Advances in Machine Learning Framework for Near-Infrared Spectroscopy: A Taxonomic Review on Food Quality Assessment

Nguyen Thi Hoang Phuong[1], Hieu Nguyen Van[2], Xuan Nguyen Thi Thanh[2], Phien Nguyen Ngoc[3,4,*]
[1]Faculty of Information Technology, Pham Van Dong University, Quang Ngai, Vietnam
[2]The University of Danang, University of Science and Technology, Viet Nam
[3]Center for Applied Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam
[4]Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam
E-mail: nthphuong@pdu.edu.vn, nvhieuqt@dut.udn.vn, nttxuan@dut.udn.vn, nguyenngocphien@tdtu.edu.vn
*Corresponding author

Keywords: Machine learning, deep learning, near-infrared spectroscopy, advances in framework, food quality

Received: November 01, 2024

This review taxonomically analyzes and evaluates recent advances in machine learning (ML) frameworks applied to near-infrared spectroscopy (NIRS) for food quality assessment. Through a comprehensive literature search across IEEE Explore, ScienceDirect, and Springer (2021-2024), we examine key framework components: data acquisition, public datasets, preprocessing, wavelength selection, and advanced ML architectures. Our analysis reveals the current state: miniaturized devices and multi-device data collection are expanding spectral coverage, while public datasets focus mainly on nutritional indices, lacking safety-related data. Framework-wide challenges persist in device compatibility, dataset comprehensiveness, and model interpretability. Recent advances show promising developments through: specialized deep learning architectures achieving 97-100% accuracy, data transformation techniques (2D-COS, GAFD) enhancing interpretability, hybrid traditional-deep learning models, and effective transfer learning for cross-device applications. Based on these insights, we propose three critical research directions: expanding food safety datasets through regulatory partnerships, developing multi-level fusion for heterogeneous device data, and creating automated techniques for model optimization and interpretability. These directions are vital for advancing ML-NIRS applications in food quality assessment, improving both efficiency and reliability.

Povzetek: Analizirani so napredki v strojnih učnih modelih za spektroskopijo bližnjega infrardečega spektra (NIRS) pri oceni kakovosti hrane. Pregled obsega ključne komponente, kot so zbiranje podatkov, predprocesiranje in izbira valovnih dolžin. Predlagane so tri raziskovalne smeri: širitev podatkovnih zbirk, razvoj fuzije večnivojskih podatkov in avtomatizacija optimizacije modelov za boljšo zanesljivost ocenjevanja kakovosti hrane.

1 Introduction

Food quality and safety have emerged as critical concerns for both the food industry and global consumers [1]. The burden of foodborne diseases and economic losses due to poor quality or spoiled food at the production and distribution stages is enormous [2], requiring careful monitoring of food composition regularly. NIRS, as an analytical technique that can provide complex "chemical fingerprints" of food samples related to their composition, quality, and safety [3–5], has been combined with classical statistical methods and advanced ML to address this issue.

ML techniques and IoT development have brought about considerable changes in many fields [6–8]. Applying ML, mainly supervised learning, to multivariate NIRS spectral analysis has significantly changed food quality assessment and assurance. These studies have been diverse across various data types and increased rapidly in the past two years [9]. From 2022 to 2024, along with traditional ML methods such as Principal Component Analysis (PCA), Partial Least Squares (PLS), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Deep Learning (DL) such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Autoencoders (AE), as in Figure 1, there are four major trends in NIRS applications. This includes (1) detecting contaminated and adulterated food, identifying adulterants, and determining the level and residual concentration of chemicals in agricultural and livestock products [10, 11]; (2) developing sustainable agriculture through monitoring crop growth, soil nutrients, and various components of crops to improve care and early detection and treatment of crop diseases [12]; (3) determining the optimal harvest time to achieve maximum economic yield [13]; and (4) evaluating product quality, particularly for high-value economic items [14–19], etc.

However, ML for NIRS is still inceptive compared to other fields for several reasons. First, ML on NIRS spectra requires specialized data, which is difficult to collect due

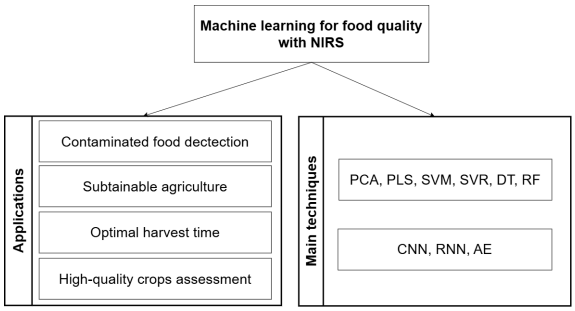

Figure 1: NIRS applications and ML techniques for NIRS

to expensive spectrometers and reference chemical data, in chemical content determination problems [9]. Second, studies are often published in agricultural or interdisciplinary chemometrics journals that combine data science and chemistry, as in Figure 2, so computer scientists' access to technical developments is more limited. With the potential of developing ML to solve social food quality problems, this needs to be further promoted by supporting a technical overview.

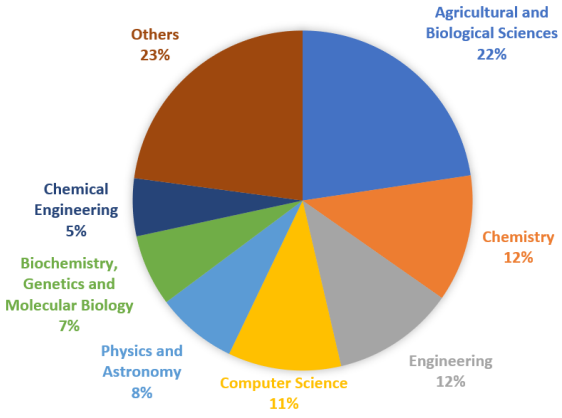

Figure 2: Subject area from 2021 to 2024 according to Scopus analysis in ML for NIRS food quality

Therefore, this study aims to shift the focus to technical surveys, technological advances, availability of public datasets, and development potentials not mentioned in previous review studies. We conduct two main review bases:

1. First, we summarize and discuss the contributions and limitations of recent review articles on ML in NIRS and food analysis in particular. From there, we find gaps that need to be exploited and further evaluated on techniques and data.

2. Second, we synthesize and evaluate recent new research articles, classifying and highlighting information on data (NIRS tools, data fusion techniques, public datasets), as well as ML techniques (preprocessing, wavelength selection, and advanced ML architectures) that have not been covered in existing reviews.

From this background, gaps were identified from a computer science perspective to conduct future research in food inspection.

2 Methodology

In the fourth quarter of 2023, a thorough investigation was conducted through IEEE Explore, ScienceDirect, and Springer, employing controlled vocabulary in ML, NIRS, and food quality analysis, as in Figure 3. The search focused on emerging ML techniques for NIRS in food quality assessment, utilizing specific terms such as "deep learning", "chemometrics", "NIR spectroscopy", and "food". In addition to these above primary keywords, we conducted deeper searches focusing on specific components of our ML framework. Each framework component served as secondary keywords - notably "preprocessing" and "wavelength selection" - to thoroughly identify recent studies focusing on improvements in these critical areas. For machine learning algorithms, we specifically searched for both traditional methods (PCA, PLS, SVM) and emerging deep learning architectures (CNN, RNN, AE, GAN) to track their evolution and applications in NIR analysis. Additionally, the NIR dataset was also searched extensively on Mendeley and Zenodo. This hierarchical search approach, structured according to our ML framework taxonomy, enabled us to systematically evaluate recent advances in specific methodological aspects rather than just general applications. Through this focused search strategy, we could better assess how recent research has contributed to advancing different components of the ML framework in NIR spectral analysis.

The literature review methodology branches into two distinct paths: (1) comprehensive review papers and (2) original research articles. The first branch focuses on conducting rigorous analyses of existing literature to identify key challenges, significant contributions, and unexplored territories within the machine learning domain for NIR analysis. The second branch encompasses original research papers, systematically categorized according to their contributions to the ML framework - spanning from data acquisition and public datasets to preprocessing/wavelength selection and advanced architectural innovations. This dual-branch approach ensures both a broad understanding of the field's current state through synthesized reviews and a detailed examination of specific technical advancements through original research contributions.

Stringent filters were applied, including English language restriction and consideration of peer-reviewed articles, reviews, books, book chapters, and conference papers from the four years (2021-2024). The research database has been updated to include publications up to September 2024 to ensure the most informed and up-to-date discussion. The

strategy aimed to capture the latest innovations at the intersection of machine learning and NIRS for non-destructive and rapid evaluation of diverse food quality traits, excluding older publications beyond the scope of emerging techniques.

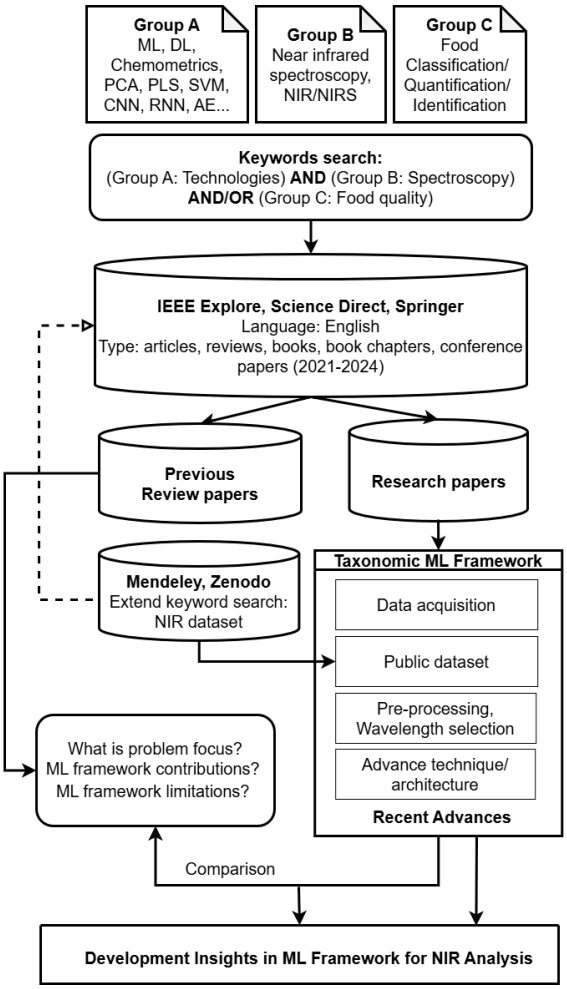

Figure 3: Research process flow chart

3 Previous review studies: an overview

Our analysis of previous ML in NIR spectroscopy for food quality and safety draws from 18 review articles published between 2021 and 2024. This comprehensive overview reveals significant progress in applying advanced ML techniques to NIR data and exposes critical challenges, as in Table 1, 2. While studies demonstrate the potential of methods ranging from traditional multivariate analysis to sophisticated deep learning algorithms, they also underscore persistent limitations in datasets, model optimization, and real-world applicability.

Building upon this overview, we conduct an additional review in the next section to explore unaddressed aspects that significantly impact machine learning trends in food spectroscopy. This supplementary analysis aims to fill crucial gaps and provide insights into emerging directions that could shape the future of ML-driven NIRS in food analysis.

These studies are mainly related to ML for food quality and safety and cover a range of applications or techniques. Recent review articles have examined hyperspectral imaging and NIR spectroscopy combined with advanced algorithms for non-invasive assessment of parameters, including nutritional composition (e.g., protein, moisture, fatty acids), adulteration/defect detection, and geographical origin discrimination in various food products. Both traditional multivariate analysis methods, like PCA and PLS, and increasingly sophisticated ML algorithms, including SVM, ANN, and CNNs, have been explored for relevant tasks such as multi-class food classification and quality prediction. This demonstrates the general feasibility of data-driven modeling approaches on spectroscopic data for food evaluation. However, significant limitations persist regarding dataset availability and model optimization, transferability, and interpretability, specifically for NIR food applications using advanced machine learning.

Despite the widespread application of NIRS in food quality and safety assessment, significant challenges remain in developing and sharing suitable datasets for machine learning research in this field. Firstly, current technique reviews tend to cover NIR datasets broadly without an in-depth analysis tailored to the food domain. Secondly, the mentioned datasets primarily consist of Vis-NIR spectral range ($<$ 1000 nm) stored in MATLAB data files, which poses challenges for developing ML research applications (typically developed with Python). However, this spectral range is often considered less informative in chemical information than the 1000-2500 nm [4]. Thirdly, other current shared datasets with broader coverage (400 - 2500 nm) are just suitable for simple classification tasks, lacking the detailed chemical information required for laboratory-based quality assessment regression problems. Besides, researchers also highlighted significant limitations in collecting valuable data, labeling, data enrichment, and practical deployment due to high costs. Last but not least, many previous studies utilizing NIRS for food quality and safety inspection have relied on small datasets, often with fewer than 200 samples, limiting model robustness and generalizability. Therefore, building NIRS spectral datasets with appropriate wavelength bands, relevant chemical parameterization, and proper labeling would be more practical when addressing real-world problems.

Additionally, reviews focusing specifically on deep learning also need more details regarding optimal network architectures, data requirements regarding sample size and variability, and quantitative benchmarking on relevant food NIR datasets. There is no in-depth discussion of deep learning or other advanced machine learning methods, nor is

Table 1: Previous Review Studies (1)

The issues	Review focus	Related Contributions	Limitations
Quantification of food bioactives by NIR spectroscopy: Current insights, long-lasting challenges, and future trends [3]	Factors affecting model performance. Algorithm used: Mostly PLS; SVM, MLR, BP-ANN, CNN	Effects of sample prep, analyte concentration, instrument features on performance. Compares benchtop/portable NIR. Proposes FAIR data management. Suggests theoretical calculations for interpretation.	Limited datasets ($<$ 200 samples). Difficulty in choosing preprocessing/regression methods. Interpretability/transferability issues. Lacks DL focus.
Food quality 4.0: From traditional approaches to digitalized automated analysis [5]	Traditional vs emerging techniques. Algorithm used: Mostly PLSR, PLS-DA; SVM	Industry 4.0 innovations (AI, DL, sensors) in spectroscopic. Portable/miniaturized NIR-AI for evaluation. HSI as non-destructive quality technique.	Brief NIRS-food analysis mention. No in-depth NIRS/DL applications with food datasets. Mostly Vis-NIR datasets.
DL for NIRS data modeling: Hypes and benefits [9]	Potential benefits and pitfalls of using DL for modeling NIRS	DL auto-transforms spectral data without preprocessing. Shallow DL success with small datasets ($<$ 1000). DL efficiency for complex food analysis tasks (multi-class, multi-response).	Small, under-optimized datasets in DL-chemometrics comparisons. Limited food NIR spectra DL modeling. No large food quality/safety datasets for DL benchmarking.
Multivariate analysis of food fraud: A review of NIR based instruments in tandem with chemometrics [10]	Chemometrics with NIRS, HSI. Algorithm used: Mostly PLSR, PLS-DA; SVM, PCA, SIMCA, ANN, KNN	Overview of NIR/HSI principles. Summarizes chemometrics for spectral processing. Reviews NIR/HSI with chemometrics for adulterant detection. Compares classification/regression models. Discusses advantages/limitations for food authenticity.	No DL discussion. Focuses on PCA, PLS, and SVM. No specific NIR datasets for food. Lab-prepared samples, not real-world fraud. Limited assessment of method robustness and applicability.
AI-based techniques for adulteration and defect detections in food and agricultural industry: A review [11]	AI techniques combined with sensors. Algorithm used (NIR-specific): Mostly SVM, PLSR, ANN, PCA, CNN, Random Forest.	AI for food authentication/quality. CNN (2015-2022). Challenges: technique standardization, algorithm selection, data fusion, fast detection, severity quantification, framework development.	Lack of sensor device, data acquisition, preprocessing details. Impact on model performance not discussed.
Computer vision and DL in insects for food and feed production [20]	Applications	CV and NIRS for non-invasive assessment of nutritional composition, moisture, protein, fat, fatty acids in live insects.	Limited NIRS applications mentioned. No technical aspects discussed.
Application of NIRS for the nondestructive analysis of wheat flour: A review [21]	Application. Algorithm used: MPLS, PLSR, RF, RBF, LDA.	NIR fundamentals, recent developments for wheat flour quality/safety assessment. Four development areas: data quality, chemometrics, affordable tools, data integration.	Focuses on traditional ML for classification/regression.
Quality analysis and authentication of nutraceuticals using near IR (NIR) spectroscopy: trends and applications [22]	Novel analytical trends and applications from a chemistry/metabolomics perspective	NIR trends for nutraceutical quality control (HSI, portable devices). Targeted/untargeted metabolomics applications. Geographical classification.	No DL discussion for NIR data. No specific NIR nutraceutical datasets were analyzed.

Table 2: Previous Review Studies (2)

The issues	Review focus	Related Contributions	Limitations
A research review on DL combined with HSI in multiscale agricultural sensing [23]	Applications and limitations. Algorithm used (specific to NIR): Mostly CNN, AE	DL models for food quality, ripeness, moisture, nitrogen, chlorophyll, sugar prediction. HSI range (250-2500 nm) is broader than NIRS.	Data collection/real-world application challenges. Limited food authentication focus. CNN, SAE, and RNN without detailed evaluation.
Efficient extraction of deep image features using CNN for applications in detecting and analyzing complex food matrices [24]	Principle, architectures, applications of feature extraction methods, CNNs	1D CNN feasibility for NIRS food classification/defect detection. CNN features outperform traditional ML. HSI-CNN examples for cereal variety/quality classification.	Not NIRS-specific. Limited NIRS-based food analysis datasets. Lacks DL challenges for food NIRS data.
A Review of ML for NIRS [25]	ML, especially DL. Algorithm used: Mostly PLS, ELM, SVR, SVM, SLFN, DT, RF, AE, CNN, RNN, LSTM, GRU, GAN.	Summarizes NIR modes, instruments, preprocessing, datasets, feature selection. Covers traditional ML (PLS, SVM, ELM) and DL (CNN, RNN, autoencoders) for NIR food data.	No in-depth ML-NIR food analysis. Limited food spectroscopy dataset details. No model performance comparison. No data augmentation/transfer learning discussion.
Are standard sample measurements still needed to transfer multivariate calibration models between NIR spectrometers? [26]	Recent developments in calibration transfer (CT) methods	Mentions DL for multivariate calibration and transfer learning in NIRS model updating.	No in-depth DL-NIR food analysis discussion. No public DL-NIR food datasets were mentioned. Brief food applications (temperature/form adaptation).
Recent advances and application of ML in food flavor prediction and regulation [27]	Algorithm used: SVM, DT, RF, KNN, ELM, ANN	Pinciples, advantages, application, challenges of ML for food flavor prediction/regulation.	Traditional ML focus. Few NIRS flavor prediction studies. Small sample sizes. Limited public NIRS flavor datasets.
ML applications for multi-source data of edible crops[28]	Fusion of multi-source data with ML techniques	CNN and ResNet for edible crop classification using 2D spectral/HSI.	Not NIRS-specific. No NIRS algorithm performance
AI in sensory and consumer studies of food products [29]	Applications. Algorithm used: ANN, SVM, CNN	ML, particularly NIRS, for predicting sensory responses from physicochemical data.	ANN common, but basic supervised learning prevalent. More DL research needed for complex spectroscopic data. Lack of public NIRS-sensory datasets limits validation/research.
DL in analytical chemistry [30]	Applications. Algorithm used: CNN, DNN, LSTM, GAN, AE	DL applications in analytical chemistry. AEs for cereals, CNNs for vibrational spectral data classification/quantification, chemical component determination, geographical discrimination.	No in-depth NIRS-food DL analysis. Lacks food application datasets discussion. No focus on DL techniques/datasets for food NIRS.
Recent advances in assessing qualitative and quantitative aspects of cereals using nondestructive techniques[31]	Chemometrics based AI & ML. Algorithm used: PLS, PCR, LDA, PLSR, PCA, KNN	Chemometrics for cereal analysis. NIRS for amylose, moisture, texture, authentication. Preprocessing, models, performance parameters for NIRS.	Lacks DL/large dataset focus for NIRS. No public datasets. Limited portable NIRS discussion for on-site cereal analysis.

there any analysis provided comparing the performance of different algorithms specifically for NIRS data. The reviews mainly focus on conventional chemometrics techniques like PCA, PLS, SVM, etc. Even where public reference datasets exist, few studies thoroughly validate and compare deep learning techniques using these resources.
Based on the mentioned contributions and limitations, in these review studies, we focus on exploring key aspects related to measurement devices, data collection, publicly available NIR food datasets, and new ML architectures applied to NIRS spectral data. This review aims to contribute to expanding multidisciplinary progress at the intersection of NIRS, data science, food science, and industrial applications.

4 Recent advances of machine learning for NIRS

Based on the comparison with previous research results in section 3, this section classifies new contributions to machine learning frameworks in processing NIR spectral data in classification, pattern recognition, and component content regression problems, as in Figrure 4.

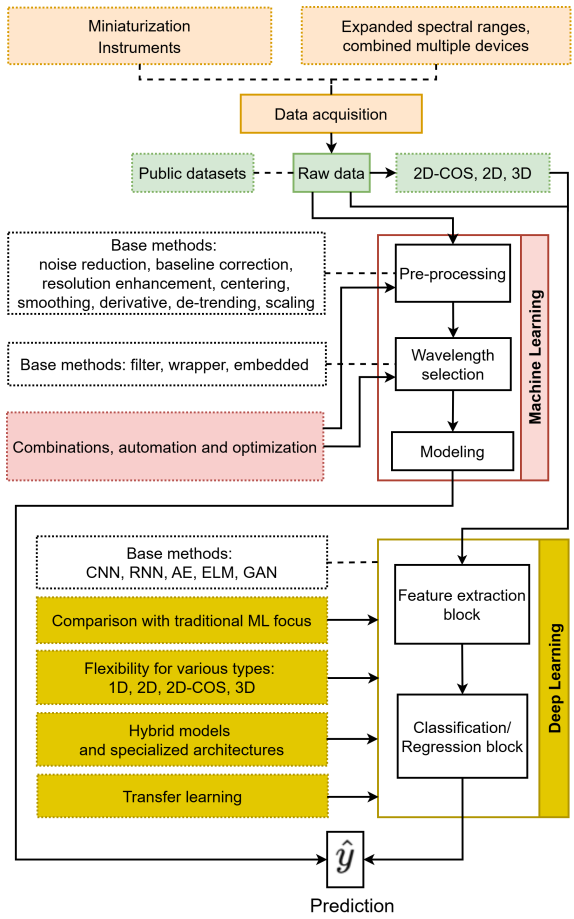

Figure 4: Advances in ML framework for NIRS

4.1 Data acquisition

Effective and accurate data acquisition using near-infrared spectrometers is crucial for food quality assurance. This relies on appropriate spectrometer selection, compatible measurement modes, and relevant reference components.

4.1.1 Miniaturization NIR spectral instruments

In recent years, NIR spectral instruments have undergone a notable transformation in size and portability, progressing from traditional benchtops to various compact, handheld, pocket-sized, miniaturized, and real-time versions [4, 32]. Traditionally, benchtop systems provide standardized design, broad spectral range, and high performance. For instance, the NIRSystems 6500 covers the full NIR range from 400 nm to 2500 nm, and the FT-NIR Frontier operates between 900-2500 nm [33]. The NIR spectral dataset covering the entire spectral range is typically measured using this instrument.
On the other hand, portable NIR spectrometers prioritize practical applications and offer greater flexibility, revolutionizing modern measurement techniques. However, this comes at the cost of design uniformity and spectral range. The diverse designs and technologies used in portable systems lead to variations in spectral coverage and resolution compared to their benchtop counterparts. For example, the SCiO operates within a narrow 740-1070 nm range [34], while the IAS-3120 [35], NIR-S-G1 [36], and NIR-M-R2 [37] function within the more common 900-1700 nm range. Specialized instruments like the Spectromètre portable NIR [38] cover a distinct 1750-2150 nm range. Many other devices from various manufacturers in [32, 34, 39–41] also have differing spectral ranges. This diversity challenges the development of multivariate analysis models with heterogeneous data. Despite these challenges, the ongoing miniaturization of NIR spectral instruments continues to expand their applicability across diverse fields, driving innovation in portable spectroscopic analysis.

4.1.2 Expanded spectral ranges and combined multiple devices

NIR spectrometers are capable of operating in various measurement modes: diffuse reflectance for solid surfaces; transmittance for gases, liquids, or cuvette-contained semi-solids; transflectance combining reflectance and transmittance for semi-solids without cuvettes; interactance using a specialized probe for enhanced solid sample analysis; and a transmittance mode accounting for scattering in dense samples [25, 42]. Each mode varies in cost, signal quality, and speed, and the selection depends on the specific needs of sample material and analysis requirements. These modes are determined by the detector, wavelength selector, and light source. For instance, the portable SCiO spectrometer uses a silicon array detector, a bandpass filter as the wavelength selector, and an LED light source, making it suitable for transmittance measurements of liquid or thin solid sam-

ples like vanilla solutions to determine vanillin content [34]. The DLPR NIRscan Nano employs an InGaAs detector, reflective diffraction grating, and tungsten lamps, enabling diffuse reflectance analysis of solid samples such as adulterated almond flour [43].

NIR datasets are typically collected by scanning multiple times and averaging values for each sample, potentially including 1500-1700 variables for the full NIR range [4]. However, achieving such comprehensive datasets is uncommon, due to challenges in sensor sensitivity across the broad spectral range and the high cost of instruments for full spectrum. NIRS data in various types of foods typically collected exclusively from either the short-wave (800-1100 nm, often together with the Vis-NIR region) or long-wave (1100-2500 nm) regions or the middle region [44]. The selection of NIR spectral range is crucial for analysis effectiveness. The key difference lies in signal strength across wavelengths [4]. In short-wave regions (700-1000 nm), signals become progressively weaker with 3rd and 4th overtones and show extensive band overlap, making detection and analysis more challenging. In contrast, long-wave regions (1000-2500 nm) contain stronger first and second overtones with better peak separation. This spectral range is especially effective at detecting absorption from specific molecular bonds - primarily those containing hydrogen atoms (O-H, C-H, N-H) and certain strong bonds (C=O, C≡N) - which produce characteristic absorption patterns useful for chemical analysis and machine learning model development. While long-wave regions typically provide better analytical performance, the choice of spectral range often requires balancing between analysis accuracy and economic considerations, as instruments for shorter wavelengths can use simple glass components, making them significantly more cost-effective than specialized equipment needed for longer wavelengths. The sub-1700 nm range is the most common due to its sufficiency in capturing key chemical components of the sample, as demonstrated by numerous studies [4, 44].

Recently, obtaining comprehensive and valuable information across the entire NIR spectral range often requires employing at least two instruments to capture data from both short-wave and long-wave regions. For instance, the NIR spectral of almond flour [43] were obtained from 900-2500 nm using three portable spectrometers: DLPR NanoNIRscan (900-1700 nm), MicroNIR 1700 (950-1650 nm), and NeoSpectra FT-NIR (1350-2500 nm) with 16 nm resolution. The classification models using portable near-infrared devices achieved 100% sensitivity and over 95% specificity in identifying almond flour adulterations above 5% (w/w), while the PLS regression models obtained coefficients of determination above 0.90 and RMSEP values between 3.2-4.8% for quantifying almond flour purity. Moreover, many real-time datasets are also collected with additional white reference and black reference spectra inside the instrument. These are two reference environments with no light at all, and light is reflected at 99.99%, which can be designed as supplementary components within standard measurement devices [45]. Compared to datasets measured in laboratory settings, as done previously, this trend of real-world data collection is more practical and is being targeted by food companies.

Therefore, the trend of using handheld or ultracompact devices for direct measurements at the production site, collecting the wide NIR spectrum by many devices, and supplementing reference spectra within the devices, are notable highlights in current food authentication data collection devices. This convenience creates significant opportunities for generating valuable datasets and improving food safety control.

Discussion: Development insights

The using multiple NIR spectral devices trend in food analysis offers opportunities for more comprehensive data collection but also presents challenges due to data heterogeneity. To address this issue and create larger datasets, developing effective data fusion techniques becomes crucial.

In NIRS, these multilevel fusion techniques enable integration and standardization of data from diverse instruments, effectively leveraging data resources from multiple institutions and organizations to build large, valuable datasets without depending on homogeneous equipment. Data fusion, as in [46–48], occurs at three levels: low-level fusion directly combines raw data from multiple sources, mid-level fusion integrates extracted features from different data sources, and high-level fusion combines decisions or interpretations made from separate data sources to conclude. Promising approaches for data fusion include data normalization, selection of important variables, application of advanced machine learning methods such as neural networks or transfer learning, utilization of multi-block data analysis techniques, and development of advanced spectral correction methods. These multilevel fusion techniques enable integration and standardization of data from diverse NIR instruments, effectively leveraging data resources from multiple institutions and organizations to build valuable large datasets without depending on homogeneous equipment, thereby significantly improving the effectiveness of multivariate analysis models in food quality control.

Moreover, compared to using a single handheld device with a narrow spectral range, fusing data from multiple devices to expand the spectral range offers advantages similar to laboratory benchtop with full-range spectra. Specifically, there are more additional spectral information from different spectral regions, facilitating better discrimination and quantification of components in complex samples with high spectral overlap while maintaining the mobility and convenience of handheld devices in practical applications.

4.2 Public datasets

This section reviews publicly accessible NIR spectroscopy food datasets, addressing the limitations of proprietary datasets in developing robust, generalizable machine learning models. These datasets typically include spectral data (often in .csv, .xlsx, or .unsb formats) spanning vari-

ous wavelength ranges, predominantly 900-2500 nm, with some datasets covering both visible and near-infrared regions (Vis-NIR). In NIR spectral archiving, the .unsb file format is a standard format similar to .csv and .xlsx. This is a proprietary file format of The Unscrambler software. To open and process this file, it is necessary to use The Unscrambler or convert it to a more familiar format for machine learning. Corresponding reference values for key food quality parameters are usually provided. While hyperspectral imaging (HSI) offers both spatial and spectral information, its large file sizes and complex processing limit widespread use in portable devices. Therefore, this review focuses on spectral data from NIRS, balancing information richness and practicality for rapid, non-destructive food quality assessment. The datasets cover diverse food products, highlighting the need for larger, more diverse, and standardized spectral databases to advance NIRS applications in food science.

4.2.1 Milk composition in transmittance mode

Collected over eight weeks on a dairy farm, the milk dataset in [45] includes transmittance mode spectra spanning the 960-1690 nm range for 1224 raw milk samples from 41 cows, each with a 2.86 nm/pixel resolution. The dataset incorporates raw milk spectra and white and dark reference spectra used for calibration. With accompanying laboratory reference values for essential milk components such as fat, protein, lactose, urea, and somatic cell count, this dataset goes beyond spec-tral information by including details like cow ID, milk yield, and time intervals between milkings. Formatted as a .csv file with comprehensive variable descriptions, the dataset aims to facilitate chemometric analysis and the development of multivariate calibration models for predicting milk parameters.

4.2.2 Handheld NIR for chicken breast filets

In a non-destructive manner, portable miniaturized NIR spectrometry captured diffuse reflectance data (908–1676 nm, with an evenly distributed spectral resolution, resulting in 125 variables/measurement) from chicken breast filets in [49]. This data helped differentiate fresh and thawed filets and assess bird growth conditions. NIR measurements were taken from 153 commercial chicken filet samples in three modes: direct contact with meat and through the top foil (with or without an air pocket). Thawed samples were generated by freezing and thawing. Multivariate statistics were applied to the 4590 raw NIR spectra.

4.2.3 SpectroFood dataset

The SpectroFood dataset in [50] is a comprehensive hyperspectral meta-dataset aimed at non-destructive estimation of dry matter content across multiple crops. It comprises visible/near-infrared (VIS/NIR) hyperspectral data coupled with corresponding dry matter measurements for four crops - apples, broccoli, leeks, and mushrooms. In total, 1028 samples were measured using four different calibrated hyperspectral imaging cameras across the spectral range of 398-1717 nm, with all measurements capturing the VIS/NIR range of 470-900 nm. Specifically, 240 apple samples were measured in the 430-990 nm range with 141 bands, 250 broccoli samples in the 470-900 nm range with 150 bands, 288 leek samples in the 398-1717 nm range with 421 bands, and 250 mushroom samples in the 400-998 nm range with 204 bands. The dataset provides the mean reflectance spectrum extracted for each sample in a tabular (.csv) format, along with the corresponding dry matter percentage which ranges from 8.1% to 87%. Additionally, the raw hyperspectral image data for each crop is also provided as .mat files. This multi-crop, multi-sensor dataset aims to facilitate the development of generalized AI/ML models for dry matter estimation that can robustly handle data from different imaging systems and crops.

4.2.4 Reflectance spectral dataset of pre-cooked pasta

This dataset in [51] comprises 1200 Vis-SWIR reflectance spectra (350-2500 nm, with 2151 variables/measurement) of 6 Pennette 72 and 6 Mezze Penne pre-cooked pasta samples with varying salt levels, measured in both frozen and thawed states. The spectra were non-destructively acquired using a portable ASD FieldSpec 4 Standard-Res spectrophotometer, with 50 spectra collected per sample. The data is provided as a .mat file containing a dataset object with rows labeled for sample ID, dry matter content (42.8%, 46.7%, 47.5%), pasta type, and physical state. The averaged spectra highlight differences in the visible region based on salt content, while frozen and thawed samples differed in reflectance intensities across most wavelength ranges, especially 350-1450 nm, 1600-1850 nm, 2100-2400 nm. This annotated Vis-SWIR dataset has valuable reuse potential for developing multivariate classification and regression models to rapidly inspect pre-cooked pasta quality by combining portable spectroscopy and chemometric techniques.

4.2.5 Sugar content measurements of grapes berries in various maturity stage

This dataset in [52] involves 274 samples, each composed of 100 grapes, representing three grape varieties: Syrah, Fer, and Mauzac. The dataset is structured as a CSV file, where rows represent samples and columns include variables such as tray keys, grape varieties, sugar content, and reflectance spectra. Sugar content ranges from 100 to 300 g/L across varie-ties. Grape sorting was performed using NaCl densimetric baths, followed by hyperspectral acquisition. A total of 274 reflectance spectra were obtained, covering red (Syrah, Fer Servadou) and white (Mauzac) grape varieties.

4.2.6 Mango fruits

The dataset in [53] comprises 186 NIR spectra (1000-2500 nm, log(1/R) absorbance) of intact mangoes from 4 cultivars, acquired using FT-NIR with 64 coadded scans/sample. Raw spectral data in .xls and .unsb formats. Reference data includes vitamin C (mg/100g), soluble solids (°Brix), and total acidity (mg/100g). This dataset enables the development of prediction models for rapid, non-destructive quality evaluation of whole mangoes using NIR spectroscopy.

4.2.7 Enhanced NIR spectra of intact mangoes

This dataset in [54] provides original and enhanced near-infrared (NIR) spectral data (1000-2500 nm, 1557 wavelength variables) of 58 intact Kent mango samples. The spectra were acquired using a Fourier transform NIR spectrometer, with 32 coadded scans per sample. The raw absorbance spectra were enhanced using algorithms like multiplicative scatter correction (MSC), baseline linear correction (BLC), and their combination MSC+BLC. The original and enhanced spectral data in .unsb and .xlsx formats for predicting two key internal quality traits - total acidity (TA) and vitamin C content. Model performances were evaluated against reference TA and vitamin C values measured by standard methods, using metrics such as coefficient of determination (R2), correlation (r), root mean square error (RMSE), and residual predictive deviation (RPD).

4.2.8 Cocoa beans

The dataset in [55] contains NIR absorbance spectra (1000-2500 nm) with 32 co-added scans at 0.2 nm resolution for a total of 72 bulk samples of intact cocoa beans, with each sample amounting to 50g. The spectra data is provided in both .xlsx and .unsb file formats. For each sample, the actual moisture content (%) and fat content (%) were measured using standard laboratory methods like thermogravimetry and Soxhlet extraction, respectively. The measured moisture content ranged from 6.74% to 12.08% with a mean of 9.04%, while the fat content ranged from 35.26% to 45.75% with a mean of 40.32%.

4.2.9 Vis-NIR spectra for sugarcane across multiple spectrometers

This dataset in [56] provides Vis-NIR absorbance spectra and corresponding chemical reference data for 60 sugarcane samples, which were analyzed using 8 different spectrometers. These include one laboratory spectrometer (LabSpec 4) and seven micro-spectrometers (NIRscan Nano, F750, MicroNIR1700, MicroNIR2200, NIRONE 2.2, SCIO, TellSpec). The spectral ranges covered span from 350-2500 nm for the LabSpec 4, while the micro-spectrometers capture narrower ranges, such as 1750-2150 nm for the NIRONE 2.2 device. The reference chemical data encompasses total sugar content (ranging from 1.1-51% dry matter), crude protein content (0.9-9.6%), acid detergent fiber (26-59.3%), and in-vitro organic matter digestibility (13-66.6%). This open-access dataset facilitates comparing prediction performance across the various spectrometers employed.

4.2.10 Enhanced Vis/NIR spectral dataset of intact Cucurbitaceae fruits

The dataset in [57] comprises Vis/NIR absorbance spectra (381-1065 nm) of 300 samples from 6 Cucurbitaceae fruit types, including zucchini, bitter gourd, ridge gourd, melon, chayote, and cucumber. The spectra were acquired using a NirVana AG410 portable spectrometer, with each sample scanned 6 times. The data is provided in .xls and .unsb formats. Reference data on soluble solids and water content were determined by standard wet chemistry methods.

Discussion: Development insights

Analysis of existing public NIR datasets in the food industry reveals a clear trend: prioritizing nutritional indices and product quality. This trend reflects economic development goals through enhancing nutritional value and optimizing production processes. However, a notable gap exists - the relative scarcity of data related to food safety factors, particularly in identifying chemical toxins and other hazards. Based on the above survey, to the best of our knowledge, there are currently no relevant public NIR datasets. This could be due to some factors: the high cost of preparing absolutely safe samples for reference, the cost of chemicals for testing and measuring unsafe levels, the sensitivity of data related to food contamination, and the technical challenges of detecting low concentrations of substances using NIR methods.

The first approach to closing this gap is to promote the creation, development, and sharing of NIR spectral datasets through active partnerships with food safety regulatory bodies. Through one project supported by the People's Committee, we are currently collecting NIR spectral data in collaboration with regional food safety authorities to build comprehensive datasets covering 19 common chemical residues found in daily food products, following Ministry of Health standards. This NIR data is validated through independent testing using traditional chemical reference methods. This data collection strategy aligns with regulatory requirements and enables standardized datasets for market surveillance while advancing ML applications in food quality.

The second approach is to improve small food safety datasets, creating richer, larger-scale, and more valuable datasets. Methods such as data enrichment with Generative Adversarial Networks (GANs) and spectral diffusion models that can generate synthetic spectra are untapped potential directions. This approach can increase both the quantity and quality of available data, partly addressing the difficulty of high cost in collecting spectral data related to chemicals and chemical residues in food. This will support the

development of comprehensive, rapid, and non-destructive analytical methods, bringing double benefits: economic development in parallel with consumer health protection.
In these recent datasets containing over 1000 spectra, techniques are mainly in traditional ML, such as PLS for [50] in [58], PLS-DA, SVM, ANN or combine methods Random Subspace Discriminant Ensemble (RSDE) for [49] in [59]. In [59], RSDE demonstrated superior performance with over 95% classification accuracy. Its innovative ensemble architecture combines multiple submodels through random subspace projection and majority voting, delivering enhanced accuracy and reliability while inherently reducing noise sensitivity and overfitting risks. The trend across studies shows a shift from single methods to ensemble and hybrid approaches for handling complex spectral data. Another instance, a milk composition study using NIR transmittance spectra [45] showed that combining SO-PLS with appropriate preprocessing improved prediction accuracy by 5-25%. While these studies demonstrated the effectiveness of hybrid approaches, they examined different food products under varying conditions, making direct comparisons challenging. Notably, DL approaches were not extensively explored, suggesting potential opportunities for investigating modern architectures like CNNs and transformer models that have proven effective in other computer vision and spectral analysis applications. Further research exploring traditional and DL techniques across standard datasets would be valuable to establish their relative effectiveness and generalizability.

4.3 Pre-processing and wavelength selection

While deep learning has developed increasingly, traditional ML models still dominate in NIR tasks. In this context, data preprocessing and feature selection are two commonly exploited factors, improving the quality of input data, reducing data dimensionality, extracting important information from NIR spectra, and enhancing the performance of models. This review focuses on the recent trends and notable developments in data preprocessing and feature selection for NIR spectral analysis in the last few years.
The identification of recent advances in preprocessing and wavelength selection methodologies was conducted systematically and compared with established baseline methods (as in Figure 4).

4.3.1 Pre-processing

Pre-processing of NIR data is important to improve model performance. These common techniques have been shown in [25, 44], mainly including noise reduction, baseline correction, resolution enhancement, centering, smoothing, derivative, de-trending, and scaling methods. Most studies apply only one or two pre-processing methods, chosen based on experience rather than a systematic assessment of optimal methods. This is a highly complex process that requires expert knowledge to select the most appropriate pre-processing method for each specific dataset. Choosing the wrong pre-processing method can lead to the loss of important information or add more unwanted noise, thereby negatively affecting model performance. Therefore, developing a systematic approach to evaluate and select optimal pre-processing techniques for NIR spectrum data is an important direction for future research.

Recently, the trend in NIR spectral data pre-processing has seen significant advancements, particularly towards automation and optimization of procedures. Methods such as Synergy Adaptive Moving window algorithm based on the Immune Support Vector Machine (SA-MW-ISVM) [62], Automatically generating pre-processing strategy (AgoES) [65], and Sequential preprocessing through ORThogonalization (SPORT) [66] have been developed to automate the selection and combination of optimal pre-processing methods, thereby reducing manual intervention and experimentation time. Although the Self-expansion Full Information Optimization strategy (SFIOS) [57] is not entirely automatic in pre-processing selection, it nevertheless provides a comprehensive optimization strategy that includes preprocessing, as in Figure 5. Simultaneously, the trend of combining multiple pre-processing methods, as in Table 3, such as Savitzky-Golay (SG) with Standard Normal Variate (SNV), has become popular to leverage the advantages of each method. Researchers in [65] developed strategies to find optimal pre-processing pipelines, evaluating 150 different combinations of 16 common techniques. Similarly, another study [64] explored an automated method to combine up to 4 out of 9 pre-processing types for NIR data of coconut milk. Using various models (Multi-Layer Perceptron (MLP), k-Nearest Neighbors (KNN), and Partial Least Squares (PLS)), they achieved the best results with KNN on Micro-NIR data, obtaining a classification accuracy of 0.97 and a regression RPD of 16.108. Both studies [64, 65] emphasize the importance of automated optimization in pre-processing pipelines, as no single method consistently outperforms others across all scenarios.
Notably, there is a close integration between pre-processing and machine learning algorithms, as demonstrated in SA-MW-ISVM, where both pre-processing and SVM parameters are optimized simultaneously. Similarly, ensemble learning techniques are widely applied, as in AGoES and SPORT, to combine various pre-processing models. Additionally, many methods focus on selecting relevant spectral variables, thus contributing to dimensionality reduction and improving model performance.
Not limited to a single data type, new methods have been developed with adaptability for various data types, including both solid and liquid data, as well as spectral types beyond NIR. Moreover, the trend of sharing open-source preprocessing algorithms is increasing, thereby enabling the research community to continue developing and improving these methods. Interestingly, while many methods focus on large datasets, some approaches like Extended Multiplicative Signal Augmentation (EMSA) emphasize improving performance on smaller datasets, meeting the needs of

Table 3: Preprocessing methods and their applications

Preprocessing methods	Single	Combine	Remarks
SG smoothing, MSC, SNV, 1st Der, 2nd Der [60]	x	Multiple combinations of SG with other methods	SG alone or combined with SNV gave the best results for most models; Improved classification accuracies compared to raw spectra; SG+SNV optimal for predicting most physicochemical properties; Preprocessing crucial for developing robust NIR models for melon seed powder authentication
SM, DF, NM, CT, DE [61]	x	Multiple schemes from multiple methods	SFIOS auto-optimization strategy combines methods, and provides statistical info; en-iViSSA ensemble improves model performance; DF highlights spectral info; CT and DE good for solid samples; MSC and SNV minimize scattering effects; Open source
SG smoothing, SG derivatives, MSC, SNV, Autoscale, Normalization [62]	x	Multiple schemes from 6 methods	SA-MW-ISVM algorithm optimizes preprocessing, wavelength selection, and SVM parameters simultaneously; Improves prediction accuracy by up to 44% compared to PLS; Selects relevant wavelengths, reducing variables to ∼30% of full spectrum; Chooses appropriate preprocessing combinations; Applicable to NIR and other spectroscopy data
SG, EMSC, EMSA, Batch Aug [63]	x	Multiple schemes from 4 methods	EMSA can replace pre-processing for CNNs; Combination of pre-processing and augmentation improves results for conventional classifiers on small datasets; CNNs benefit from traditional pre-processing; Augmentation especially beneficial for small datasets
BSO3, 1st Der, 2nd Der, SNV, MSC, MS, SG filter, SS [64]	x	Multiple schemes from 9 methods	Automatic strategy combines preprocessing and ML hyperparameter tuning; Improves classification (up to 98% accuracy) and regression (RPD up to 16) for coconut milk adulteration detection using FT-NIR and Micro-NIR
Baseline correction, Scatter correction, Scaling [65]	x	150 combinations from 16 single methods	AGoES automatically generates and evaluates all preprocessing combinations; Different optimal combinations are found for each ML algorithm and property predicted; Improved model performance compared to raw spectra for most cases; SVM with AGoES preprocessing performed best overall
SG smoothing/derivatives, SNV, VSN [66]	x	Multiple combinations via SO-PLS	SPORT method automatically combines and selects optimal preprocessing sequence; Outperformed single preprocessing and stacking approaches; Selected parsimonious combinations of 2-3 methods; Order of combination had minor impact on performance

many practical applications.

Discussion: Development insights

Effective preprocessing depends on the type of data. However, the variety of handheld measurement devices available today, as discussed in Section 4.1, and the rapid development of deep learning raise questions about the future role and necessity of data preprocessing. Possible research directions include: (1) more flexible preprocessing automation, adapting to different types of data, serving complex multi-class classification problems in practice, such as classifying unsafe fruits and vegetables in food safety inspection, (2) developing preprocessing methods suitable for many different measuring devices, aiming at a model that can transfer technology between localities with asynchronous devices, and (3) evaluating if the need for preprocessing, with fluctuating data in different measuring environments, such as markets and supermarkets, where humidity, light, and temperature can all affect measurement values, thereby affecting the effectiveness of machine learning models.

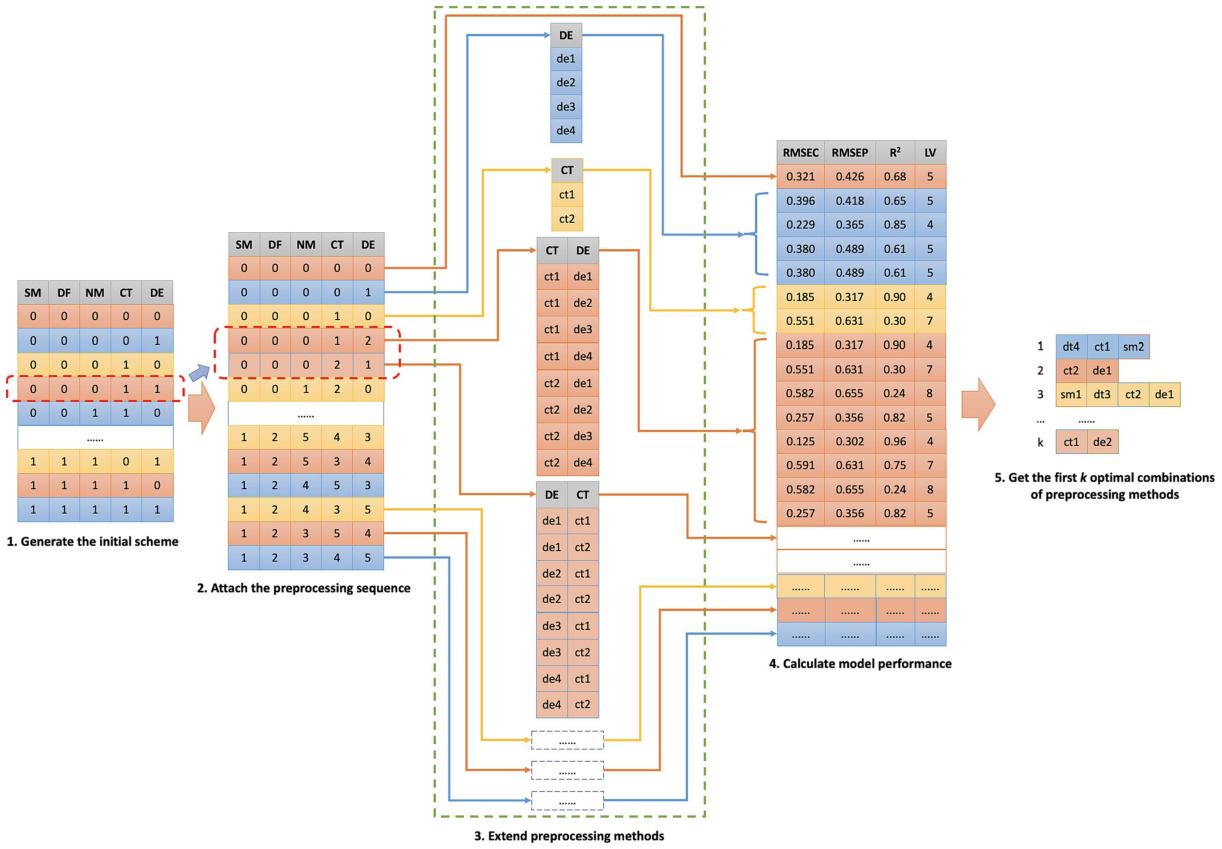

Figure 5: The optimal preprocessing scheme of SFIOS [61]

4.3.2 Wavelength selection

Wavelength selection is critical for NIR analysis. By identifying the most informative spectral regions, it simplifies complex data, improves model performance, and focuses on key information, as in Table 4, 5. These methods are categorized by their approach and can be broadly grouped as follows:

1. Filter methods utilize statistical criteria to evaluate the relevance of individual wavelengths to the target variable, including Variable Importance in Projection (VIP), Correlation-based Feature Selection (CFS), Relief, Fisher score, and Chi-squared test;

2. Wrapper methods assess the quality of wavelength subsets based on the performance of a prediction model, such as Genetic Algorithm (GA), Successive Projections Algorithm (SPA), Recursive Feature Elimination (RFE), Sequential Forward Selection (SFS), and Sequential Backward Selection (SBS);

3. Embedded methods integrate feature selection within the model-building process, including Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net, Ridge Regression, Random Forest, and Competitive Adaptive Reweighted Sampling (CARS);

4. New hybrid methods combine the strengths of different approaches, such as Multi-Feature Extraction combined with LASSO (MFE-LASSO), Maximal information coefficient - Successive projections algorithm combined with Extreme learning machine - Genetic algorithm (MIC-SPA-GA-ELM), and Beluga whale optimization with Iterative variable subset optimization (BWO-IVSO).

Each method has its strengths and limitations, making it suitable for specific data types and analytical objectives. Filter methods efficiently screen large variable spaces to rank features. For instance, VIP was used to select 13 key wavelengths out of 209 initial variables for adulterant detection in quinoa flour [67], improving the R_p^2 from 0.94 to 0.98 and reducing RMSEP from 3.04% to 1.60%. Wrapper methods, such as GA and SPA, directly optimize subsets based on model performance. In a study on durian fruit quality assessment [19], GA selected 23 wavelengths for dry matter prediction and 19 for total soluble solids, improving the model's accuracy. Hybrid methods combine filter and wrapper approaches for robust performance, as seen in the MIC-SPA-GA-ELM [68] combination used for tobacco and corn samples, which showed the best accuracy and robustness. However, most techniques retain spectroscopic variables related to key functional groups and structural chemistry to develop broadly applicable, physically

interpretable models, as in Table 5.

Wavelength selection can also be divided into interval or peak selection, as in Figure 6. Peak selection involves choosing specific, individual wavelengths that are most informative. For example, four peaks (1428, 1704, 1892, 1912 nm) were identified as crucial in a vineyard water status prediction study [69]. On the other hand, interval selection chooses continuous ranges of wavelengths. In the same vineyard study, three intervals (1402-1508, 1676-1750, 1870-1926 nm) were selected. This approach can be particularly useful when certain regions of the spectrum are known to be associated with specific molecular structures or properties of interest.

The current trend is to combine multiple variable selection methods to optimize results. For instance, the BWO-IVSO approach applied in aflatoxin B1 analysis in peanuts significantly improved model performance compared to using the full spectrum [72]. Another example is the two-step approach using RReliefF and MIC, followed by Elastic Net, for azodicarbonamide detection in wheat flour [74]. Besides, recent studies also emphasize the importance of optimizing model parameters. Algorithms like Harris Hawks Optimization (HHO) and Rime Optimization Algorithm (RIME) have been used to fine-tune parameters of models such as Kernel-based Extreme Learning Machine (KELM), significantly improving model performance [78].

The number of selected wavelengths or spectral regions typically varies based on the complexity of the sample and the specific analytical objectives. For instance, 18 wavelengths were selected for sugar in tobacco leaves, while 24, 34, 26, and 16 wavelengths were chosen for moisture, oil, protein, and starch in corn, respectively [68]. In some cases, a few carefully selected wavelengths can provide comparable or even superior results to models using the full spectrum, while significantly reducing computational requirements. ***Discussion: Development insights***

Wavelength selection is a popular method in NIR spectral processing. This is based on the interpretability of the results, which are related to the functional groups representing the sample. Although this method is effective, as with preprocessing, whether wavelength selection is necessary or using the raw data itself with deep learning architectures is more effective needs to be further compared.

Combining both methods is also a possible direction for development. This method takes advantage of both the nonlinear learning capabilities of the neural network and maintains the interpretability of the final model.

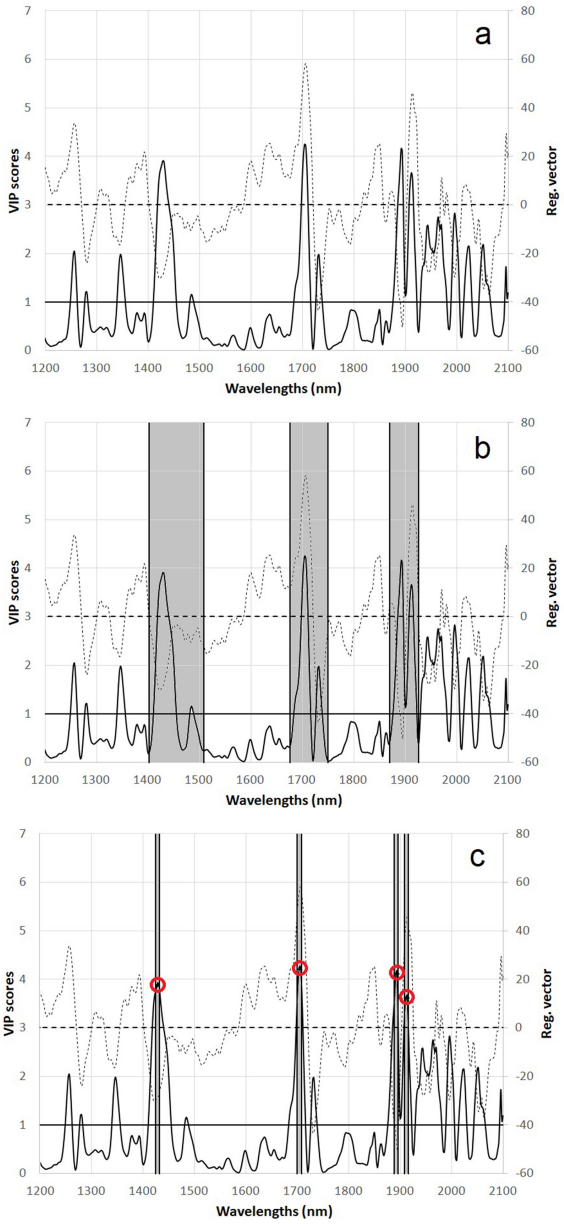

Figure 6: Wavelength selection methods comparison: (a) Manual VIP-based selection after preprocessing, (b) Interval selection approach, and (c) Peak selection method highlighting key wavelengths and bandwidths [69]

4.4 Advanced NIR food spectral analysis techniques

4.4.1 Traditional machine learning

Traditional ML methods play a crucial role in NIR spectral analysis, focusing on addressing multicollinearity issues and improving generalization capabilities. In a typical pipeline, these techniques involve data pre-processing, feature selection/ extraction, and applying traditional ML algorithms to model the selected features and generate outputs. Popular techniques include PCA, PLS, ELM, SVM, SVR, DT, and RF. These methods concentrate on extracting important features, minimizing data redundancy, and building effective predictive models for various NIR applications. The recent highlights of traditional ML on NIR spectra largely lie in the improvements in pre-processing strategies as well as wavelength selection, and the effective feature extraction methods mentioned earlier.

However, traditional methods often face limitations in

Table 4: Wavelength selection methods for NIR spectra (1)

Task	Dataset	Variable Selection Methods	Selected Wavelengths	Results	Remarks
Quinoa flour adulteration [67]	54 samples, 941-1674 nm, 209 vars	VIP	13 wavelengths	Initial (209 vars): R_p^2=0.94, RMSEP=3.04%. Selected (13 vars): R_p^2=0.98, RMSEP=1.60%	PLSR model improved for quinoa flour adulteration prediction
Vineyard water status [69]	288 samples, 1200-2100 nm, 501 vars	Interval, Peak, IPLS	3 intervals: 1402-1508, 1676-1750, 1870-1926 nm. 4 peaks: 1428, 1704, 1892, 1912 nm	Initial (501 vars): R_p^2=0.84, RMSEP=0.167 MPa. Selected (9-33 vars): R_p^2=0.77-0.78, RMSEP=0.186-0.201 MPa	3 methods for key wavelengths, simplified model, comparable accuracy
Durian quality [19]	278 samples, 860-1750 nm	SPA, GA, VIP	GA: DM 23, TSS 19 wavelengths	Full: R^2=0.83, RMSEP=4.96% (DM); R^2=0.81, RMSEP=3.71% (TSS). Selected: R^2=0.85, RMSEP=4.50% (DM); R^2=0.66, RMSEP=5.15% (TSS). Accuracy: 94.20%	3 methods improved DM and TSS prediction, GA best
Robust NIR model [68]	Corn: 80, 1100-2498 nm	LARS, CARS, SPA, UVE, MIC-combined	24 (moisture), 34 (oil), 26 (protein), 16 (starch)	MIC-SPA-GA-ELM: Best accuracy, robustness	Combined methods improve model accuracy, stability
S-ovalbumin in eggs [70]	150 samples, 900-1700 nm, 390 vars	SPA, IRIV	SPA: 16, IRIV: 14	PLSR (16): R_c^2=0.90, RMSECV=8.92%, R_p^2=0.84, RMSEP=9.98%. PLSR (14): R_c^2=0.91, RMSECV=8.44%, R_p^2=0.86, RMSEP=9.33%	IRIV-selected (14) PLSR best for S-ovalbumin prediction
Adaptive PLS [71]	Corn: 80, 2498-1100 nm. Wine: 44, 900-5000 cm^{-1}	CARS, MCUVE, LARS, ABUSE	Corn: 4 regions. Wine: 4 wavelengths	ABUSE PLS: Corn (25): RMSECV=0.003, RMSEP=0.004. Corn (26): RMSECV=0.009, RMSEP=0.013. Wine (4): RMSECV=5.12, RMSEP=4.66. Wine (3): RMSECV=5.13, RMSEP=4.63	ABUSE selects key peaks, best performance, fewer variables, improved accuracy, reduced time
Aflatoxin B1 in peanuts [72]	100 samples, 955-1702 nm, 128 vars	IVSO, BWO-IVSO	IVSO: 32, BWO-IVSO: 18	SVM (full): RMSEP=31.4602, R_p=0.9608, RPD=3.6799. SVM (IVSO): RMSEP=30.4587, R_p=0.9633, RPD=3.8009. SVM (BWO-IVSO): RMSEP=24.6322, R_p=0.9761, RPD=4.6999	BWO-IVSO removes redundancy, noise, enhances AFB1 analysis accuracy

learning complex and nonlinear features when dealing with complex and high-dimensional NIR spectral data. This leads to the need for deep learning approaches, which can automatically extract complex features and efficiently process high-dimensional data, thereby improving accuracy and generalization capabilities in many NIR spectral analysis tasks.

Table 5: Wavelength selection methods for NIR spectra (2)

Task	**Dataset**	**Variable Selection Methods**	**Selected Wavelengths**	**Results**	**Remarks**
Dual-sPLS NIR [73]	Rice: 447, 12481-3595 cm^{-1}, 1153 vars	PLS, iPLS, SiPLS, mw-PLS	149 optimal vars	SiPLS (149 vars): RMSEP reduced 0.2284 to 0.1952	Variable selection improves model, saves computation time
ADA in wheat flour [74]	101 samples, 0-300 mg/kg, 7012 vars	RReliefF, MIC, EN	Two-step: 500, then 40	PLSR (7012): RMSEP=2.53%, r=0.975. PLSR (500 MIC): RMSEP=1.32%, r=0.992. PLSR (40 MIC+EN): RMSEP=0.78%, r=0.997	MIC+EN eliminates irrelevant vars, retains key info, improves accuracy
Talcum in wheat flour [75]	123 samples, 1050 vars	EN, GA	EN+GA: 55	EN+GA (55/1050): GBDT: R^2=0.9778, RMSEP=0.8905, RPD=6.8099	Detects low talcum concentrations in wheat flour
GBM-PLS for corn [76]	120 samples, 7 countries, 867-2535 nm, 949 vars	RC, CARS, XGBoost, LightGBM, CatBoost	Moisture: 6 (CatBoost). Protein: 6 (LightGBM)	Best: Moisture - R^2_V=0.97, RMSEV=0.45%, RPDV=6.20. Protein - R^2_V=0.82, RMSEV=0.51%, RPDV=2.41	GBMs good for wavelength selection. CatBoost best for moisture, LightGBM for protein. SHAP identified key wavelengths
Strawberry SSC [77]	630 samples. Reflectance: 600-1080 nm (949). Transmittance: 600-950 nm (805)	SI, SPA, UVE, CARS	Transmittance (CARS): 33	Best (Transmittance CARS-PLS): R_p=0.928, RMSEP=0.412 °Brix, RPD=2.670	Transmittance with CARS best. 3 strawberries/sec. More research needed
Zearalenone in wheat (CSA-NIR) [78]	131 samples, 901-1701 nm, 228 vars	VCPA, BOSS, CARS	CARS: 107 (best)	Best (CARS-RIME-KELM): R^2_p=0.9900, RMSEP=18.4610 μg/kg	CARS best. RIME improved KELM. CSA-NIR effective for zearalenone detection
Zearalenone in wheat (FT-NIR) [79]	116 samples, 10,000-4000 cm^{-1}, 3112 vars	CARS, SVM-RFE, MFE-LASSO	MFE-LASSO: 38 (best)	Best (MFE-LASSO-PLS): R^2_p=0.9545, RMSEP=18.6442 μg/kg, RPD=4.3198	MFE-LASSO best. FT-NIR+MFE-LASSO-PLS effective for zearalenone detection

4.4.2 Deep learning architecture

In NIRS analysis, DL architectures have demonstrated great potential in enhancing the efficiency and accuracy of analytical processes. Each architecture possesses unique operational mechanisms and advantages suitable for different challenges in NIR analysis, as shown in [25]. Stacked Autoencoders (SAE) excel at learning low-dimensional representations of input data, effectively reducing noise and focusing on essential information. Variational Autoencoders (VAE) function as generative models capable of producing new samples, proving valuable for data augmentation. CNN are adept at learning local features of NIR data, while RNN process NIR data as time series, capturing sequential relationships. ELM show high efficiency in scenarios with limited training samples, and GAN can generate new training data, addressing data scarcity issues. A thorough understanding of the mechanisms and advantages of each architecture enables researchers to select the most appropriate method for specific applications in NIR analy-

sis, ultimately leading to more accurate and reliable results in this crucial field of spectroscopy. In this study, we focus only on recent notable DL research and the salient value points not discussed in previous review articles. In particular, papers that have comparisons with traditional ML are prioritized, as in Table 6, 7, 8.

Firstly, DL models achieve superior performance over traditional ML methods in NIRS analysis through their innovative architectures such as automatic hierarchical feature extraction, attention mechanisms, deep temporal learning via recurrent networks, and intelligent feature fusion through dense connections, enabling more accurate classification, regression, and anomaly detection tasks. Previously, many DL studies were done without comparison with traditional methods that have achieved high performance. This has received more attention recently. These results confirm the potential of DL in enhancing performance for regression, classification, and anomaly detection tasks in NIR spectral analysis, opening up possibilities for wide-ranging applications across fields requiring high accuracy and reliability.

In classification tasks, for instance, Convolutional Neural Network-Attention (CNN-ATT) achieved 100% accuracy in categorizing chickpeas into HTC and ETC classes [84], surpassing traditional SVM models. In this study, CNN-ATT enhances performance through its attention block mechanism that dynamically weighs and focuses on the most relevant features in input data, allowing the network to adaptively prioritize important spectral information while filtering out less significant signals. Similarly, the Transformer in [92] achieves remarkable classification accuracy (99.31%) through its three-layer encoder and multi-head attention mechanism that effectively extracts semantic information from both vibration and Vis/NIR spectral data while dynamically adjusting feature weights via an attention feature fusion module to focus on the most relevant information for apple moldy core detection, better than PLS-DA, SVM, ELM. In another study, for corn variety recognition [87], CNN reached 99.2% accuracy, outperforming traditional methods like KNN, SVM, and PLS by 25.78%. In a more complex scenario of identifying adulterated beef and mutton [85], ResNet with 2DCOS, as in Figrure 7, achieved 100% accuracy, significantly surpassing PLS-DA's 32-50% accuracy. A total of 1,878 synchronous and asynchronous 2D-COS spectra were obtained from transforming 1D spectra across 23 diverse adulteration patterns (5 pure meats and 18 mixed samples with varying proportions of 25%, 50%, and 75%) into 2D images to enhance resolution and analytical sensitivity while providing multi-dimensional information through auto and cross-correlation peaks, enabling accurate detection of both components and mixing ratios in meat samples, with characteristic markers at different wavelengths. Besides, ResNet achieves superior performance through its innovative skip connection technique that allows direct data flow between layers, effectively preventing gradient vanishing and enabling faster, more accurate training than traditional CNNs when handling this complex multi-class classification problem.

For regression tasks, DL models consistently showed superior performance. For instance, in predicting lead content in oilseed rape [81], the Transfer Stacked Auto-encoder (T-SAE) model achieved R^2 values of 0.9215 and 0.9349 for leaves and roots respectively, outperforming PCA-SVM and SAE. A highlight of this research is that T-SAE achieves high performance through its dual-model transfer mechanism, where network weights are initialized from pre-trained SAE models while allowing deep feature layers to be trained from scratch with random weights, effectively combining information from both leaf and root spectral data to achieve superior classification accuracy (98.75%) in lead stress detection. In another study, for estimating soluble solids content in pears [83], SpectraNet-32 achieved the best results, surpassing classical methods like PLS, MLR, and SVM. In predicting cooking time for chickpeas [84], 1D-CNN also outperformed traditional regression methods. Regarding anomaly detection and complex analysis, DL models also excelled. For pesticide residue recognition on garlic chive leaves [90], 1D CNN achieved 97.9% accuracy, outperforming traditional models. In analyzing complex organic compounds [89], the proposed DL model achieved R^2 values between 0.9574 and 0.9996, improving upon PLSR and BPNN by significant margins. This architecture achieves superior dynamic feature extraction through its innovative dual-module design, as in Figure 8, where the short-term feature extraction utilizes multi-rate dilated convolutions with dense connections to capture short-term spectral patterns while the long-term feature extraction employs Gated Recurrent Unit (GRU) enhanced by temporal attention mechanism to comprehensively merge features across all timesteps, complemented by a linear bypass path and two-stage quality regression approach that effectively prevents overfitting in NIR spectral analysis.

Thus, the outstanding advantages of DL architecture, such as automatic feature extraction, attention mechanisms, multi-scale temporal feature learning through dilated convolutions, comprehensive time-series analysis via GRU networks, and intelligent fusion through dense connections, have led to superior performance compared to traditional ML approaches.

Second highlight, DL architectures demonstrate high flexibility, successfully applied to various data types ranging from 1D spectra to 2D images, 2D correlation spectra (2D-COS), and even 2D dynamic data. These studies highlight the versatility of DL architectures in NIRS, effectively processing various data formats from simple to complex data, and even enabling advanced techniques like transfer learning across different devices. This flexibility opens up the potential for developing architectures in computer vision on NIR spectral data transformed into image data formats.

There are many methods to convert 1D NIR spectra into 2D. First of all, the 2D-COS technique [82, 91], as in Figure 9 transforms one-dimensional NIR spectra into two-dimensional correlation spectra (synchronous and asynchronous) by calculating cross-correlations between spec-

Table 6: Advanced DL Architectures for NIR food spectral analysis (1)

NIR Tasks	**Datasets**	**Pre-processing**	**Models**	**DL Architecture**	**Results**
ADF and IVOMD in sugarcane [80]	60 NIR x 3 devices, 600/device, 3:1:1 split	WS, Interpolation, SNV	1D-Inception-ResNet, PLS	8 Conv, 4 FC, Residual, Softplus, Dropout 0.15	ADF/IVOMD: $R^2 > 0.96$, RMSEP < 2.75. *Outperformed PLS, Successful inter-device transfer*
Lead content in oilseed rape [81]	500 samples (leaves/roots), 3:1:1 split (480.46-1001.61 nm)	SNV, 1st Der, 2nd Der, PCA	T-SAE, SAE, SVM, SVR, PCA-SVM	Best T-SAE: 411-148-108-60 (leaves), 410-140-91-56 (roots)	$R^2 = 0.9215$, RMSEP = 0.0302 mg/kg (leaves); $R^2 = 0.9349$, RMSEP = 0.0278 mg/kg (roots) *Outperformed PCA-SVM, SAE; successful transfer learning*
Total phenolic content in boletes [82]	187 samples (3 species), 90% model, 10% valid	SNV, HCA, Folin-Ciocalteu	ResNet, 2D-COS, SVM, PLS-DA	12-layer ResNet, identity & conv blocks, BatchNorm, ReLU	100% accuracy (train & test). *Outperformed traditional methods, rapid & non-destructive*
SSC and temp in "Rocha" pear [83]	3300 spectra (1650 pears), 499.73-1101.83 nm, 5 valid sets	QNV, Savitzky-Golay (1st & 2nd), PLS wrapper (BVE-PLS), PLS-VIP	SpectraNet-32/53, DeepSpectra, PLS, MLR, SVM, MLP	SpectraNet-32: 32-layer ResNet, 3 Residual Units, BatchNorm, GELU, Global Avg Pooling, Dropout	Best: RMSEP = 1.08%, R^2 = 0.58 (SSC). *Outperformed classical methods, predicted SSC & temperature, 8000 spectra/s*
Chickpea HTC/ETC classification, Cooking time prediction [84]	864 seeds (8 varieties), 900-2500 nm	SNV, 1st and 2nd derivatives; CARS, IRIV, CNN-FS	PLSDA, SVC, CNN-ATT, 1D-CNN	CNN-ATT: ATT block, 3 1D conv, 2 dense, SoftMax. 1D-CNN: 1 input, 1 conv, 4 dense, 1 output	SVC & CNN-ATT: 100% acc (full spectrum). 1D-CNN: $R_p^2 = 0.880$, RMSEP = 0.662 (cooking time) *Non-destructive, rapid detection. Effective for both classification and prediction*
Adulterated beef/mutton ID [85]	1878 samples, 0-100% adulteration, 400-2500 nm	Raw, FD, SD, MSC-SG	PLS-DA, ResNet with 2DCOS	ResNet: 12 hidden layers, ReLU, global avg pooling	ResNet+2DCOS: 100% acc. PLS-DA: 32-50% acc. *2DCOS enhances spectral resolution. ResNet extracts 2DCOS features effectively.*
Subsurface bruises in plums [86]	1125 HSI, 430-1000 nm	Standardization, Data augmentation; PCA (10 wavelengths)	HSCNN, ResNet, 3D-CNN, PLS-DA	HSCNN. ResNet: Adapted for HSI. 3D-CNN: 3 conv, 3 max-pool, 3 batch norm, global avg pool, 2 dense	Best: HSCNN (full spectrum), F1 90%. 3 wavelengths: F1 89%. *Detected invisible bruises. Reduced to 3 wavelengths with similar performance*

Table 7: Advanced DL Architectures for NIR food spectral analysis (2)

NIR Tasks	Datasets	Pre-processing	Models	DL Architecture	Results
Corn variety recognition [87]	450 NIR (5 varieties), 2:1 split, 11542-3940 cm^{-1}	DT for baseline drift; CARS (114/1845 wavelengths)	CNN-LeNet-5, BP, KNN, SVM, PLS	3 Conv: 32, 64, 128 kernels (7, 4, 4 windows). 3 Pool: Max & Glob. Avg. ReLU, Dropout 0.1, FC, softmax	CNN: 99.2% test acc. *Combining NIR, CNN enables accurate, rapid recognition; 25.78% higher accuracy than traditional models*
Watermelon SSC [88]	1440 Vis-NIR (317-1117 nm), 60:30:10 split	PLSR: SG+2nd der. BPNN: MSC	PLSR, BPNN, 1D-CNN	1D-CNN: 5 Conv1D, 3 MaxPool, Flatten, 3 Dense, BatchNorm, Dropout 0.1	1D-CNN: $R_p^2 = 0.97$, RMSEP=0.21. +14.1% vs PLSR, +6.6% vs BPNN *High R_p^2, Low RMSEP. Features at 720, 810nm*
Quality of complex organics [89]	2D NIR, 1408 spectra (12500-3950 cm^{-1}), 844:432:564 windows	SG (NIR). SG+2nd der for PLSR. MSC for BPNN	PLSR, BPNN, Proposed DL	MDFE: SDFE (3 dilated 2D CNN) + LDFE (GRU with attention). Regression: FC (1024, 500), ReLU, BatchNorm, Dropout 0.1	DL: R^2=0.9574-0.9996, RMSE=0.0013-0.4374. +14.1% vs PLSR, +6.6% vs BPNN. *Short/long-term dynamics, Dilated CNN extracts multi-level short-term features; Temporal attention redistributes GRU features; Nonlinear fitting of complex NIR-quality mapping; Visualization shows model extracting relevant spectral bands*
Pesticide on garlic chives [90]	SWIR HSI, 30 leaf spectra, 920-1700nm, 90:5:5 split	Mean filter (SNR), Isolated Forest (outliers)	PLS, MLR, BPNN, KNN, LDA, NB, RF, SVM, 1D CNN	1D CNN: 3 conv (3x1, stride 2x1, pad 1), avg/max pool, flatten, 2 FC (256, 128), 4-node output, ReLU, BatchNorm	97.9% test acc. Recall $>$ 97.7%, AUC $>$ 0.99. 0.208 Hamming loss (mixed). vs KNN (91.2%), LDA (77.5%), NB (41.4%), RF (90.9%), SVM (92.8%).*Non-destructive, Rapid, Outperformed traditional models, Exploiting pixel-wise spectra for a large dataset; Successful mixed residue identification*

tral intensities $\tilde{y}(v_1)$ and $\tilde{y}(v_2)$ at wavelengths v_1 and v_2 as external perturbations (like geographical origins, storage time) change. The synchronous spectrum reveals peaks that change in the same direction through auto-correlation peaks (diagonal) and cross-correlation peaks (off-diagonal), while the asynchronous spectrum only shows cross-peaks indicating spectral changes from different molecular sources. This technique is particularly valuable for NIR spectral analysis when spectral peaks overlap due combinations, and complex molecular interactions in food samples, as it improves spectral resolution and helps distinguish overlapping features that are not observable in one-dimensional spectra. Unlike 2D-COS method which focuses on molecular interactions through correlation analysis, the Gramian Angular Difference Field (GADF) [93] converts one-dimensional NIRS into two-dimensional images by first normalizing

Table 8: Advanced DL Architectures for NIR food spectral analysis (3)

NIR Tasks	Datasets	Pre-processing	Models	DL Architecture	Results
Wolfberry origin identification [91]	NIR-HSI (900-1700nm, 256 bands), 700 samples from 5 regions, 525:175 split	2D-COS to resolve overlapping peaks. CARS/IRIV/iVISSA for wavelength selection. GLCM for texture features	LDA, PLS-DA, SVM, CNN	CNN: 3 conv layers (16,32,64 filters), 3×3 kernel, ReLU, BatchNorm, MaxPool(2), GAP, FC(128), Dropout(0.2), Sigmoid	CNN+iVISSA: Acc=96.67%. CNN+texture: 97.71%. +9.71% vs PLS-DA, +7.42% vs SVM. *2D-COS resolves peaks, iVISSA selects wavelengths, Texture features improve accuracy*
Apple moldy core detection [92]	Vibration signals (100-1500Hz) + Vis/NIR (350-1150nm), 725 samples (180 normal + 545 diseased) split 3:1:1	CEEMDAN, Vibration + Vis/NIR fusion	PLS-DA, SVM, ELM, MobileNet, DMLPT	DMLPT: 3-layer Transformer Encoder for each input, AFF for fusion, MLP with residual connection. Multi-head attention	DMLPT+fusion: Acc=99.31% (normal/moderate/severe: 100%). +11.03% vs PLS-DA, +5.52% vs MobileNet. *Multi-modal fusion and multi-head attention excel at severity detection*
Apple SSC prediction [93]	Vis/NIR transmission spectra (589-1120nm, 1468 bands), 1450 spectra from 290 apples, 1020:430 split	GADF, SNV-UVE	PLS, MLR, VGG16, ResNet50, ShuffleNetv2, MobileViT	MobileViT: Conv + Transformer hybrid, CA mechanism for spatial features, multi-head attention. 3 MobileNet blocks, SiLU activation	GADF-MobileViT: R^2=0.938, RMSE=0.532. +6.6% vs PLS, +2.1% vs ResNet50. *GADF enables 2D transform, CA enhances features, Model focuses on key bands*
Tea quality classification [94]	NIR (1000-1800nm, 800 points), 1000 samples (50 grades, 21 brands), 750:250 split	SNV preprocessing. Transform 1D spectra to 2D pseudo-images (2×20×20)	PLS-DA, SVM, RF, TeaNet variants	TeaNet series: 3 conv/ residual/ inverted blocks, BatchNorm, ReLU, GAP, FC. Feature extraction + classification	TeaResNet+SNV: Acc=100%, TeaMobileNet: 99.6%. +31.2% vs SVM, +2.4% vs RF. *2D transform enables CNN, SNV increases variance, Models excel at multi-category*
Pb detection in oilseed rape [95]	FHSI 480-980nm, 2400 samples (1200 per environment), 3:1 split	SNV preprocessing, T-SCAE for cross-environment transfer	SVR with SPA/ CARS/ IRIV/ VISSA, SCAE, T-SCAE	Pre-trained SCAEs + extended layers (822-423-301-155)	T-SCAE+SNV: R^2=0.939, RMSE=0.020. +6.51% vs traditional. *Transfer learning enables cross-environment prediction*

spectral data to [-1,1], then transforming them into polar coordinates through angular cosine encoding, and finally generating a GADF matrix by calculating the sine differences of angular values between each pair of spectral points, resulting in a matrix that preserves spectral relationships while optimizing for pattern recognition and machine learning applications, as in Figure 10. Different from 2D-COS, which focuses on correlation analysis, and GADF, which uses polar coordinate transformation, TeaNet's approach [94] simply transforms one-dimensional NIR spectra with 800 spectral points into two-dimensional pseudo-images of size 2x20x20 through direct matrix reshaping, requiring no complex mathematical calculations while still preserving all spectral information and enabling spatial rela-

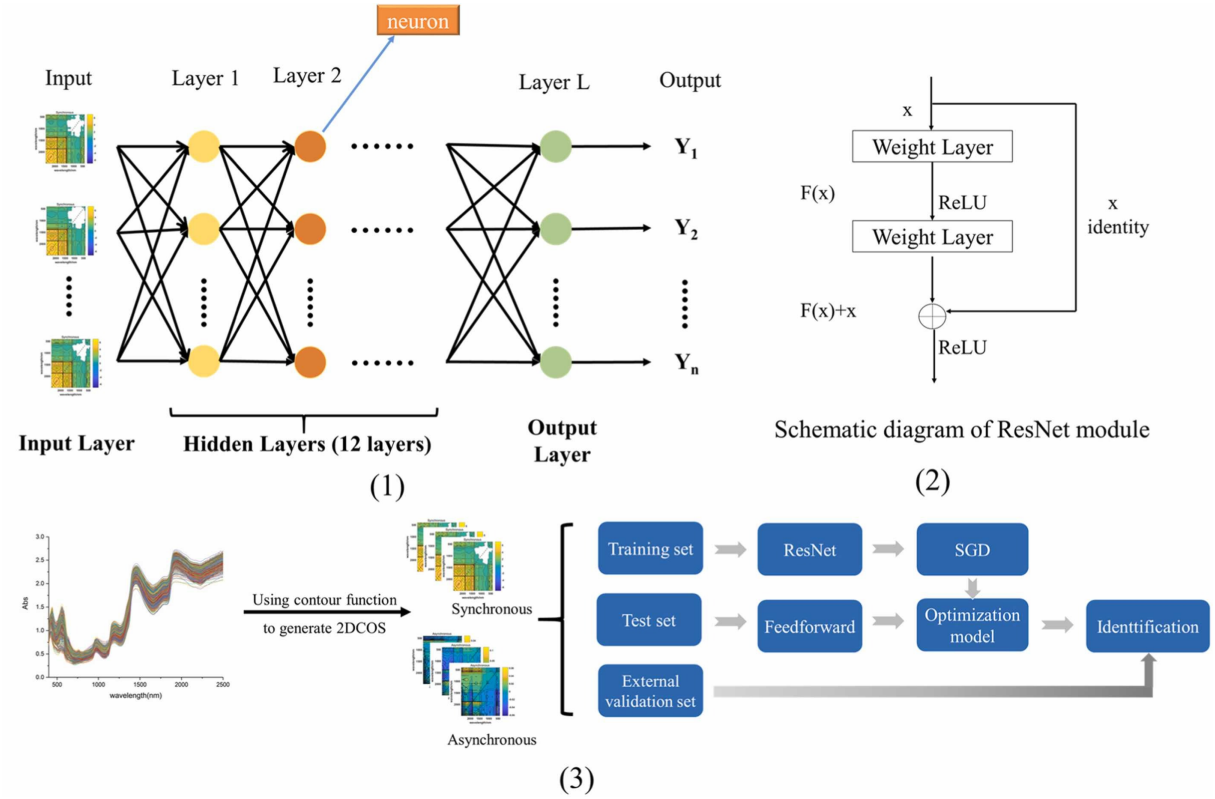

Figure 7: 2D-COS combined with ResNet process for identifying adulterated beef and mutton [85]

tionship analysis through CNN's convolutional operations across spectral bands.

DL demonstrates superior performance across these diverse data types. For 1D spectral data instance, in watermelon soluble solids content analysis [88], a 1D-CNN with five convolutional layers effectively processed 1D Vis-NIR spectra (317 to 1117 nm), achieving superior results compared to traditional methods. Similarly, for corn variety recognition [87], a CNN based on the LeNet-5 architecture successfully handled 1D NIR spectra (11542-3940 cm^{-1}), achieving 99.2% accuracy.

With 2D spectral image data, the study on subsurface bruise detection in plums [86] employed various CNN architectures, including HSCNN and ResNet, to process 2D hyperspectral images (430-1000 nm). These models effectively extracted spatial and spectral features, with HSCNN achieving the best F1 score of 90% using the full spectrum. For 2D dynamic spectral data, in the quality prediction of complex organic compounds [89], a novel DL model was designed to handle 2D NIR dynamic spectral matrices (time x wavenumbers). This model incorporated multi-level dynamic feature extraction, including short-term (using dilated 2D CNN) and long-term (using GRU with temporal attention) feature extraction, effectively capturing both spatial and temporal characteristics of the spectral data. Additionally, for transformed 2D data, in identifying adulterated beef and mutton [85], researchers innovatively transformed 1D spectral data into 2D-COS before feeding it into a ResNet model. This approach achieved 100% accuracy, demonstrating the potential of DL in processing transformed spectral data. Finally, the study on quantitative analysis of ADF and IVOMD in sugarcane [80] showcased the ability of a 1D-Inception-ResNet to handle data from multiple devices, achieving $R^2 > 0.96$ and $RMSEP < 2$ for both devices, demonstrating successful inter-device transfer learning.

The third highlight is the development of hybrid models and specialized architectures that have significantly advanced NIR spectral analysis, leading to improved performance across various tasks. These innovative approaches demonstrate the potential of tailored deep learning solutions in spectroscopy. These examples demonstrate how hybrid models and specialized architectures are pushing the boundaries of NIR spectral analysis, offering improved accuracy, robustness, and applicability across diverse tasks and data types.

Hybrid models combine DL techniques or integrate traditional methods with neural networks, leveraging the strengths of multiple approaches. For example, the T-SAE (Transfer Stacked Auto-Encoder) model used for lead content prediction in oilseed rape [81] combines transfer learning with auto-encoder architectures. This hybrid approach achieved impressive results with R^2 values of 0.9215 and 0.9349 for leaves and roots respectively, outperforming traditional methods. Another notable example is the multi-level dynamic feature extraction model for quality predic-

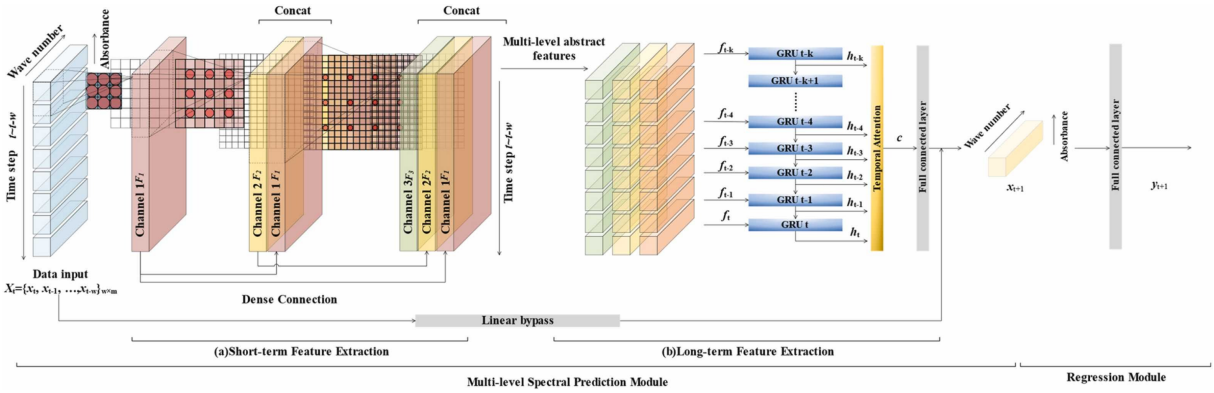

Figure 8: Framework of multi-level dynamic feature-based near-infrared quality prediction [89]

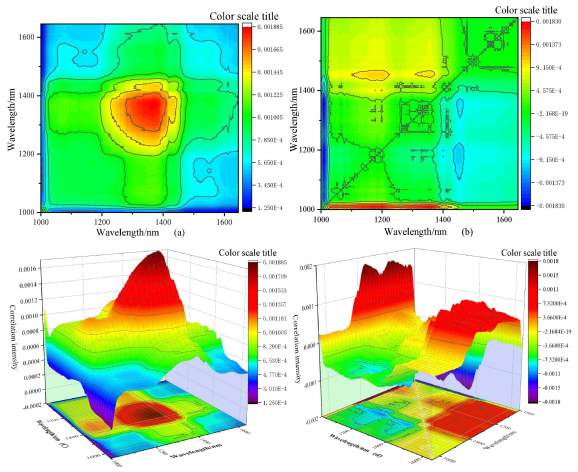

Figure 9: 2D-COS of wolfberries [91]

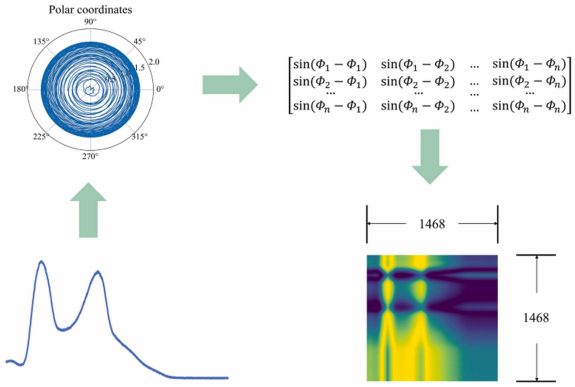

Figure 10: GADF transformation [91]

tion of complex organic compounds [89]. This hybrid approach combined dilated 2D CNNs for short-term feature extraction and GRU with temporal attention for long-term feature extraction, as mentioned above. The model's sophisticated architecture included three dilated 2D CNN with varying dilated rates and kernel sizes, followed by a GRU layer with temporal attention. Besides, specialized architectures are designed to address specific challenges in NIR spectral analysis, often incorporating domain knowledge. The CNN-ATT model used for chickpea classification [84] incorporated an attention mechanism specifically designed to focus on relevant spectral regions. This model, consisting of an attention block followed by three 1D convolutional blocks and two dense layers, achieved 100% accuracy in classification.

Fourth highlight, transfer learning has emerged as a new powerful technique in NIR spectral analysis, allowing models to leverage knowledge from one task or dataset to improve performance on another. This approach is particularly valuable in scenarios with limited training data or when dealing with complex spectral relationships.

Several studies in the provided table demonstrate the effectiveness of transfer learning in various NIR applications. Inter-device transfer learning was successfully implemented in [80], with a 1D-Inception-ResNet model achieving consistently high performance across different spectrometers ($R^2 > 0.96$, $RMSEP < 2.75$). This architecture combines the Inception module for extracting diverse features at multiple scales from spectral data with Residual connections for efficient deep network training, successfully enabling model transfer across three different NIR spectrometers. Other studies have demonstrated effective applications of transfer learning in predicting lead content in oilseed rape. Using a cross-sample transfer approach, research by [81] employed the T-SAE model to predict lead content across different plant samples, achieving strong results for both leaves ($R^2 = 0.9215$, RMSEP = 0.0302 mg/kg) and roots ($R^2 = 0.9349$, RMSEP = 0.0278 mg/kg). In terms of cross-environment transfer, [95] developed the Transfer Stacked Convolutional Auto-Encoder (T-SCAE) architecture, as in Figure 11 to create a model that could work across different growing conditions. By combining pre-trained SCAE models from both silicon and silicon-free environments, this approach effectively extracted deep features for predicting lead concentrations in

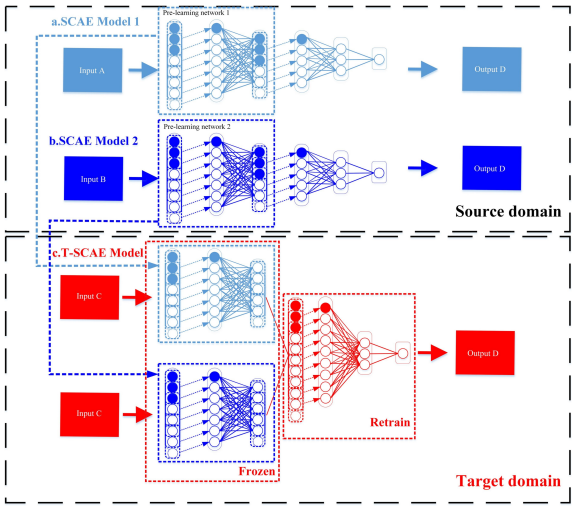

Figure 11: T-SCAE transfer network [95]

oilseed rape leaves, achieving excellent performance in the target domain with an R^2 of 0.9385, RMSEP of 0.02017 mg/kg, and RPD of 3.291. Additionally, spectral range transfer was also illustrated in [86], with the HSCNN model maintaining high performance (F1 score 89%) when reducing from full spectrum to just 3 wavelengths. The model was trained using the AdamW optimizer with a decaying learning rate strategy, and incorporated data augmentation techniques including intensity changes and spatial transformations to prevent overfitting, underscoring the versatility of transfer learning across different analytical dimensions. These examples highlight the versatility and power of transfer learning in NIR spectroscopy, significantly enhancing the adaptability and generalization capabilities of deep learning models in spectral analysis.

Discussion: Development insights

DL architectures demonstrate superior performance over traditional ML methods in NIR spectral analysis through advanced features like automatic hierarchical extraction, attention mechanisms, temporal learning, and dense feature fusion, significantly outperforming conventional approaches with accuracies of 97-100% in various analytical tasks (classification, regression, anomaly detection) across diverse spectral data formats (1D spectra, 2D correlation spectra, dynamic data). The development of hybrid models, specialized architectures, and effective transfer learning capabilities further enhances its robustness in handling complex spectral relationships and limited data scenarios, with proven success in cross-device ($R^2 > 0.96$) and cross-domain applications, marking a significant advancement in spectroscopic analysis. In summary, DL has significantly advanced NIR spectroscopy analysis through superior performance, versatility in data handling, innovative architectures, and effective transfer learning. However, several critical gaps remain that require further research.

Firstly, processing diverse data from multiple sources and devices continues to be a challenge, necessitating the development of more robust methods to ensure consistency and accuracy.

Secondly, the interpretability of DL models in NIR spectral analysis needs improvement, particularly in developing specialized interpretation methods that incorporate expert knowledge in chemistry and spectroscopy.

Lastly, efficient learning from limited data, especially in applications such as food quality assessment and hazardous chemical detection, remains a significant challenge to be addressed.

With ongoing data collection efforts through collaborations with the People's Committee and other governmental agencies, coupled with the accumulation of diverse spectral datasets from multiple devices and regions, as mentioned in Section 4.2, ML and DL techniques will be comprehensively evaluated on larger-scale, heterogeneous data. This expanded evaluation scope will help address several key challenges: processing diverse data from multiple sources and devices to ensure consistency and accuracy, improving DL model interpretability while incorporating domain expertise in chemistry and spectroscopy, and developing efficient learning strategies for limited but critical data scenarios like food quality assessment and hazardous chemical detection.

5 Conclusion

This comprehensive review has analyzed recent advancements in the application of machine learning to near-infrared spectroscopy for food quality assessment. We have identified significant trends across various aspects of this field. In data collection, the trend towards using handheld or ultracompact NIR devices for direct on-site measurements, combined with multiple devices to collect broad NIR spectra, has significantly expanded the applicability of this technology. Regarding pre-processing and wavelength selection, automated and optimized processing techniques, along with the trend of combining multiple methods, have substantially improved model performance. In the field of deep learning, specialized architectures, and hybrid models have been developed, often outperforming traditional ML methods in many NIR spectral tasks. Additionally, transfer learning techniques have shown remarkable potential in addressing challenges related to interdevice variability, cross-sample analysis, and adaptation to new tasks or spectral ranges.

However, significant challenges remain to be addressed. Most notably, there is a lack of comprehensive datasets for food safety applications, a need to improve the interpretability of complex models, and the necessity to develop efficient learning methods from limited data. Additionally, processing diverse data from multiple sources and devices remains a major challenge to be resolved.

Based on these findings, we propose three important research directions for the future: Based on these research directions, we propose three important directions for future

research:

1. Develop larger and more diverse public datasets, with a particular focus on food safety parameters. This can be achieved through collaborations between food safety regulatory authorities and support from local government agencies to build NIR spectral datasets, following Ministry of Health standards. This includes collecting comprehensive spectral data accompanied by reference concentrations validated through independent laboratory testing.

2. Enhance machine learning models' processing capabilities and interpretability for heterogeneous data sources. This involves developing multi-level fusion techniques for integrating data from different devices, creating visualization methods for model interpretability, and establishing comprehensive cross-validation strategies using stratified sampling and bootstrapping techniques to ensure model reliability across diverse operating conditions. These approaches require systematic testing across devices and environments to establish standardized protocols for real-world applications.

3. Develop intelligent automation frameworks that integrate preprocessing selection, feature engineering, and model optimization. These frameworks should adapt to different device types and measurement conditions while maintaining the interpretability of results. The systems should include standardized evaluation metrics and clear protocols to enable seamless integration across NIR platforms in real-world applications.

These proposals aim to establish standardized approaches for NIR spectroscopy with machine learning, improving both the efficiency and reliability of food quality assessment processes. The implementation of these directions will help bridge the gap between theoretical advances and practical applications while addressing current challenges in real-world deployment.

Acknowledgement

This study is funded and implemented for the project with number "24/HĐ-SKHCN, 2023". This work is supported by the People's Committee, Da Nang, and the University of Science and Technology, University of Danang.

References

[1] H. Pu, J. Yu, D.-W. Sun, et al. "Feature construction methods for processing and analysing spectral images and their applications in food quality inspection". In: *Trends in Food Science & Technology* 138 (2023), pp. 726–737. DOI: 10.1016/j.tifs.2023.06.036.

[2] S. M. Pires, H. G. Redondo, J. Pessoa, et al. "Risk ranking of foodborne diseases in Denmark: Reflections on a national burden of disease study". In: *Food Control* 158 (2024), p. 110199. DOI: 10.1016/j.foodcont.2023.110199.

[3] W. Tian, Y. Li, C. Guzman, et al. "Quantification of food bioactives by NIR spectroscopy: Current insights, long-lasting challenges, and future trends". In: *Journal of Food Composition and Analysis* 124 (2023), p. 105708. DOI: 10.1016/j.jfca.2023.105708.

[4] Y. Ozaki et al. *Near-Infrared Spectroscopy: Theory, Spectral Analysis, Instrumentation, and Applications*. Singapore: Springer Singapore, 2021.

[5] A. Hassoun, S. Jagtap, G. Garcia-Garcia, et al. "Food quality 4.0: From traditional approaches to digitalized automated analysis". In: *Journal of Food Engineering* 337 (2023), p. 111216. DOI: 10.1016/j.jfoodeng.2022.111216.

[6] I. Latreche, S. Slatnia, O. Kazar, et al. "A Review on Deep Learning Techniques for EEG-Based Driver Drowsiness Detection Systems". In: *Informatica* 48.3 (2024). DOI: 10.31449/inf.v48i3.5056.

[7] H. A. Mohammed and I. M. Husien. "A Deep Transfer Learning Framework for Robust IoT Attack Detection". In: *Informatica* 48.12 (2024). DOI: 10.31449/inf.v48i12.5955.

[8] J. Ravničan et al. "A Prestudy of Machine Learning in Industrial Quality Control Pipelines". In: *Informatica* 46.2 (2022). DOI: 10.31449/inf.v46i2.3938.

[9] P. Mishra, D. Passos, F. Marini, et al. "Deep learning for near-infrared spectral data modelling: Hypes and benefits". In: *TrAC Trends in Analytical Chemistry* 157 (2022), p. 116804. DOI: 10.1016/j.trac.2022.116804.

[10] H. Nobari Moghaddam, Z. Tamiji, M. Akbari Lakeh, et al. "Multivariate analysis of food fraud: A review of NIR based instruments in tandem with chemometrics". In: *Journal of Food Composition and Analysis* 107 (2022), p. 104343. DOI: 10.1016/j.jfca.2021.104343.

[11] S. Othman, N. R. Mavani, M. A. Hussain, et al. "Artificial intelligence-based techniques for adulteration and defect detections in food and agricultural industry: A review". In: *Journal of Agriculture and Food Research* 12 (2023), p. 100590. DOI: 10.1016/j.jafr.2023.100590.

[12] S. A. D. M. Zahir, A. F. Omar, M. F. Jamlos, et al. "A review of visible and near-infrared (Vis-NIR) spectroscopy application in plant stress detection". In: *Sensors and Actuators A: Physical* 338 (2022), p. 113468. DOI: 10.1016/j.sna.2022.113468.

[13] S. A. D. M. Zahir, M. F. Jamlos, A. F. Omar, et al. "Review – Plant nutritional status analysis employing the visible and near-infrared spectroscopy spectral sensor". In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 304 (2024), p. 123273. DOI: 10.1016/j.saa.2023.123273.

[14] H. Ni, W. Fu, J. Wei, et al. "Non-destructive detection of polysaccharides and moisture in Ganoderma lucidum using near-infrared spectroscopy and machine learning algorithm". In: *LWT* 184 (2023), p. 115001. DOI: 10.1016/j.lwt.2023.115001.

[15] M. I. S. Mohd Hilmi Tan, M. F. Jamlos, A. F. Omar, et al. "Ganoderma boninense classification based on near-infrared spectral data using machine learning techniques". In: *Chemometrics and Intelligent Laboratory Systems* 232 (2023), p. 104718. DOI: 10.1016/j.chemolab.2022.104718.

[16] D. Wang et al. "Determination of polysaccharide content in shiitake mushroom beverage by NIR spectroscopy combined with machine learning: A comparative analysis". In: *Journal of Food Composition and Analysis* 122 (2023), p. 105460. DOI: 10.1016/j.jfca.2023.105460.

[17] Y.-Q. Zhong, J.-Q. Li, X.-L. Li, et al. "Near infrared spectroscopy for simultaneous quantification of five chemical components in Arnebiae Radix (AR) with partial least squares and support vector machine algorithms". In: *Vibrational Spectroscopy* 127 (2023), p. 103556. DOI: 10.1016/j.vibspec.2023.103556.

[18] Z. Guo, Y. Zhang, J. Wang, et al. "Detection model transfer of apple soluble solids content based on NIR spectroscopy and deep learning". In: *Computers and Electronics in Agriculture* 212 (2023), p. 108127. DOI: 10.1016/j.compag.2023.108127.

[19] C. Saenphon, S. Ditcharoen, C. Malai, et al. "Total soluble solids, dry matter content prediction and maturity stage classification of durian fruit using long-wavelength NIR reflectance". In: *Journal of Food Composition and Analysis* 124 (2023), p. 105667. DOI: 10.1016/j.jfca.2023.105667.

[20] S. Nawoya, F. Ssemakula, R. Akol, et al. "Computer vision and deep learning in insects for food and feed production: A review". In: *Computers and Electronics in Agriculture* 216 (2024), p. 108503. DOI: 10.1016/j.compag.2023.108503.

[21] S. Zhang, S. Liu, L. Shen, et al. "Application of near-infrared spectroscopy for the nondestructive analysis of wheat flour: A review". In: *Current Research in Food Science* 5 (2022), pp. 1305–1312. DOI: 10.1016/j.crfs.2022.08.006.

[22] M. M. Nagy, S. Wang, and M. A. Farag. "Quality analysis and authentication of nutraceuticals using near IR (NIR) spectroscopy: A comprehensive review of novel trends and applications". In: *Trends in Food Science & Technology* 123 (2022), pp. 290–309. DOI: 10.1016/j.tifs.2022.03.005.

[23] L. Shuai, Z. Li, Z. Chen, et al. "A research review on deep learning combined with hyperspectral Imaging in multiscale agricultural sensing". In: *Computers and Electronics in Agriculture* 217 (2024), p. 108577. DOI: 10.1016/j.compag.2023.108577.

[24] Y. Liu, H. Pu, and D.-W. Sun. "Efficient extraction of deep image features using convolutional neural network (CNN) for applications in detecting and analysing complex food matrices". In: *Trends in Food Science & Technology* 113 (2021), pp. 193–204. DOI: 10.1016/j.tifs.2021.04.042.

[25] W. Zhang, L. C. Kasun, Q. J. Wang, et al. "A Review of Machine Learning for Near-Infrared Spectroscopy". In: *Sensors* 22 (2022), p. 9764. DOI: 10.3390/s22249764.

[26] P. Mishra, R. Nikzad-Langerodi, F. Marini, et al. "Are standard sample measurements still needed to transfer multivariate calibration models between near-infrared spectrometers? The answer is not always". In: *TrAC Trends in Analytical Chemistry* 143 (2021), p. 116331. DOI: 10.1016/j.trac.2021.116331.

[27] H. Ji, D. Pu, W. Yan, et al. "Recent advances and application of machine learning in food flavor prediction and regulation". In: *Trends in Food Science & Technology* 138 (2023), pp. 738–751. DOI: 10.1016/j.tifs.2023.07.012.

[28] Y. Zhang and Y. Wang. "Machine learning applications for multi-source data of edible crops: A review of current trends and future prospects". In: *Food Chemistry: X* 19 (2023), p. 100860. DOI: 10.1016/j.fochx.2023.100860.

[29] C. A. Nunes, M. N. Ribeiro, T. C. de Carvalho, et al. "Artificial intelligence in sensory and consumer studies of food products". In: *Current Opinion in Food Science* 50 (2023), p. 101002. DOI: 10.1016/j.cofs.2023.101002.

[30] B. Debus et al. "Deep learning in analytical chemistry". In: *TrAC Trends in Analytical Chemistry* 145 (2021), p. 116459. DOI: 10.1016/j.trac.2021.116459.

[31] M. Zareef, M. Arslan, M. M. Hassan, et al. "Recent advances in assessing qualitative and quantitative aspects of cereals using nondestructive techniques: A review". In: *Trends in Food Science & Technology* 116 (2021), pp. 815–828. DOI: 10.1016/j.tifs.2021.08.012.

[32] M. Hernández-Jiménez, I. Revilla, A. M. Vivar-Quintana, et al. "Performance of benchtop and portable spectroscopy equipment for discriminating Iberian ham according to breed". In: *Current Research in Food Science* 8 (2024), p. 100675. DOI: 10.1016/j.crfs.2024.100675.

[33] L. Yuan, X. Meng, K. Xin, et al. "A comparative study on classification of edible vegetable oils by infrared, near infrared and fluorescence spectroscopy combined with chemometrics". In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 288 (2023), p. 122120. DOI: 10.1016/j.saa.2022.122120.

[34] Widyaningrum, Y. A. Purwanto, S. Widodo, et al. "Rapid assessment of vanilla (Vanilla planifolia) quality parameters using portable near-infrared spectroscopy combined with random forest". In: *Journal of Food Composition and Analysis* 133 (2024), p. 106346. DOI: 10.1016/j.jfca.2024.106346.

[35] R. Chen, S. Li, H. Cao, et al. "Rapid quality evaluation and geographical origin recognition of ginger powder by portable NIRS in tandem with chemometrics". In: *Food Chemistry* 438 (2024), p. 137931. DOI: 10.1016/j.foodchem.2023.137931.

[36] J. P. Cruz-Tirado, M. S. S. Vieira, O. O. V. Correa, et al. "Detection of adulteration of Alpaca (Vicugna pacos) meat using a portable NIR spectrometer and NIR-hyperspectral imaging". In: *Journal of Food Composition and Analysis* 126 (2024), p. 105901. DOI: 10.1016/j.jfca.2023.105901.

[37] R. Zhu et al. "High-accuracy classification and origin traceability of peanut kernels based on near-infrared (NIR) spectroscopy using Adaboost - Maximum uncertainty linear discriminant analysis". In: *Current Research in Food Science* 8 (2024), p. 100766. DOI: 10.1016/j.crfs.2024.100766.

[38] J. Liang et al. "Integrating portable NIR spectrometry with deep learning for accurate Estimation of crude protein in corn feed". In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 314 (2024), p. 124203. DOI: 10.1016/j.saa.2024.124203.

[39] K. Yao, J. Sun, B. Zhang, et al. "On-line monitoring of egg freshness using a portable NIR spectrometer combined with deep learning algorithm". In: *Infrared Physics & Technology* 138 (2024), p. 105207. DOI: 10.1016/j.infrared.2024.105207.

[40] S. Ghidini, M. O. Varrà, D. Bersellini, et al. "Real-time and non-destructive control of the freshness and viability of live mussels through portable near-infrared spectroscopy". In: *Food Control* 160 (2024), p. 110353. DOI: 10.1016/j.foodcont.2024.110353.

[41] Z. Wu, C. Li, H. Liu, et al. "Quantification of caffeine and catechins and evaluation of bitterness and astringency of Pu-erh ripen tea based on portable near-infrared spectroscopy". In: *Journal of Food Composition and Analysis* 125 (2024), p. 105793. DOI: 10.1016/j.jfca.2023.105793.

[42] K. B. Beć, J. Grabska, and C. W. Huck. "Miniaturized NIR Spectroscopy in Food Analysis and Quality Control: Promises, Challenges, and Perspectives". In: *Foods* 11 (2022), p. 1465. DOI: 10.3390/foods11101465.

[43] J. M. Netto et al. "Authenticity of almond flour using handheld near infrared instruments and one class classifiers". In: *Journal of Food Composition and Analysis* 115 (2023), p. 104981. DOI: 10.1016/j.jfca.2022.104981.

[44] X. Chu et al. *Chemometric Methods in Analytical Spectroscopy Technology*. Singapore: Springer Nature Singapore, 2022.

[45] J. A. Diaz-Olivares, A. Van Nuenen, M. J. Gote, et al. "Near-infrared spectra dataset of milk composition in transmittance mode". In: *Data in Brief* 51 (2023), p. 109767. DOI: 10.1016/j.dib.2023.109767.

[46] D. Li et al. "Research on Data Fusion and Sharing Based on Power Big Data". In: *2023 9th Annual International Conference on Network and Information Systems for Computers (ICNISC)*. Wuhan, China, 2023, pp. 287–290. DOI: 10.1109/ICNISC60562.2023.00070.

[47] L. Zhang et al. "Multi-source heterogeneous data fusion". In: *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*. Chengdu, China, 2018, pp. 47–51. DOI: 10.1109/ICAIBD.2018.8396165.

[48] Tianzhe Jiao et al. "A Comprehensive Survey on Deep Learning Multi-Modal Fusion: Methods, Technologies and Applications". In: *Computers, Materials and Continua* 80.1 (2024), pp. 1–35. DOI: 10.32604/cmc.2024.053204.

[49] G. Van Kollenburg, Y. Weesepoel, H. Parastar, et al. "Dataset of the application of handheld NIR and machine learning for chicken fillet authenticity study". In: *Data in Brief* 29 (2020), p. 105357. DOI: 10.1016/j.dib.2020.105357.

[50] I. Malounas, W. Vierbergen, S. Kutluk, et al. "SpectroFood dataset: A comprehensive fruit and vegetable hyperspectral meta-dataset for dry matter estimation". In: *Data in Brief* 52 (2024), p. 110040. DOI: 10.1016/j.dib.2024.110040.

[51] G. Bonifazi et al. "A dataset of visible – Short wave InfraRed reflectance spectra collected on precooked pasta products". In: *Data in Brief* 36 (2021), p. 106989. DOI: 10.1016/j.dib.2021.106989.

[52] M. Ryckewaert, D. Héran, C. Feilhes, et al. “Dataset containing spectral data from hyperspectral imaging and sugar content measurements of grapes berries in various maturity stage”. In: *Data in Brief* 46 (2023), p. 108822. DOI: 10.1016/j.dib.2022.108822.

[53] A. A. Munawar, Kusumiyati, and D. Wahyuni. “Near infrared spectroscopic data for rapid and simultaneous prediction of quality attributes in intact mango fruits”. In: *Data in Brief* 27 (2019), p. 104789. DOI: 10.1016/j.dib.2019.104789.

[54] R. Hayati, A. A. Munawar, and F. Fachruddin. “Enhanced near infrared spectral data to improve prediction accuracy in determining quality parameters of intact mango”. In: *Data in Brief* 30 (2020), p. 105571. DOI: 10.1016/j.dib.2020.105571.

[55] Agussabti et al. “Data analysis on near infrared spectroscopy as a part of technology adoption for cocoa farmer in Aceh Province, Indonesia”. In: *Data in Brief* 29 (2020), p. 105251. DOI: 10.1016/j.dib.2020.105251.

[56] A. Zgouz, D. Héran, B. Barthès, et al. “Dataset of visible-near infrared handheld and microspectrometers – comparison of the prediction accuracy of sugarcane properties”. In: *Data in Brief* 31 (2020), p. 106013. DOI: 10.1016/j.dib.2020.106013.

[57] K. Kusumiyati et al. “Enhanced visible/near-infrared spectroscopic data for prediction of quality attributes in Cucurbitaceae commodities”. In: *Data in Brief* 39 (2021), p. 107458. DOI: 10.1016/j.dib.2021.107458.

[58] I. Malounas et al. “Evaluation of a hyperspectral image pipeline toward building a generalisation capable crop dry matter content prediction model”. In: *Biosystems Engineering* 247 (2024), pp. 153–161. DOI: 10.1016/j.biosystemseng.2024.09.009.

[59] H. Parastar et al. “Integration of handheld NIR and machine learning to ”Measure & Monitor” chicken meat authenticity”. In: *Food Control* 112 (2020), p. 107149. DOI: 10.1016/j.foodcont.2020.107149.

[60] J.-L. Z. Zaukuu, A. A. Nkansah, E. T. Mensah, et al. “Non-destructive authentication of melon seed (Cucumeropsis mannii) powder using a pocket-sized near-infrared (NIR) spectrophotometer with multiple spectral preprocessing”. In: *Journal of Food Composition and Analysis* 134 (2024), p. 106425. DOI: 10.1016/j.jfca.2024.106425.

[61] S. Wang, M. Lin, Y. Meng, et al. “Self-expansion full information optimization strategy: Convenient and efficient method for near infrared spectrum autoanalysis”. In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 303 (2023), p. 123224. DOI: 10.1016/j.saa.2023.123224.

[62] S. Wang, P. Zhang, J. Chang, et al. “A powerful tool for near-infrared spectroscopy: Synergy adaptive moving window algorithm based on the immune support vector machine”. In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 282 (2022), p. 121631. DOI: 10.1016/j.saa.2022.121631.

[63] U. Blazhko et al. “Comparison of augmentation and pre-processing for deep learning and chemometric classification of infrared spectra”. In: *Chemometrics and Intelligent Laboratory Systems* 215 (2021), p. 104367. DOI: 10.1016/j.chemolab.2021.104367.

[64] A. Sitorus and R. Lapcharoensuk. “Development of automatic tuning for combined preprocessing and hyperparameters of machine learning and its application to NIR spectral data of coconut milk adulteration”. In: *Food Chemistry* 457 (2024), p. 140108. DOI: 10.1016/j.foodchem.2024.140108.

[65] N. D. Arianti, E. Saputra, and A. Sitorus. “An automatic generation of pre-processing strategy combined with machine learning multivariate analysis for NIR spectral data”. In: *Journal of Agriculture and Food Research* 13 (2023), p. 100625. DOI: 10.1016/j.jafr.2023.100625.

[66] J.-M. Roger, A. Biancolillo, and F. Marini. “Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy”. In: *Chemometrics and Intelligent Laboratory Systems* 199 (2020), p. 103975. DOI: 10.1016/j.chemolab.2020.103975.

[67] Z. Wang, Q. Wu, and M. Kamruzzaman. “Portable NIR spectroscopy and PLS based variable selection for adulteration detection in quinoa flour”. In: *Food Control* 138 (2022), p. 108970. DOI: 10.1016/j.foodcont.2022.108970.

[68] Y. Qin, K. Song, N. Zhang, et al. “Robust NIR quantitative model using MIC-SPA variable selection and GA-ELM”. In: *Infrared Physics & Technology* 128 (2023), p. 104534. DOI: 10.1016/j.infrared.2022.104534.

[69] M. Marañón, J. Fernández-Novales, J. Tardaguila, et al. “NIR attribute selection for the development of vineyard water status predictive models”. In: *Biosystems Engineering* 229 (2023), pp. 167–178. DOI: 10.1016/j.biosystemseng.2023.04.001.

[70] K. Yao, J. Sun, J. Cheng, et al. “Monitoring S-ovalbumin content in eggs during storage using portable NIR spectrometer and multivariate analysis”. In: *Infrared Physics & Technology* 131 (2023), p. 104685. DOI: 10.1016/j.infrared.2023.104685.

[71] B. Mahanty. "Adaptive Bottom-Up Space Exploration in model population analysis: An agile variable selection algorithm for PLS models". In: *Chemometrics and Intelligent Laboratory Systems* 203 (2020), p. 104057. DOI: 10.1016/j.chemolab.2020.104057.

[72] J. Li, J. Deng, X. Bai, et al. "Quantitative analysis of aflatoxin B1 of peanut by optimized support vector machine models based on near-infrared spectral features". In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 303 (2023), p. 123208. DOI: 10.1016/j.saa.2023.123208.

[73] X. Miao, Y. Miao, Y. Liu, et al. "Measurement of nitrogen content in rice plant using near infrared spectroscopy combined with different PLS algorithms". In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 284 (2023), p. 121733. DOI: 10.1016/j.saa.2022.121733.

[74] C. Du, L. Sun, H. Bai, et al. "Quantitative detection of azodicarbonamide in wheat flour by near-infrared spectroscopy based on two-step feature selection". In: *Chemometrics and Intelligent Laboratory Systems* 219 (2021), p. 104445. DOI: 10.1016/j.chemolab.2021.104445.

[75] C. Du, L. Sun, H. Bai, et al. "Quantitative detection of talcum powder in wheat flour based on near-infrared spectroscopy and hybrid feature selection". In: *Infrared Physics & Technology* 123 (2022), p. 104185. DOI: 10.1016/j.infrared.2022.104185.

[76] R. Zheng, Y. Jia, C. Ullagaddi, et al. "Optimizing feature selection with gradient boosting machines in PLS regression for predicting moisture and protein in multi-country corn kernels via NIR spectroscopy". In: *Food Chemistry* 456 (2024), p. 140062. DOI: 10.1016/j.foodchem.2024.140062.

[77] Z. Guo, L. Zhai, Y. Zou, et al. "Comparative study of Vis/NIR reflectance and transmittance method for on-line detection of strawberry SSC". In: *Computers and Electronics in Agriculture* 218 (2024), p. 108744. DOI: 10.1016/j.compag.2024.108744.

[78] Z. Ji, J. Zhu, J. Deng, et al. "Quantitative determination of zearalenone in wheat by the CSA-NIR technique combined with chemometrics algorithms". In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* (2024), p. 124858. DOI: 10.1016/j.saa.2024.124858.

[79] J. Zhu et al. "Improve the accuracy of FT-NIR for determination of zearalenone content in wheat by using the characteristic wavelength optimization algorithm". In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 313 (2024), p. 124169. DOI: 10.1016/j.saa.2024.124169.

[80] A. Tan et al. "1D-inception-resnet for NIR quantitative analysis and its transferability between different spectrometers". In: *Infrared Physics & Technology* 129 (2023), p. 104559. DOI: 10.1016/j.infrared.2023.104559.

[81] X. Zhou, C. Zhao, J. Sun, et al. "Detection of lead content in oilseed rape leaves and roots based on deep transfer learning and hyperspectral imaging technology". In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 290 (2023), p. 122288. DOI: 10.1016/j.saa.2022.122288.

[82] X. Chen et al. "Rapid identification of total phenolic content levels in boletes by two-dimensional correlation spectroscopy combined with deep learning". In: *Vibrational Spectroscopy* 121 (2022), p. 103404. DOI: 10.1016/j.vibspec.2022.103404.

[83] J. A. Martins, D. Rodrigues, A. M. Cavaco, et al. "Estimation of soluble solids content and fruit temperature in "Rocha" pear using Vis-NIR spectroscopy and the SpectraNet–32 deep learning architecture". In: *Postharvest Biology and Technology* 199 (2023), p. 112281. DOI: 10.1016/j.postharvbio.2023.112281.

[84] D. Saha, T. Senthilkumar, C. B. Singh, et al. "Rapid and non-destructive detection of hard to cook chickpeas using NIR hyperspectral imaging and machine learning". In: *Food and Bioproducts Processing* 141 (2023), pp. 91–106. DOI: 10.1016/j.fbp.2023.07.006.

[85] L. Wang, J. Liang, F. Li, et al. "Deep learning based on the Vis-NIR two-dimensional spectroscopy for adulteration identification of beef and mutton". In: *Journal of Food Composition and Analysis* 126 (2024), p. 105890. DOI: 10.1016/j.jfca.2023.105890.

[86] S. Castillo-Girones, R. Van Belleghem, N. Wouters, et al. "Detection of subsurface bruises in plums using spectral imaging and deep learning with wavelength selection". In: *Postharvest Biology and Technology* 207 (2024), p. 112615. DOI: 10.1016/j.postharvbio.2023.112615.

[87] J. Yang, X. Ma, H. Guan, et al. "A recognition method of corn varieties based on spectral technology and deep learning model". In: *Infrared Physics & Technology* 128 (2023), p. 104533. DOI: 10.1016/j.infrared.2022.104533.

[88] G. Wang, X. Jiang, X. Li, et al. "Determination of watermelon soluble solids content based on visible/near infrared spectroscopy with convolutional neural network". In: *Infrared Physics & Technology* 133 (2023), p. 104825. DOI: 10.1016/j.infrared.2023.104825.

[89] Z. Chen, X. Luan, and F. Liu. “Deep learning near-infrared quality prediction based on multi-level dynamic feature”. In: *Vibrational Spectroscopy* 123 (2022), p. 103450. DOI: 10.1016/j.vibspec.2022.103450.

[90] W. He, H. He, F. Wang, et al. “Non-destructive detection and recognition of pesticide residues on garlic chive (Allium tuberosum) leaves based on short wave infrared hyperspectral imaging and one-dimensional convolutional neural network”. In: *Food Measure* 15 (2021), pp. 4497–4507. DOI: 10.1007/s11694-021-01012-7.

[91] Fujia Dong et al. “Identification of the proximate geographical origin of wolfberries by two-dimensional correlation spectroscopy combined with deep learning”. In: *Computers and Electronics in Agriculture* 198 (2022), p. 107027. ISSN: 0168-1699. DOI: 10.1016/j.compag.2022.107027.

[92] Z. Liu et al. “Detection of apple moldy core disease by fusing vibration and Vis/NIR spectroscopy data with dual-input MLP-Transformer”. In: *Journal of Food Engineering* 382 (2024), p. 112219. DOI: 10.1016/j.jfoodeng.2024.112219.

[93] Y. Li et al. “Combined gramian angular difference field image coding and improved mobile vision transformer for determination of apple soluble solids content by Vis-NIR spectroscopy”. In: *Journal of Food Composition and Analysis* 131 (2024), p. 106200. DOI: 10.1016/j.jfca.2024.106200.

[94] J. Yang et al. “TeaNet: Deep learning on Near-Infrared Spectroscopy (NIR) data for the assurance of tea quality”. In: *Computers and Electronics in Agriculture* 190 (2021), p. 106431. DOI: 10.1016/j.compag.2021.106431.

[95] Xin Zhou et al. “Determination of lead content in oilseed rape leaves in silicon-free and silicon environments based on deep transfer learning and fluorescence hyperspectral imaging”. In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 311 (2024), p. 123991. ISSN: 1386-1425. DOI: 10.1016/j.saa.2024.123991.