Laboratory Systems* 69, no. 1–2 (2003): 21–33, https://doi.org/10.1016/S0169-7439(03)00110-2.

47. A. C. Olivieri, "Analytical Figures of Merit: From Univariate to Multiway Calibration," *Chemical Reviews* 114, no. 10 (2014): 5358–5378, https://doi.org/10.1021/cr400455s.

48. D. Chicco, M. J. Warrens, and G. Jurman, "The Coefficient of Determination R-Squared Is More Informative Than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation," *PeerJ Computer Science* 7 (2021): 1–24, https://doi.org/10.7717/PEERJ-CS.623/SUPP-1.

49. C. Hartmann, J. Smeyers-Verbeke, W. Penninckx, and D. L. Massart, "Detection of Bias in Method Comparison by Regression Analysis," *Analytica Chimica Acta* 338, no. 1–2 (1997): 19–40, https://doi.org/10.1016/S0003-2670(96)00341-8.

50. M. J. Page, J. E. McKenzie, P. M. Bossuyt, et al., "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews," *Journal of Clinical Epidemiology* 134 (2021): 178–189, https://doi.org/10.1016/j.jclinepi.2021.03.001.

51. G. D. Smith and K. H. M. Ho, "Systematic Reviews: When Should They Be Updated?," *Journal of Clinical Nursing* 32, no. 9–10 (2023): e17–e18, https://doi.org/10.1111/jocn.16547.

52. S. Heo, A. Y. Chan, P. Diaz Peralta, L. Jin, C. R. Pereira Nunes, and M. L. Bell, "Impacts of the COVID-19 Pandemic on Scientists' Productivity in Science, Technology, Engineering, Mathematics (STEM), and

Medicine Fields," *Humanities and Social Sciences Communications* 9, no. 1 (2022): 1–11, https://doi.org/10.1057/s41599-022-01466-0.

53. W. Zhang, L. C. Kasun, Q. J. Wang, Y. Zheng, and Z. Lin, "A Review of Machine Learning for Near-Infrared Spectroscopy," *Sensors* 22, no. 24 (2022): 9764, https://doi.org/10.3390/S22249764.

54. K. H. Esbensen and P. Geladi, "Principles of Proper Validation: Use and Abuse of Re-Sampling for Validation," *Journal of Chemometrics* 24, no. 3–4 (2010): 168–187, https://doi.org/10.1002/cem.1310.

55. P. Udomkun, P. Rungpichayapichet, N. Phuangcheen, and B. Innawong, "Rapid Determination of Fructooligosaccharide in Solar-Dried Banana Syrup by Using Near-Infrared Spectroscopy," *Journal of Food Measurement and Characterization* 15, no. 4 (2021): 3397–3407, https://doi.org/10.1007/s11694-021-00911-z.

56. B. Minasny and A. Mcbratney, "Why You Don't Need to Use RPD," *Pedometron*. 33 (2013): 14–15.

57. P. Taradolsirithitikul, P. Sirisomboon, and S. C. Dachoupakan, "Qualitative and Quantitative Analysis of Ochratoxin a Contamination in Green Coffee Beans Using Fourier Transform Near Infrared Spectroscopy," *Journal of the Science of Food and Agriculture* 97, no. 4 (2017): 1260–1266, https://doi.org/10.1002/jsfa.7859.

58. M. W. Liemohn, A. D. Shane, A. R. Azari, A. K. Petersen, B. M. Swiger, and A. Mukhopadhyay, "RMSE Is Not Enough: Guidelines to Robust Data-Model Comparisons for Magnetospheric Physics," *Journal of Atmospheric and Solar - Terrestrial Physics* 218 (2021): 105624, https://doi.org/10.1016/j.jastp.2021.105624.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.