Seed Science Research

cambridge.org/ssr

Spectral imaging and chemometrics applied at phenotyping in seed science studies: a systematic review

Thomas B. Michelon*, Elisa Serra Negra Vieira and Maristela Panobianco

Department of Plant Science, Federal University of Paraná, R. dos Funcionários, 1540, CEP 80035-050, Curitiba, PR, Brazil

Review Paper

Cite this article: Michelon TB, Serra Negra Vieira E, Panobianco M (2023). Spectral imaging and chemometrics applied at phenotyping in seed science studies: a systematic review. *Seed Science Research* **33**, 9–22. https://doi.org/10.1017/S0960258523000028

Received: 7 July 2022
Revised: 16 December 2022
Accepted: 4 January 2023
First published online: 26 January 2023

Keywords:
deep learning; hyperspectral imaging; machine learning; multispectral imaging; seed classification; spectroscopy

***Author for Correspondence:**
Thomas B. Michelon,
E-mail: thomasbrunomichelon@gmail.com

Abstract

The evaluation of the genetic quality of a seed lot is crucial for the quality control process in its production and commercialization, as well as in the identification of superior genotypes and the verification of the correct crossing in plant breeding programmes. Current techniques, based on the identification of seed morphological characteristics, require skilled analysts, while biochemical methods are time-consuming and costly. The application of spectral imaging analysis, which combines digital imaging with spectroscopy, is gaining ground as a fast, accurate and non-destructive method. The success of this technique is closely linked to chemometric techniques, which use statistical and mathematical tools in data processing. The aim of the work was to evaluate the main procedures in terms of spectral image analysis and chemometric procedures applied in seed phenotyping and its practical application. A systematic review was conducted using the PRISMA methodology, in which a total of 1304 articles were identified and screened to the inclusion of 44 articles pertaining to the scope. It was concluded that spectral image analysis has a high ability to classify seeds of different genotypes (93.33%) in a range of situations: between cultivars; hybrids and progenitors; and hybrids and lines, as well as in the separation of coated seeds. Accurate classification can be obtained by different strategies, such as the choice of the equipment type, the spectrum range and extra features, guided by the characteristics of the species, as well as in the choice of algorithms and dimensionality reduction procedures for the optimization of models when there is a large amount of data. Despite the fact that the practical application of this technique in seed phenotyping still needs to be developed for use in laboratories with large volumes of analyses, lots, genotypes and harvests. Research has been accelerated to overcome the practical challenges of this method, as seen in works using model update algorithms, online classification systems, and real-time classification maps. Thus, there are strong indications that the application of multispectral image analysis will reach the routine of seed analysis laboratories.

Introduction

Varietal sorting is an essential part of the quality control process of a seed lot, either in germplasm bank management, production or commercialization, in order to identify its genetic quality and avoid species mixture (Elmasry et al., 2019). For plant breeding programmes, cultivar discrimination is also crucial to prove the correct crossing between plants, identify superior genotypes and guarantee seed homogeneity according to their minimum descriptors for the purposes of registering new cultivars. For all these purposes, the process of separating seeds by its morphological characteristics, such as colour, texture and shape, requires well-trained analysts and sometimes time-consuming and expensive biochemical and molecular techniques (Hansen et al., 2016; Zhu et al., 2020).

Thus, non-destructive, rapid and non-subjective methods are of great interest in determining seed quality (Elmasry et al., 2019; Xia et al., 2019b). In this regard, multispectral image analysis is a promising alternative that combines spectroscopy with digital imaging. The technique is based on the reflectance of an object – the intensity that a given surface reflects a wavelength. An object can be illuminated by different wavelengths [e.g. visible light, near-infrared (NIR)], and when combined with a digital image, the reflectance of each pixel of this object's image can be measured to differentiate it from another (Boelt et al., 2018; Xia et al., 2019a,b).

Since each pixel contains a dataset (reflectance from each wavelength), the result is a large amount of data proportional to the number of wavelengths used and the size of the image. As these data are considered chemical information, the role of chemometrics is to use statistical and mathematical tools to obtain the most important information from the dataset of each object (Amigo, 2020).

https://doi.org/10.1017/S0960258523000028 Published online by Cambridge University Press

Spectral image analysis is considered one of the major emerging technologies in seed analysis and technology. Its versatility, non-destructive characteristics and rapid determination of quality attributes of a seed lot, combined with data science, make it possible to automate the entire seed sorting process (Elmasry et al., 2019; Xia et al., 2019a,b; Amigo, 2020; Zhou et al., 2020a,b).

The success in applying the technique lies in combining experimental issues with the process of extracting information from the seeds and the chemometric strategy used, which may include from the choice of classification algorithms to data dimensionality reduction processes. Therefore, the process of choosing each aspect involved in the analysis is not trivial, and thus, this systematic review aims to evaluate the main procedures in terms of spectral imaging analysis and chemometric procedures applied in seed phenotyping as well as its practical application.

Methodology

The study followed the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) methodology (Moher et al., 2009; Page et al., 2021), as it presents a clear and systematic research method with a focus on reproducibility.

Inclusion and exclusion criteria

The inclusion and exclusion criteria were based on literature type, access, period, language and scope (Table 1). The 15-year period was chosen to limit the search to new papers, given the recent expansion of spectral imaging technology in seed science. Regarding the scope, only papers on spectral imaging analysis (multispectral and hyperspectral) in seeds were considered; thus, papers using material not considered as seeds (i.e. grains) were not considered. Articles using spectral analysis only to quantify chemical components (e.g. oil, protein content) of seeds but did not classify them into different genotypes (e.g. cultivars or varieties) were not considered. Language was considered as a criterion to avoid bias in the translation of non-English language papers.

Search methodology

The keywords for the present work, as well as their synonyms, were obtained through prior review in studies related to the areas of seed science and technology and spectroscopy (Table 2). The databases used were the Web of Science Core Collection (WOS) and Scopus and were chosen according to previous research on the number of articles related to the scope present in each one. WOS is the database of Clarivate Analytics and has indexed more than 21,000 papers covering 256 disciplines, while the Scopus database belongs to Elsevier and is one of the most related to plant science with peer-reviewed articles. In addition to covering a large quantity of articles related to the topic, these databases allow the inclusion of Boolean operators for the search strategy, as well as symbols that allow the inclusion of all possible terms with the same root.

The search consisted of three steps: identification of potential articles, screening, and inclusion of articles (Fig. 1). A total of 1304 articles were identified and duplicates were removed with the aid of the Mendeley Reference Manager management programme (Dearden et al., 2011). A total of 308 articles were eliminated based on their characteristics as per the exclusion criteria, while 508 articles were excluded as per the scope from the evaluation of the title and abstract. A total of 60 articles were evaluated in full and 44 were included in the review.

Table 1. Inclusion and exclusion criteria

Criterion	Eligibility	Exclusion
Literature type	Article	Reviews, conference paper and book chapter
Access	Full-text available	
Period	Between 2006 and 2021	<2006
Language	English	Non-English
Scope	Uses spectral imaging (e.g. hyperspectral, multispectral imaging) applied to seed phenotyping	Did not use seeds; did not combine spectroscopy to image analysis; or just quantify certain components but did not differ cultivars, varieties, etc.

Statistical analysis

From the articles evaluated, data were collected regarding the experiment, the best classification model obtained in each study, as well as other information deemed relevant (Table 3), to identify possible factors influencing the accuracy of seed classification through spectral imaging analysis. A multiple generalized linear regression model with gamma distribution and log-link function was used, due to the non-normality of the data, in conjunction with the stepwise feature selection algorithm (backward and forward) to select the final model. The algorithm adds and removes features and compares the models by means of Akaike's Selection Criterion (AIC), in order to obtain a final model with the feature (or the combination of features) best-fitted (with the lowest AIC value) to predict the accuracy of spectral imaging analysis applied to seed phenotyping.

Results and applications

A total of 44 articles from the systematic review were included; since the authors reported more than one experiment in some papers, data from all the experiments performed were listed,

Table 2. The search strategy used for the systematic review process

Database	Search criteria
Web of Science	TS = ((seed OR seeds) AND (multispectral OR hyperspectral OR spectral OR spectroscopy OR NIR OR 'near infrared' OR nearinfrared OR 'near-infrared' OR reflectance OR chemometrics) AND (variet* OR cultiv* OR phenot* OR breed* OR hybrid* OR transgenic*) AND (classification OR discrimination OR identification OR determination OR phenotyping))
Scopus	TITLE-ABS-KEY((seed OR seeds) AND (multispectral OR hyperspectral OR spectral OR spectroscopy OR NIR OR 'near infrared' OR nearinfrared OR 'near-infrared' OR reflectance OR chemometrics) AND (variet* OR cultiv* OR phenot* OR breed* OR hybrid* OR transgenic*) AND (classification OR discrimination OR identification OR determination OR phenotyping))

https://doi.org/10.1017/S0960258523000028 Published online by Cambridge University Press

including data from the best-performing classification model (Table 4).

Accuracy, data splitting and validation methods

The average accuracy of the reviewed studies (considering all experiments listed in Table 4) was 93.33% (±7.07%). In some studies, the application of spectral image analysis resulted in 100% classification accuracy, as in Zhu et al. (2019a), on 10 soybean seed varieties using the Ensemble Learning classification algorithm. A similar result was found for Liu et al. (2014b), whose study on transgenic and non-transgenic rice seeds, by means of the Least-squares support vector machine (LS-SVM) algorithm, used both spectral information and biometric data regarding seed morphology. It was also the case of the study of Kong et al. (2013) on four rice seed varieties, using the Random Forest (RF) algorithm, and the study of Rodríguez-Pulido et al. (2013), which separated four grape varieties using general discriminant analysis (GDA).

Their high accuracy suggests a promising feature of spectral image analysis in distinguishing genotypes, but there are some concerns. The first is about the amount of classification groups: only 46 and 23% of the experiments had more than 5 and 10 categories, respectively. In works that used many categories, for example, Fabiyi et al. (2020), with 90 cultivars, although the overall accuracy was relatively high (79.64%) using the RF algorithm, for some cultivars accuracy was only 30–50%. The same result was found in the study of Zhou et al. (2020a), in which the overall accuracy was 93.10% using a deep learning algorithm (convolutional neural network – CNN) for the classification of 30 cultivars, while there was a variation of more than 20% in classification accuracy for certain cultivars.

It is not clear, in most of the reviewed papers, if spectral image analysis was applied owing to its agility and automation or because of the ability to classify cultivars in situations in which classification by visual morphological characteristics was not possible. This point is important because knowing whether the genotypes were chosen randomly or whether they were chosen arbitrarily from characteristics where separation would be possible even by eye allows one to establish to what extent spectral image analysis is applicable in situations of cultivar diversity, as occurs in the seed industry.

Another point of concern is test and validation data: in most works, there was (1) the absence of test data and (2) the absence of test and/or validation lots (e.g. seeds from other harvests). Ideally, when enough data are available, seed samples should be divided into training, validation and test data. Training data are used by the algorithm to estimate the model; validation data are not to be used in training, in order to gather unbiased information about the quality of the models developed. Validation data are used for predicting errors in each model for the purpose of selecting the best model. Since validation data are used

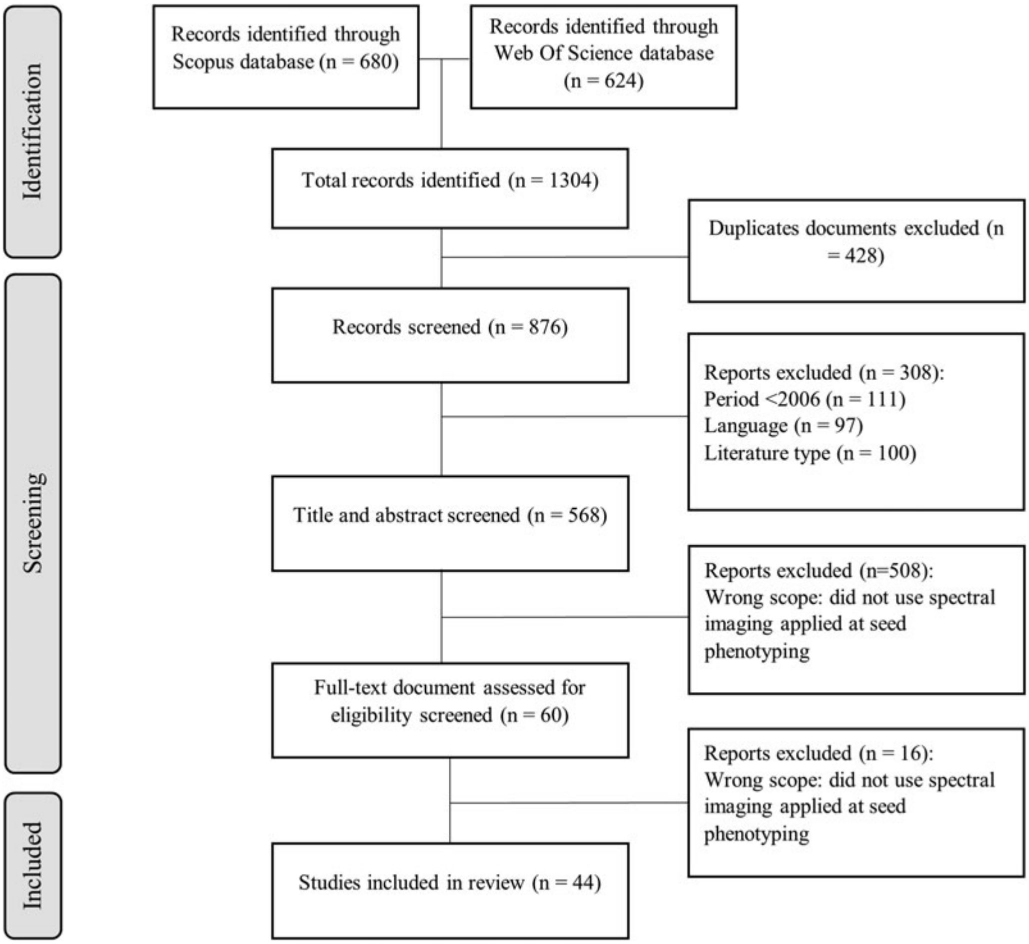
Identification
Screening
Included
Records identified through Scopus database (n = 680)
Records identified through Web Of Science database (n = 624)
Total records identified (n = 1304)
Duplicates documents excluded (n = 428)
Records screened (n = 876)
Reports excluded (n = 308):
Period <2006 (n = 111)
Language (n = 97)
Literature type (n = 100)
Title and abstract screened (n = 568)
Reports excluded (n=508):
Wrong scope: did not use spectral imaging applied at seed phenotyping
Full-text document assessed for eligibility screened (n = 60)
Reports excluded (n = 16):
Wrong scope: did not use spectral imaging applied at seed phenotyping
Studies included in review (n = 44)

Fig. 1. Selection of articles according to the PRISMA framework.

https://doi.org/10.1017/S0960258523000028 Published online by Cambridge University Press

Table 3. Features that might affect the accuracy of seed distinction in spectral analysis applied to seed phenotyping

Features	Levels
Crop type	Agricultural crops, horticultural crops, fruit production, others
Application	Varietal discrimination, haploid, transgenic/non-transgenic, hybrid/progenitors
Spectrum	NIR, VIS–NIR
Sensor	Multispectral imaging, hyperspectral imaging
Number of wavelengths	19–700
Number of seed groups	2–90
Total seeds used	376–147,096
Algorithm class	Machine learning, deep learning
Extra features (e.g. morphology, texture, colour)	Present (1); Absent (0)
Wavelength selection and/or dimensionality reduction	Present (1); Absent (0)
Wavelength preprocessing	Present (1); Absent (0)

constantly (depending on the number of models to be tested), test data (i.e. data not yet used) are commonly used to obtain the true error of the final model (i.e. generalization error), and these data are used only once so as not to overestimate accuracy (Hastie et al., 2017).

In most works of the present review, despite the large number of seeds being used, test data were not used – only validation data (even when it was referred to as test data in the studies, owing to different definitions), which may lead to high accuracy. In studies with a small number of seeds, an alternative is to use cross-validation, in which samples go through *n* data splitting cycles (in training and validation), model building and error computation, and final accuracy is determined from the average error of the *n* models obtained (Hastie et al., 2017). However, of the experiments that used approximately less than 100 seeds per classification category (referring to the first quartile of the variable number of seeds in Table 4), 37% did not perform cross-validation, which may cause overestimation of the resulting accuracy.

Another aspect regarding data division is that only 6% of the studies used validation and/or test lots (i.e. from other harvests and/or regions). In the works that did not use validation lots, high accuracy may have been due to a model overfitted to the lot; consequently, there may not be such accuracy in the classification of the same cultivars from other harvests and regions (He et al., 2016; Huang et al., 2016a). For example, Huang et al. (2016a), when classifying seeds of four wheat varieties and using – as test data – seeds from the same year as those used for training, found 100% classification accuracy using the LS-SVM algorithm. However, when using seeds from other years, accuracy was only 75.4%. Similarly, Shrestha et al. (2016a) used tomato seeds of four cultivars from three harvest years, in experiments with seeds only from the same year and with the mixture of seeds from the other years, both in the test and training data. For the fitted and validated models with seeds from the same year, they found accuracy per cultivar from 73 to 100%, whereas for the sample with mixed seeds from other harvests, accuracy ranged from 34 to 88%.

There was a great variation in the total number of seeds used per classified genotype, as observed in each experiment, even in those that used the same species. For example, in the works performed on wheat seeds, the number of seeds used per class ranged from 20 to 5100 seeds. Few of the reviewed studies evaluated the influence of seed quantity on training samples. For instance, Qiu et al. (2018) tested training samples with different amounts for designing their classification models, ranging from 100 to 3000 seeds, for each of the four cultivars. They found when using more than 1500 seeds, the increment in accuracy was not significant. Certainly, the accuracy determined in experiments that used larger samples leads to more confidence, but the approach of studying the ideal number of seeds has more practical applicability, since the increase in the amount of samples generates extra processing costs, without necessarily leading to a significant increase in the accuracy of the models. Thus, stipulating the optimal number of seeds is important to achieve a balance between the cost and performance of a model, which would, thus, facilitate the applicability of the analysis (Qiu et al., 2018).

One way to obtain more data without necessarily increasing the number of seeds in a sample is by using the spectral information of each seed pixel (i.e. pixel-wise spectrum), as opposed to averaging the seed spectrum (i.e. object-wise), as evaluated by Zhu et al. (2019c) in classifying three soybean cultivars. The authors used the pixel-wise spectrum of 60 seeds and reported the equivalent performance of a sample with 810 seeds using object-wise spectrum. However, this technique requires a great deal of data processing, as there is a significant increase in the amount of information (i.e. equivalent to the number of pixels). Moreover, it also needs to be explored in different situations (e.g. species, cultivars, crops).

Crop-type application

Out of the 44 articles evaluated, approximately 80% performed an analysis of agricultural crops species (e.g. soybean, maize, wheat), 11% of horticultural seeds, 5% of fruit production and 5% of other classes (pasture and medicinal plants), while there was no work on forest seeds (Fig. 2).

As with the present study, Rahman and Cho (2016), in a narrative review with 32 papers that applied seed variety identification using image analysis techniques, 31 focused on agricultural crops. Given the emerging feature of the spectral imaging analysis technique in seed phenotyping, the use of it in agricultural crop seeds over other seeds may be mainly linked to the economic appeal of these species, as well as to the greater number of plant breeding programmes related to them.

Wavelength spectrum

Of the evaluated studies, 75% used hyperspectral equipment, and according to the density plot, the frequency distribution of the wavelengths applied in the studies varies according to the type of equipment (Fig. 3). Commercial hyperspectral equipment operates in bands with greater amplitude in the NIR spectrum (750–2500 nm), whereas commercial multispectral imaging equipment concentrates on the visible light range up to the beginning of the NIR (350–950 nm).

The wavelength range used is closely linked with the components measured in the seeds, and the visible spectrum is related

Table 4. Characteristics and applications of spectral imaging (HIS – hyperspectral imaging; MSI – multispectral imaging) applied in seed phenotyping

Species	Method	No. of wavelengths	Spectrum	Application	No. of groups	Total seeds	Training (Tr), testing(Te) and validation (V) proportion	Best classifier	Extra features	Wavelength (WL) selection/ dimensionality reduction	Spectral preprocessing	Accuracy	References
Alfafa	MSI	19	365–970	Varietal discrimination	12	2400	70% Tr; 30% Te	SVM	Morphological; colour	–	–	93.47%	Yang et al. (2020)
Cotton	HIS	256	1100–2500	Varietal discrimination	4	807	2:1:1	PLS-DA	–	–	Smoothing	98.00%	Carreiro Soares et al. (2016)
Cotton	HIS	200	942–1646	Varietal discrimination	7	13,160	3:1:1	CNN-SoftMax	–	–	Smoothing; normalization	88.84%	Zhu et al. (2019b)
Grape	HIS	240	914–1715	Varietal discrimination	4	56	60% Tr; 40% V	GDA	–	–	–	100.00%	Rodríguez-Pulido et al. (2013)
Grape	HIS	200	975–1646	Varietal discrimination	3	43,357	2:1	SVM	–	10 WL (PCA)	Smoothing	88.70%	Zhao et al. (2018a)
Jatropha curcas	HIS	256	874–1734	Origin discrimination	4	240	2:1	LS-SVM	Morphological	10 WL (SPA)	–	93.75%	Gao et al. (2013)
Looffah	HIS	200	975–1645	Varietal discrimination	6	4128	2:1	DCNN	–	–	Smoothing	95.93%	Nie et al. (2019)
Maize	HIS	649	1110–2500	Varietal discrimination	4	80	1:1	SIMCA	–	PCA	Smoothing; first derivative; normalization	97.50%	Jia et al. (2015)
Maize	HIS	380	400–1000	Varietal discrimination	3	376	70% Tr; 30% Te	LS-SVM	–	–	Detrending	91.67%	Wang et al. (2016)
Maize	HIS	94	400–1000	Varietal discrimination	4	2000	2:1	LS-SVM	–	–	–	94.40%	Huang et al. (2016a)
Maize	HIS	94	400–1000	Varietal discrimination	4	2000	2:1	LS-SVM	–	–	–	98.30%	He et al. (2016)
Maize	HIS	233	400–1000	Varietal discrimination	17	1632	3:1	LS-SVM	Morphological	11 WL (SPA); PCA	Normalization	92.65%	Huang et al. (2016b)
Maize	MSI	19	375–970	Haploid discrimination	2	240	1:1	CDA	Fluorescence excitation/ emission	–	–	85.83%	De La Fuente et al. (2017)
Maize	HIS	94	400–1000	Varietal discrimination	4	3600	5:1; 5:2; 5:2; 5:4	LS-SVM	–	–	–	85.40%	Guo et al. (2017)
Maize	HIS	219	924–1657	Varietal discrimination	14	1120	3:1	LS-SVM	–	19 WL (JSWSA)	–	96.57%	Yang et al. (2017)
Maize	HIS	256	862.9–1704.2	Haploid discrimination	2	200	2:1	BPR	–	BULDP	Smoothing; first derivative; normalization	99.85%	Wang et al. (2018)
Maize	HIS	200	975–1646	Varietal discrimination	3	12,900	2:1	RBFNN	–	15 WL (PCA)	Smoothing	91.00%	Zhao et al. (2018b)
Maize	HIS	233	400–1000	Varietal discrimination	17	1632	1:1	LS-SVM	Texture	10 WL (MLDA)	Normalization	99.13%	Xia et al. (2019a,b)
Maize	HIS	200	975–1646	Varietal discrimination	3	5400	4:1:1	LR	–	–	–	96.67%	Zhang et al. (2020)
Maize	HIS	200	975–1646		4	20,400	2:1	RBFNN	–	–	Smoothing	98.09%	Bai et al. (2020)

(*Continued*)

Table 4. (*Continued.*)

Species	Method	No. of wavelengths	Spectrum	Application	No. of groups	Total seeds	Training (Tr), testing(Te) and validation (V) proportion	Best classifier	Extra features	Wavelength (WL) selection/ dimensionality reduction	Spectral preprocessing	Accuracy	References
				Varietal discrimination									
Maize	HIS	420	450–979	Varietal discrimination	4	3200	6:2:2	DCNN	–	–	–	95.30%	Zhang et al. (2021)
Maize (silage maize)	HIS	200	975–1646	Varietal discrimination	4	20,400	2:1	RBFNN	–	–	Smoothing	99.10%	Bai et al. (2020)
Maize (sweet maize)	HIS	700	480–1020	Varietal discrimination	9	810	4:1	SVM	–	23 WL (CARS)	Smoothing; first derivative	94.86%	Zhou et al. (2020b)
Maize and silage maize	HIS	200	975–1646	Varietal discrimination	8	40,800	2:1	RBFNN	–	–	Smoothing	88.41%	Bai et al. (2020)
Maize and silage maize	HIS	200	975–1646	Varietal discrimination	2	40,800	2:1	RBFNN	–	–	Smoothing	88.41%	Bai et al. (2020)
Maize waxy maize	HIS	220	430–972	Varietal discrimination	8	800	4:1	FDA	–	t-SNE	Procrustes analysis (PA)	97.50%	Miao et al. (2018)
Oat	HIS	200	975–1646	Varietal discrimination	4	14,846	3:1	DCNN	–	–	Smoothing	99.19%	Wu et al. (2019)
Okra	HIS	200	975–1645	Varietal discrimination	6	6136	2:1	DCNN	–	–	Smoothing	98.24%	Nie et al. (2019)
Pepper	MSI	19	365–970	Varietal discrimination	3	4416	9:1	SVM	–	–	–	97.70%	Li et al. (2020a,b)
Rice	HIS	256	1039–1612	Varietal discrimination	4	225	2:1	RF	–	–	First derivative	100.00%	Kong et al. (2013)
Rice	MSI	19	365–970	Transgenic; non-transgenic	2	400		LS-SVM	Morphological	–	–	100.00%	Liu et al. (2014b)
Rice	MSI	19	365–970	Varietal discrimination	5	250	4:1	LS-SVM	Morphological; colour	–	–	94.00%	Liu et al. (2016a,b)
Rice	MSI	19	365–970	Varietal discrimination	20	600	–	k-NN + multiclass CDA	Morphological; colour	–	–	93.00%	Hansen et al. (2016)
Rice	HIS	256	874.41–1733.91	Mutant discrimination	2	2640	2:1	ELM	–	–	Smoothing	91.75%	Feng et al. (2017)
Rice	HIS	256	975–1646	Varietal discrimination	4	20,907	3:2	CNN	–	–	Smoothing	87.00%	Qiu et al. (2018)
Rice	HIS	256	385–1000	Varietal discrimination	90	8640	4:1	RF	Morphological	85 WL (LDA)	Normalization	79.64%	Fabiyi et al. (2020)
Soybean	MSI	19	365–970	Hybrid; progenitors	3	600		BPNN	Morphological	–	–	98.00%	Liu et al. (2014a)
Soybean	HIS	128	373–1043	Varietal discrimination	10	1200		EL	–	CARS	MSC	100.00%	Zhu et al. (2019a)
Soybean	HIS	200	975–1646	Varietal discrimination	3	5670		CNN	Pixel-wise	–	Smoothing; normalization	98.78%	Zhu et al. (2019c)
Soybean	HIS	128	400–1000		10	1200	3:1:1	GS-SVM	–	–	First derivative	97.20%	Zhu et al. (2020)

				Varietal discrimination									
Soybean	HIS	462	400–1000	Varietal discrimination	15	750	2:1	RSLD	–	155 WL (CCM)	–	99.20%	Wei et al. (2020)
Sunflower	MSI	-	450–1550	Varietal discrimination	6	12,000	-	DLJ4	Morphological; texture	–	–	98.20%	Bantan et al. (2020)
Tomato	MSI	19	365–970	Varietal discrimination	11	2525	3:1	PLS-DA	Morphological; colour	–	–	79.00%	Shrestha et al. (2015)
Tomato	MSI	19	365–970	Varietal discrimination	5	1236		SVM-DA	–	–	SNV; detrending	98.00%	Shrestha et al. (2016b)
Tomato	HIS	288	950–2500	Varietal discrimination	4	1366		PLS-DA	–	–	Smoothing; detrending	71.00%	Shrestha et al. (2016a)
Wheat	MSI	19	365–970	Varietal discrimination	7	1728	-	k-NN	Morphological; colour, texture	–	–	69.53%	Vrešak et al. (2016)
Wheat	HIS	200	975–1660	Varietal discrimination	5	33,494	9:1	SVM	–	–	–	87.81%	Bao et al. (2019)
Wheat	HIS	200	975–1645	Varietal discrimination	30	147,096	2:1:1	CNN-ATT	–	–	Normalization	93.10%	Zhou et al. (2020a)

to superficial characteristics, such as pigmentation (e.g. flavonoids, carotenoids, chlorophyll) and oxidation. These characteristics are ideal for distinguishing seeds with marked physical characteristics, for example, tegument colour or texture. As regards the NIR spectrum, this region is sensitive to the molecular overtone of hydrogen-containing groups, such as C–H, N–H, O–H chemical bonds, which represent seed starch, protein and oil contents, and can penetrate deeper than visible light through the subsurface layer of seed coat (Rodríguez-Pulido et al., 2013; Li et al., 2014, 2020a; Mortensen et al., 2021). For works that used hyperspectral cameras, there was a peak near the 1000 nm range. In this range, the 1122, 1200 and 1314 nm bands (related to organic C–H compounds, such as starch) stand out, while the 1402 nm wavelength is associated with the O–H region of carboxylic acids, as well as regions near the 1580 nm band (Osborne and Douglas, 1981; Lammertyn et al., 1998; Serranti et al., 2013; Zhao et al., 2014).

Shrestha et al. (2016a), using hyperspectral image analysis in the NIR region for the classification of four tomato seed varieties, found that the 1417, 1901, 2102 and 2238 nm bands, associated with protein and water content, and the 1222 and 1695 nm bands, associated with fatty acid content, represent an important spectral signature for this species. Rodríguez-Pulido et al. (2013), when separating seeds of three grape cultivars, with one coming from two different regions, using the principal component analysis (PCA) score, found that the bands at 928, 940, 1148, 1620 and 1652 nm, referring to organic compounds with C–H chemical bonds, were primarily responsible for distinguishing the seeds. Zhao et al. (2018b), based on the score of the first six principal components of PCA, selected the bands in the 1100 and 1390 nm region and the bands at 1436, 1453 and 1554 nm (with the latter three corresponding to the first overtone of O–H stretching, to classify grape seeds of three cultivars). Zhang et al. (2021), using multispectral equipment and classifying four maize cultivars, found that the wavelength bands with the greatest contribution to the distinction of cultivars were 450–700 nm, related to the chlorophyll and β-carotene content of the endosperm, 730 and 785 nm, related to organic compounds with O–H and N–H bonding, and 850–950 nm, related to C–H hydrocarbons. Huang et al. (2016b) found 92.65% accuracy when they classified 17 corn cultivars, using 11 wavelengths selected by the Successive Projection Algorithm (SPA), located in the 500–750 nm region, which are sensitive to seed starch and oil contents. Similarly, Xia et al. (2019a) classified 17 corn cultivars based on 10 optimal wavelengths, belonging to the regions of 410–470 nm, 524–790 and the wavelength of 988 nm, which represent seed texture, starch and oil content, and water content, respectively.

Thus, since the NIR region is sensitive to organic compounds in seeds in deeper layers than visible light, this region seems to be a good strategy to differentiate seeds with similar surface characteristics (i.e. where the visible spectrum region acts more intensely) (Williams and Norris, 2001; Rodríguez-Pulido et al., 2013). For example, Wang et al. (2018) separated haploid from diploid maize seeds, whose visual similarity makes it difficult to separate them by traditional or machine vision methods. Using hyperspectral image analysis in the NIR region (860–1700 nm), they were able to identify differences in oil content and other organic components, differentiating the seeds with 99.85% accuracy.

As for the equipment, the basic difference between multispectral and hyperspectral devices is in the number of wavelengths

https://doi.org/10.1017/S0960258523000028 Published online by Cambridge University Press

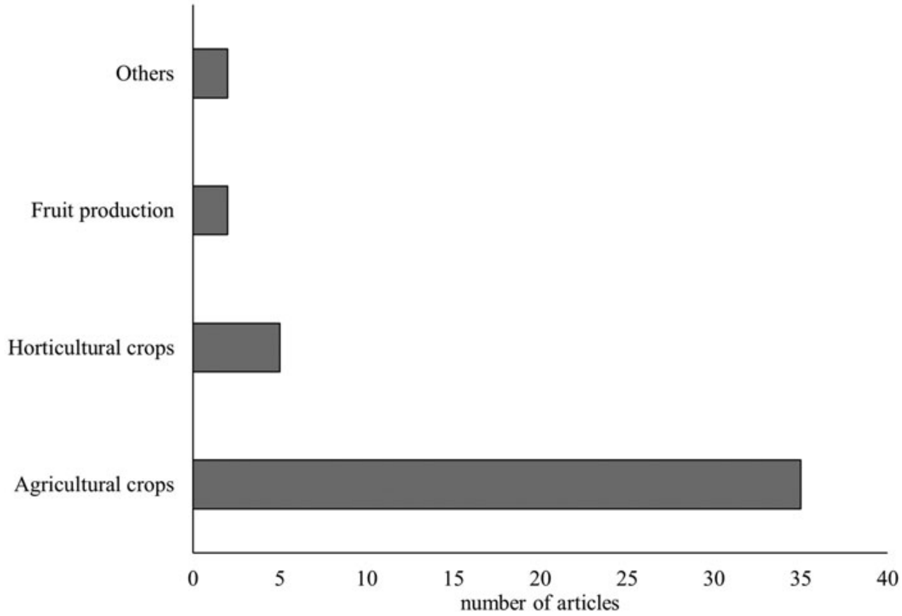

Fig. 2. Species used in each article of the review.

that each one can measure. Multispectral equipment measures up to 20 wavelengths, while hyperspectral cameras can reach higher values, as reported in the work of Zhou et al. (2020b) with sweet corn and 700 wavelengths measured. The use of hyperspectral equipment results in a larger amount of data and, consequently, more time for processing and development of the classifier models. Therefore, all the studies that performed some form of wavelength selection or dimensionality reduction used hyperspectral equipment. Dimensionality reduction aims to mitigate the problem of correlation between predictor variables and model overfitting (Wu et al., 2019; Amigo, 2020). For example, Gao et al. (2013) used SPA to reduce from 256 to 10 wavelengths and obtained 93.75% accuracy.

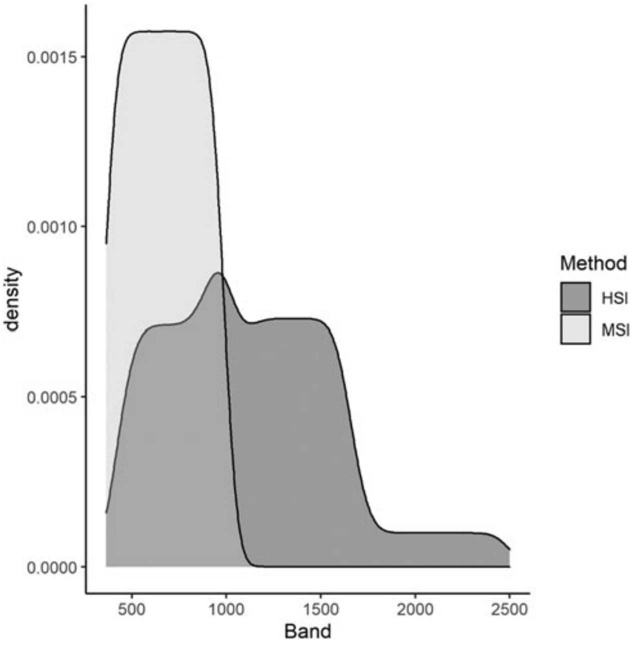

Fig. 3. Density plot of the wavelengths used according to the hyperspectral (HIS) or multispectral (MSI) method.

Classifiers

For seed classification based on the selected wavelength and other features, the evaluated papers used machine learning and deep learning class algorithms on 30 and 17 occasions, respectively; in 2017, the percentage of papers that used machine learning was 95% (Fig. 4).

In 2018 and later, the number of papers using deep learning not only increased but was proportionally higher than the number of papers using machine learning. Deep learning is an unsupervised classification method (the class of seeds is not previously provided to the algorithm) and brings the advantage of identifying abstract patterns in a large amount of data that supervised methods would not be able to find (i.e. deep features) (Gheisari et al., 2017; Wu et al., 2019). However, to achieve successful classification, deep learning algorithms preferentially need a larger volume of data, and this is represented in the average number of seeds used in the evaluated papers that applied machine learning: 3897, compared to 22,765 in deep learning.

The larger the amount of data, the greater the demand for technology and processing time, which may be linked to the low frequency of use of deep learning in previous studies. In contrast to processing time, this class of algorithms seems to be more advantageous in seed classification as highlighted by Zhu et al. (2020), who found that all tested deep learning algorithms had higher accuracy than machine learning algorithms. Similarly, Qiu et al. (2018), comparing the deep learning algorithm CNN with the machine learning algorithms SVM and K-nearest neighbour (K-NN), found that, as the training samples increased, the CNN model outperformed the others. Nie et al. (2019), when classifying hybrid okra and loofah seeds using the deep learning model deep CNN (DCNN) and comparing it to partial least-square discriminant analysis (PLS-DA) and SVM, found that the number of varieties increased from two to six. The authors reported that with increased complexity (number of varieties), the accuracy of the DCNN model remains more stable than that of the others.

Thus, the main advantage of using deep learning algorithms lies in their ability to integrate the steps of feature learning, feature extraction, dimensionality reduction and classification into just one system, which brings greater convenience in the use of data with more complexity (i.e. a larger number of features), as occurs when hyperspectral images are used (Wu et al., 2019).

Chemometric features

The adjusted model, which was identified by the stepwise algorithm with the lowest AIC (−07.9015), used the following features: number of seeds, seed classification groups and use of methods for wavelength selection and/or data dimensionality reduction. Only the first feature was significant (Table 5). According to the estimated and exponentialized coefficient (to reverse the logarithmic scale) of the number of groups (0.998), as the number of seed classification groups increases, the final model accuracy tends to decrease by approximately 0.2% with each new group.

Classification accuracy tends to naturally decrease as the number of possible groups increases, but the non-significant influence of the other factors is due to the fact that the technique can result in high classification models using different strategies in the process (e.g. method, feature selection, preprocessing), that is, an isolated factor is not enough to determine the accuracy of an analysis. It must be clear that the final model only indicates a possible relationship between the variables, since other factors not listed may be relevant to determine classification accuracy (e.g. species, seed quality); moreover, further research is needed to make a robust analysis.

Discussion

Overall strengths

In the 44 evaluated studies, it was clear that the information collected through spectral image analysis, both reflectance and biometric measures of morphology and texture, are sufficient to classify seeds of different genotypes. Although one needs to further explore the ability to generalize the use of the analysis between seeds from other regions and/or harvests, as well as make different combinations of genotypes in future work, the fact is that well-fitted classification models have high accuracy in several situations: between cultivars, hybrids and progenitors, and hybrids and lines; transgenic and non-transgenic seeds.

Spectral image analysis allows the separation of genotypes even in coated seeds, mainly by using the NIR, which penetrates beyond the surface of the seed. Coated seeds are common in the industry, as coating provides protection against fungi and microorganisms, and aids germination by supplying nutrients and amino acids, among other benefits. In the case of these treated seeds, even when dyes are applied for identification, classification using spectral image analysis has still proved possible (Jia et al., 2015; Zhang et al., 2020). However, for small and/or non-uniform seeds, which are coated with thicker layers (which occurs by the encrustation or pelleting process), the analysis may not be applicable.

Reflectance offers sufficient information for the separation of the seeds of different genotypes, and spectral information can be collected quickly, through an image or set of images according to the number of bands measured; seeds remain intact and there is no need for prior treatment. Therefore, multispectral image analysis has a huge advantage over conventional tests, because its limitations refer to treatment and processing of data rather than data collection. Traditional methods, such as molecular markers, are indeed highly reliable, robust methods; thus, they can hardly be replaced. However, in routine work in seed analysis laboratories, when identifying hybrid seeds in breeding programmes or in identifying cultivar mixtures in purity testing, traditional methods are not necessary if there is an alternative way that is reliable, fast and agile enough to meet the industry's demands (Shrestha et al., 2015; Bao et al., 2019; Zhang et al., 2020).

Challenges and limitations

Phenotypic variation

The main challenge of the spectral image analysis technique applied to seed phenotyping surely lies in the extrapolation of the fitted model to seeds coming from other harvests/regions (Zhu et al., 2020). The development of a prediction model is somewhat difficult and laborious and requires professionals

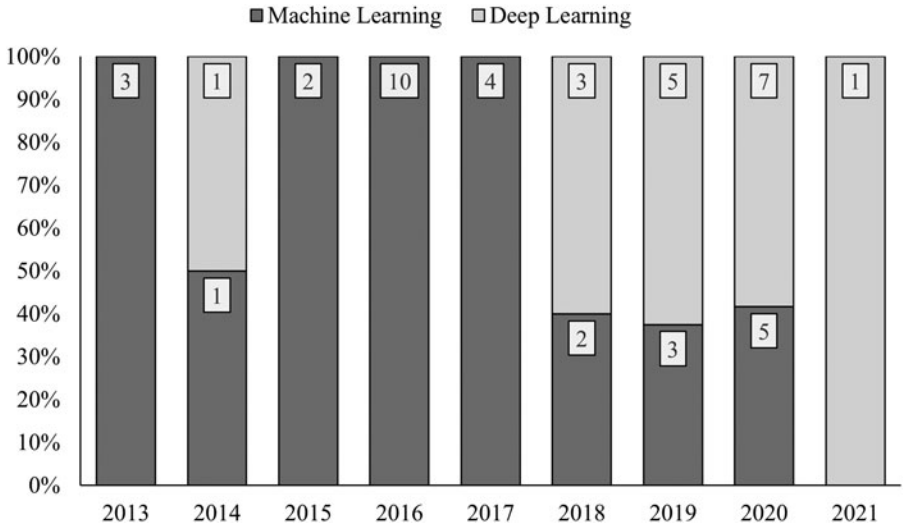

Fig. 4. Proportion and absolute quantity (number in square) of the classification algorithm classes used in each paper over the years.

Table 5. Coefficients estimated from the adjusted model identified by the stepwise algorithm with the features that may influence the accuracy of the spectral imaging analysis of the evaluated studies

Coefficients	Estimate (log scale)	Std. Error	*t*	*P*-value
Intercept	−0.0643	0.0138	−4.651	3.02×10^{-5}
No. of groups	−0.0020	0.0009	−2.263	0.0286
WL Selection1[a]	0.0378	0.0256	1.476	0.1471

[a] Papers that used wavelength selection or dimensionality reduction procedure to obtain the most accurate model.

specialized in data analysis. In the process, seeds need to be used to train the proposed model, test different forms of data processing and validate the model with new seeds. This is time-consuming and not easily adapted to the work routine in seed-producing companies. Thus, the model created to classify produced cultivars must be able to classify seeds over the years and also those grown in different regions (He et al., 2016).

When it comes to biological data, there is great variation among seeds from different years and regions, since morphophysiological characteristics are highly affected by climate, soil, parent plant characteristics, among other factors. Importantly, the biometric data obtained by spectral image analysis are sensitive to morphophysiological characteristics (e.g. pigmentation and organic compounds); therefore, variations in the characteristics of cultivars obtained in other harvests may be enough to misclassify them (Shrestha et al., 2016a,b).

The intensity of phenotypic influence varies with species and the characteristics used for the classification of cultivars. For some species, there are cultivars with outstanding characteristics that facilitate differentiation, as is the case with the tegument colour of bean seeds or peanut seeds. However, in many species, with subtle differences among cultivars, the influence on phenotypic variation can be a problem. For example, in some maize cultivars, the balance between sugar and amylopectin, which can be used as a spectral marker, can be affected by variations in water content in the seed from different regions, which can influence classification (Wang et al., 2016).

An alternative would be to use seeds from cultivars from other years and/or regions to train the classification model, but this poses some difficulties. The first is to obtain a sufficient number of seeds from other years and/or regions, since a large volume of samples is needed to overcome the effect of location. In addition, the use of seeds from a seed bank or archive samples, which were stored in different periods, could influence the classification of newly harvested seeds (Shrestha et al., 2016a,b).

Some studies suggest the use of model updating, which seems to be a promising alternative. In this method, the original model, previously prepared using seeds from the same harvest, is updated with seeds from the following harvests, in order to have a more accurate model, without the need to perform the whole process of adjusting a new model. Some studies show a 10–35% increase in overall accuracy when classifying cultivars coming from other years, when compared to a non-updated model (Guo et al., 2017; He et al., 2016; Huang et al., 2016a). Such a practice would partly solve the model portability problem; however, most model updating methods need to be updated with previously classified seeds, which requires time for sampling and classification. An alternative to updating the model is through semi-supervised classification algorithms that use the pre-label approach, that is, they use the original model to classify a new seed sample (from another year) based on retraining the model with seeds classified with a high degree of confidence. However, these methods still need to be evaluated for different species and situations (Guo et al., 2017).

Seed coat influence

The external structures of the seeds can cause a great influence on the analysis, either by using the visible spectrum, which is sensitive to their surface characteristics, or by using the NIR spectrum, which can penetrate the subsurface layer and is sensitive to the organic compounds present in these structures, for example, the bands of 1180 and 1470 nm, which are sensitive to the presence of lignin and fibre, commonly present in the seed coat of several species. Thus, among species whose external structures are the same, as occurs in palea and lemma in hybrid and self-pollinating cereals, the influence of these structures can be a limiting factor (Blackwell et al., 1977; Gao et al., 2013; Feng et al., 2017; Caporaso et al., 2021).

Thus, when it comes to the differentiation of genotypes that present external structures without enough distinguishing characteristics, the use of spectral imaging is not the best choice, since it would merely describe the composition of these structures. For this type of seed, processing would be necessary, but the commercial use of spectral image analysis requires processing, which leads to extra costs and is time-consuming. In addition, the removal of the husk in cereal seeds limits their use of them, since the husk has an important protective function against fungi and insects (Abebe et al., 2004; Mortensen et al., 2021).

Seed orientation

The area exposed to the analysis may influence classification accuracy, given the sensitivity of the spectra used by the analysis to seed surface and subsurface compounds. The influence of orientation was reported in work using models fitted with measurements obtained from corn caryopses with the embryo facing up and the face facing down. The endosperm and the embryo have different compositions; therefore, choosing between the face of the caryopsis that has both structures (i.e. the face in which the embryo is facing up) or the face with only the endosperm can influence one's ability to distinguish different genotypes. The influence of orientation can vary across genotypes, and in the case of differentiation between cultivars, there was an average variation of 10% (Miao et al., 2018; Tang et al., 2020). Sorting seeds with a certain orientation is a laborious job; therefore, when seed orientation plays a role but does not impair genotype differentiation, loss of accuracy can be accepted.

However, when the difference between genotypes is found in the embryo, as occurs between haploid and diploid maize caryopses, and separation is performed in breeding programmes of the species for different purposes, seed orientation is essential. Thus, seed orientation can be a limiting factor when it comes to haploid seed identification if there is no processing before the analysis is performed; however, on a large scale, processing the seeds may be impractical (De La Fuente et al., 2017; Wang et al., 2018).

https://doi.org/10.1017/S0960258523000028 Published online by Cambridge University Press

Specificity

Unlike other techniques, such as molecular markers, in which we can state with great certainty that a seed belongs to a particular genotype from a segment of its genetic code, in spectral image analysis we cannot use reflectance as an absolute marker of species/cultivar, because it is influenced by seed composition. In other words, to identify an unknown seed, it must always come from a sample in which all the possible classifications are known, which were previously defined considering lot-specific issues (e.g. harvest, year, storage). This is not a major limitation in a seed-producing company, where the cultivars produced are known, but in situations when one must identify possible seeds and/or adulterants from an unknown sample, the use of spectral image analysis is impractical.

Adulterant genotypes can be identified in a seed lot when this genotype is commonly used in the trade of adulterated species. In this case, spectral image analysis tends to have good applicability, as the marked difference among species allows a model fitted to a particular variety or crop of the adulterated species to be distinguished with some ease from the adulterant species, as reported by Faqeerzada et al. (2020) in separating seeds of two varieties of almonds from adulterant apricot seeds.

Perspectives

Open database and key wavelength

There has been increased interest in sharing the data collected through spectral image analysis – be it reflectance or biometric data regarding morphological characteristics of the seeds of the species used in the experiments (e.g. diameter, texture) – through online repositories. Data sharing can leverage the use of the analysis by directly allowing researchers to (1) test different chemometric techniques (e.g. preprocessing, classifier algorithms) on real data, without the need to perform a new experiment, and (2) more accurately identify key bands in certain species and cultivars when comparing different experiments.

The identification of key bands would help in the transition from hyperspectral equipment to multispectral equipment with more accurate and relevant bands in seed phenotyping. Hyperspectral equipment can measure many bands. However, many of these bands contain redundant or unnecessary information for the classification of most species; in addition, hyperspectral equipment is very expensive and more difficult to handle, since the reflectance of the various wavelengths is usually obtained by the point-by-point or the line-by-line system, in which the object moves and reflectance is obtained for every pixel or line of pixels at a time, making the process more time-consuming (Jaillais et al., 2015; Zhou et al., 2020a).

Thus, the migration to multispectral equipment seems to be the most obvious trend, since it requires fewer wavelengths that are applied to all pixels of the image at once, and it is more agile and suitable for application in the seed industry, especially in sectors that work with large numbers of cultivars and lots. A fast identification system is essential, especially in real time, and multispectral equipment is ideal for this purpose (Elmasry et al., 2019).

However, in order to efficiently develop multispectral equipment with key wavelengths, a deeper understanding is needed for the interaction of the different wavelengths with the organic compounds of the different evaluated genotypes. To this end, an open database would facilitate such understanding (Elmasry et al., 2019). Some technologies greatly benefit from an open database, for example, to share data from Raman spectrometry and X-ray diffraction, which can be combined to identify different materials (Mendili et al., 2019). Naturally, when it comes to seeds, external factors have a great influence on the analysis (e.g. environment, parent plant) and consequently on their ability to be distinguished. However, with a large amount of data, one can identify relevant patterns between genotypes and at least direct the development of equipment, even if specific to certain species, to obtain a system capable of providing sufficient information for decision-making in accepting or rejecting a seed lot, which would save a great deal of time and money (Elmasry et al., 2019; Xia et al., 2019a,b).

Field of application

One of the areas where spectral image analysis presents great potential is in breeding programmes, especially in the production of hybrid seeds. Differentiating hybrid seeds from seeds generated by unwanted pollination, either from their parents or from self-pollination, is indispensable. This means differentiating between a few classes from samples with high genetic purity and coming from areas with production control and, thus with less variability among seeds, which is ideal for applying spectral analysis (Nie et al., 2019).

Forest species, for example, have a great lack of quality control methods. For species with great economic importance, such as the species of the genus *Eucalyptus* spp., the use of seeds is especially important in breeding programmes for the production of hybrids. The correct hybridization must be confirmed, given the difficulties of controlling pollination, either in indoor orchards or in the field. Thus, spectral image analysis has great potential to meet this need and bring great advances to forest improvement programmes (Ribeiro-Oliveira and Ranal, 2014).

Another relevant point that makes spectral image analysis an important tool in breeding programmes is that images show individual morphological features of seeds, since the analysis allows the collection of both reflectance and spatial biometric data. The use of morphological features is especially important to check the homogeneity of seed morphological descriptors, because morphology is an attribute relatively unaffected by environmental issues and could be used to evaluate the genetic quality of a lot, which decreases with every generation (Mortensen et al., 2021).

Online and real-time sorting systems

Probably the most promising aspect of spectral image analysis is the possibility of integration with an online system that allows real-time estimation of the quality of a seed lot. According to the International Seed Testing Association (ISTA, 2020), a certain amount of mixing of other cultivars is allowed in a seed lot. This is evaluated through purity analysis; however, there is great difficulty in determining the presence of other cultivars mixed in a lot in certain species, since the analysis depends on the analysts experience and their ability to identify cultivars by eye (Elmasry et al., 2019).

Although each company presents a specific situation (i.e. different combinations of genotypes, number of genotypes, presence of different years and/or regions) and it is not clear to what extent spectral image analysis can handle these different situations, the fact is that in the studies identified in the present review, the analysis was effective. This means that, at least in certain situations, the analysis could be integrated into a system to estimate the

https://doi.org/10.1017/S0960258523000028 Published online by Cambridge University Press

genetic quality of a seed lot, since the alternative way (i.e. through purity analysis) is extremely laborious and, in many situations, impractical (Elmasry et al., 2019).

Some researchers, such as Faqeerzada et al. (2020), reported the feasibility of an online system for real-time classification of seeds moved by a conveyor belt, in which the classification model previously adjusted using hyperspectral images in the infrared region was transferred to an online system. However, some problems still need to be overcome; for example, the speed of the conveyor belt, the variation in light, the overlapping of the seeds on the belt, among other points described by the authors.

Several studies have shown that the models developed from spectral information of seeds are robust enough for large-scale application of the analysis in real-time seed phenotyping through the design of classification maps. In this way, the seeds are determined in real time as to the probability of belonging to a certain class based on the colour scale stipulated for each class. This approach enables decision-making by the analyst and would act as a powerful tool to differentiate cultivars that would hardly be identified with the naked eye (Wang et al., 2016; Zhao et al., 2018a, b; Zhang et al., 2021).

Conclusions

The present review evaluated 44 papers that applied spectral image analysis in seed phenotyping; they were selected among 1304 papers identified in the main journal databases. The review sought to identify the main characteristics of the experiments described in the published papers, as well as to guide researchers in the choice of strategies for experimental design and data analysis, since there are many ways to obtain a highly accurate classification model. Thus, after analysing the papers, the following points summarize the main findings:

- All the evaluated studies presented satisfactory final accuracy; however, few used test data, as well as test and/or validation lots, which may have contributed to the high accuracy reported.
- As the application of multispectral analysis is relatively new in seed phenotyping, the works are still focused on agricultural species with a greater economic appeal.
- Most studies have focused on the use of hyperspectral equipment, which works mainly in the NIR region and is sensitive to seed organic compounds and able to penetrate the subsurface layer. The use of the NIR region seems to be a good strategy to identify differences between genotypes with similar surface structures, where visible light acts with greater intensity.
- The use of deep learning algorithms has been a trend in recent years, mostly because of its ability to work with more complex data, for example, data collected by using hyperspectral cameras.
- Reflectance and biometric data on seed morphology provide sufficient information to separate different genotypes in several situations: among cultivars; hybrids and progenitors; and hybrids and lines, as well as in the separation of coated seeds.
- The main challenge of the analysis is certainly the phenotypic variation of the seeds, which implies the difficulty of using the adjusted model in the classification of cultivars from other harvest, years and/or regions. The main limitations refer to the sensitivity of reflectance to seed compounds, which are highly influenced by environmental issues; the influence of seed coat on the classification of genotypes with similar external characteristics and the influence of seed orientation when the information needed for classification is on a certain face of the seed (e.g. face with the embryo).

Thus, the present review allowed a critical analysis of the use of spectral imaging in seed phenotyping, as well as a thorough evaluation of the limitations of this method. The practical application of this technique needs to be developed for use in laboratories with large volumes of analyses, lots, genotypes and harvests. However, research has been accelerated to overcome the practical challenges of this method, as seen in work using model update algorithms, online classification systems, real-time classification maps; also, spectral information of genotypes is being shared through online repositories. Thus, there are strong indications that the application of multispectral image analysis will become a part of the routine of seed analysis laboratories.

Acknowledgements. The authors thank the Coordination for the Improvement of Higher Education Personnel (CAPES) for granting the scholarships.

Conflicts of interest. None declared.

References

Abebe T, Skadsen RW and Kaeppler HF (2004) Cloning and identification of highly expressed genes in barley lemma and palea. *Crop Science* **44**, 942–950.

Amigo J (2020) *Data handling in science and technology: hyperspectral imaging*. Amsterdam, Elsevier.

Bai X, Zhang C, Xiao Q, He Y and Bao Y (2020) Application of near-infrared hyperspectral imaging to identify a variety of silage maize seeds and common maize seeds. *RSC Advances* **10**, 11707–11715.

Bantan RAR, Ali A, Naeem S, Jamal F, Elgarhy M and Chesneau C (2020) Discrimination of sunflower seeds using multispectral and texture dataset in combination with region selection and supervised classification methods. *Chaos* **30**, 113142.

Bao Y, Mi C, Wu N, Liu F and He Y (2019) Rapid classification of wheat grain varieties using hyperspectral imaging and chemometrics. *Applied Sciences (Switzerland)* **9**, 4119.

Blackwell J, Cael JJ and Koenig JL (1977) Infrared and Raman-spectroscopy of cellulose. *American Chemical Society* **13**, 206–218.

Boelt B, Shrestha S, Salimi Z, Jørgensen J, Nocolaisen M and Cartensen JM (2018) Multispectral imaging – a new tool in seed quality assessment? *Seed Science Research* **28**, 222–228.

Caporaso N, Whitworth MB and Fisk ID (2021) Total lipid prediction in single intact cocoa beans by hyperspectral chemical imaging. *Food Chemistry* **344**, 128663.

Carreiro Soares SF, Medeiros EP, Pasquini C, De Lelis Morello C, Harrop Galvão RK and Ugulino Araújo MC (2016) Classification of individual cotton seeds with respect to variety using near-infrared hyperspectral imaging. *Analytical Methods* **8**, 8498–8505.

Dearden P, Kowalski B, Lowe J, Roland R, Surridge M, Thomas S and Jones S (2011) *Mendeley reference manager*. London, UK, Mendeley Support Team.

De La Fuente GN, Carstensen JM, Edberg MA and Lü bberstedt T (2017) Discrimination of haploid and diploid maize kernels via multispectral imaging. *Plant Breeding* **136**, 50–60.

Elmasry G, Mandour N, Al-Rajaie S, Belin E and Rousseau D (2019) Recent applications of multispectral imaging in seed phenotyping and quality monitoring — an overview. *Sensors* **19**, 1–32.

Fabiyi SD, Vu H, Tachtatzis C, Murray P, Harle D, Dao TK, Andonovic I, Ren J and Marshall S (2020) Varietal classification of rice seeds using RGB and hyperspectral images. *IEEE ACCESS* **8**, 22493–22505.

https://doi.org/10.1017/S0960258523000028 Published online by Cambridge University Press

Faqeerzada MA, Perez M, Lohumi S, Lee H, Kim G, Wakholi C, Joshi R and Cho BK (2020) Online application of a hyperspectral imaging system for the sorting of adulterated almonds. *Applied Sciences (Switzerland)* **10**, 1–16.

Feng X, Peng C, Chen Y, Liu X, Feng X and He Y (2017) Discrimination of CRISPR/Cas9-induced mutants of rice seeds using near-infrared hyperspectral imaging. *Scientific Reports* **7**, 15934.

Gao J, Li X, Zhu F and He Y (2013) Application of hyperspectral imaging technology to discriminate different geographical origins of *Jatropha curcas* L. seeds. *Computers and Electronics in Agriculture* **99**, 186–193.

Gheisari M, Wang G and Bhuiyan MDZA (2017) A survey on deep learning in big data *in* IEEE international conference on computational science and engineering (CSE) and IEEE international conference on embedded and ubiquitous computing (EUC), Guangzhou, China.

Guo D, Zhu Q, Huang M, Guo Y and Qin J (2017) Model updating for the classification of different varieties of maize seeds from different years by hyperspectral imaging coupled with a pre-labeling method. *Computers and Electronics in Agriculture* **142**, 1–8.

Hansen MAE, Hay FR and Carstensen JM (2016) A virtual seed file: the use of multispectral image analysis in the management of genebank seed accessions. *Plant Genetic Resources: Characterisation and Utilisation* **14**, 238–241.

Hastie T, Tibshirani R and Friedman J (2017) *The elements of statistical learning: data, inference, and prediction* (12th edn). New York, Springer Publishing.

He C, Zhu Q, Huang M and Mendoza F (2016) Model updating of hyperspectral imaging data for variety discrimination of maize seeds harvested in different years by clustering algorithm. *Transactions of the ASABE* **59**, 1529–1537.

Huang M, Tang J, Yang B and Zhu Q (2016a) Classification of maize seeds of different years based on hyperspectral imaging and model updating. *Computers and Electronics in Agriculture* **122**, 139–145.

Huang M, He C, Zhu Q and Qin J (2016b) Maize seed variety classification using the integration of spectral and image features combined with feature transformation based on hyperspectral imaging. *Applied Sciences (Switzerland)* **6**, 183.

ISTA (2020) *The international seed testing association*. Wallisellen, International Rules for Seed Testing.

Jaillais B, Roumet P, Pinson-Gadais L and Bertrand D (2015) Detection of Fusarium head blight contamination in wheat kernels by multivariate imaging. *Food Control* **54**, 250–258.

Jia S, An D, Liu Z, Gu J, Li S, Zhang X, Zhu D, Guo T and Yan Y (2015) Variety identification method of coated maize seeds based on near-infrared spectroscopy and chemometrics. *Journal of Cereal Science* **63**, 21–26.

Kong W, Zhang C, Liu F, Nie P and He Y (2013) Rice seed cultivar identification using near-infrared hyperspectral imaging and multivariate data analysis. *Sensors (Basel Switzerland)* **13**, 8916–8927.

Lammertyn J, Nicolai B, Ooms K, De Smedt V and De Baerdemaeker J (1998) Non-destructive measurement of acidity soluble solids and firmness of Jonagold apples using NIR spectroscopy. *Transactions of ASAE* **41**, 1089–1094.

Li L, Zhang Q and Huang D (2014) A review of imaging techniques for plant phenotyping. *Sensors (Switzerland)* **14**, 20078–20111.

Li H, Jiang D, Cao J and Zhang D (2020a) Near-infrared spectroscopy coupled chemometric algorithms for rapid origin identification and lipid content detection of *Pinus koraiensis* seeds. *Sensors (Switzerland)* **20**, 1–17.

Li X, Fan X, Lili Zhao Huang S, He Y and Suo X (2020b) Discrimination of pepper seed varieties by multispectral imaging combined with machine learning. *Applied Engineering in Agriculture* **36**, 743–749.

Liu C, Liu W, Lu X, Chen W, Chen F, Yang J and Zheng L (2014a) Non-destructive discrimination of conventional and glyphosate-resistant soybean seeds and their hybrid descendants using multispectral imaging and chemometric methods. *Journal of Agricultural Science* **154**, 1–12.

Liu C, Liu W, Lu X, Chen W, Yang J and Zheng L (2014b) Nondestructive determination of transgenic *Bacillus thuringiensis* rice seeds (*Oryza sativa* L.) using multispectral imaging and chemometric methods. *Food Chemistry* **153**, 87–93.

Liu W, Liu C, Hu X, Yang J and Zheng L (2016a) Application of terahertz spectroscopy imaging for discrimination of transgenic rice seeds with chemometrics. *Food Chemistry* **210**, 415–421.

Liu W, Liu C, Ma F, Lu X, Yang J and Zheng L (2016b) Online variety discrimination of rice seeds using multispectral imaging and chemometric methods. *Journal of Applied Spectroscopy* **82**, 993–999.

Mendili YE, Vaitkus A, Merkys A, Grazulis S, Chateigner D, Mathevet F, Gascoin S, Petit S, Bardeau JF, Zanatta M, Secchi M, Mariotto G, Kumar A, Cassetta M, Lutterotti L, Borovin E, Orberger B, Simon P, Hehlen B and Le Guen M (2019) Raman open database: first interconnected Raman–X-ray diffraction open-access resource for material identification. *Journal of Applied Crystallographic* **52**, 618–625.

Miao A, Zhuang J, Tang Y, He Y, Chu X and Luo S (2018) Hyperspectral image-based variety classification of waxy Maize seeds by the t-SNE model and procrustes analysis. *Sensors (Switzerland)* **18**, 11–14.

Moher D, Liberati A, Tetzlaff J and Altman DG (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine* **6**, 1–6.

Mortensen AK, Gislum R, Jørgensen JR and Boelt B (2021) The use of multispectral imaging and single seed and bulk near-infrared spectroscopy to characterize seed covering structures: methods and applications in seed testing and research. *Agriculture* **11**, 1–18.

Nie P, Zhang J, Feng X, Yu C and He Y (2019) Classification of hybrid seeds using near-infrared hyperspectral imaging technology combined with deep learning. *Sensors and Actuators B: Chemical* **296**, 126630.

Osborne BG and Douglas S (1981) Measurement of the degree of starch damage in flour by near infrared reflectance analysis. *Journal of the Science of Food and Agriculture* **32**, 328–332.

Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald SmcGuiness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P and McKenzie JE (2021) PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BJM* **372**, 1–36.

Qiu Z, Chen J, Zhao Y, Zhu S, He Y and Zhang C (2018) Variety identification of single rice seed using hyperspectral imaging combined with convolutional neural network. *Applied Sciences (Switzerland)* **8**, 1–12.

Rahman A and Cho BK (2016) Assessment of seed quality using nondestructive measurement techniques: a review. *Seed Science Research* **26**, 285–305.

Ribeiro-Oliveira JP and Ranal MA (2014) Brazilian forest seeds: a precarious beginning, a heady present and the future, will it be promising? *Ciência Florestal* **24**, 771–784.

Rodríguez-Pulido FJ, Barbin DF, Sun DW, Gordillo B, González-Miret ML and Heredia FJ (2013) Grape seed characterization by NIR hyperspectral imaging. *Postharvest Biology and Technology* **76**, 74–82.

Serranti S, Cesare D, Marini F and Bonifazi G (2013) Classification of oat and groat kernels using NIR hyperspectral imaging. *Talanta* **103**, 276–284.

Shrestha S, Deleuran LC, Olesen MH and Gislum R (2015) Use of multispectral imaging in varietal identification of tomato. *Sensors* **15**, 4496–4512.

Shrestha S, Knapič M, Žibrat U, Deleuran LC and Gislum R (2016a) Single seed near-infrared hyperspectral imaging in determining tomato (*Solanum lycopersicum* L.) seed quality in association with multivariate data analysis. *Sensors and Actuators B: Chemical* **237**, 1027–1034.

Shrestha S, Deleuran LC and Gislum R (2016b) Classification of different tomato seed cultivars by multispectral visible-near infrared spectroscopy and chemometrics. *Journal of Spectral Imaging* **5**, a1.

Tang Y, Cheng Z, Miao A, Zhuang J, Hou C, He Y, Chu X and Luo S (2020) Evaluation of cultivar identification performance using feature expressions and classification algorithms on optical images of sweet corn seeds. *Agronomy* **10**, 1268.

Vrešak M, Olesen MH, Gislum R, Bavec F and Jørgensen JR (2016) The use of image-spectroscopy technology as a diagnostic method for seed health testing and variety identification. *PLoS One* **11**, 1–10.

Wang L, Sun DW, Pu H and Zhu Z (2016) Application of hyperspectral imaging to discriminate the variety of maize seeds. *Food Analytical Methods* **9**, 225–234.

Wang Y, Lv Y, Liu H, Wei Y, Zhang J, An D and Wu J (2018) Identification of maize haploid kernels based on hyperspectral imaging technology. *Computers and Electronics in Agriculture* **153**, 188–195.

https://doi.org/10.1017/S0960258523000028 Published online by Cambridge University Press

Wei Y, Li X, Pan X and Li L (2020) Nondestructive classification of soybean seed varieties by hyperspectral imaging and ensemble machine learning algorithms. *Sensors (Switzerland)* **20**, 1–12.

Williams P and Norris K (2001) *Near-infrared technology: in the agricultural and food industries* (2nd edn). Saint Paul, Minessota, American Association of Cereal Chemists.

Wu N, Zhang Y, Na R, Mi C, Zhu S, He Y and Zhang C (2019) Variety identification of oat seeds using hyperspectral imaging: investigating the representation ability of deep convolutional neural network. *RSC Advances* **9**, 12635–12644.

Xia C, Yang S, Huang M, Zhu Q, Guo Y and Qin J (2019a) Maize seed classification using hyperspectral image coupled with multi-linear discriminant analysis. *Infrared Physics and Technology* **103**, 103077.

Xia Y, Xu Y, Li J, Zhang C and Fan S (2019b) Recent advances in emerging techniques for non-destructive detection of seed viability: a review. *Artificial Intelligence in Agriculture* **1**, 35–47.

Yang S, Zhu QB, Huang M and Qin JW (2017) Hyperspectral image-based variety discrimination of maize seeds by using a multi-model strategy coupled with unsupervised joint skewness-based wavelength selection algorithm. *Food Analytical Methods* **10**, 424–433.

Yang L, Zhang Z and Hu X (2020) Cultivar discrimination of single alfalfa (*Medicago sativa* l.) seed via multispectral imaging combined with multivariate analysis. *Sensors (Switzerland)* **20**, 1–14.

Zhang C, Zhao Y, Yan T, Bai X, Xiao Q, Gao P, Li M, Huang W, Bao Y, He Y and Liu F (2020) Application of near-infrared hyperspectral imaging for variety identification of coated maize kernels with deep learning. *Infrared Physics and Technology* **111**, 103550.

Zhang J, Dai L and Cheng F (2021) Corn seed variety classification based on hyperspectral reflectance imaging and deep convolutional neural network. *Journal of Food Measurement and Characterization* **15**, 484–494.

Zhao H, Guo B, Wei Y and Zhang B (2014) Effects of grown origin genotype harvest year and their interactions of wheat kernels on near infrared spectral fingerprints for geographical traceability. *Food Chemistry* **152**, 316–322.

Zhao Y, Zhang C, Zhu S, Gao P, Feng L and He Y (2018a) Non-destructive and rapid variety discrimination and visualization of single grape seed using near-infrared hyperspectral imaging technique and multivariate analysis. *Molecules* **23**, 1352.

Zhao Y, Zhu S, Zhang C, Feng X, Feng L and He Y (2018b) Application of hyperspectral imaging and chemometrics for variety classification of maize seeds. *RSC Advances* **8**, 1337–1345.

Zhou L, Zhang C, Taha MF, Wei X, He Y, Qiu Z and Liu Y (2020a) Wheat kernel variety identification based on a large near-infrared spectral dataset and a novel deep learning-based feature selection method. *Frontiers in Plant Science* **11**. doi:10.3389/fpls.2020.575810.

Zhou Q, Huang W, Fan S, Zhao F, Liang D and Tian X (2020b) Non-destructive discrimination of the variety of sweet maize seeds based on hyperspectral image coupled with wavelength selection algorithm. *Infrared Physics and Technology* **109**, 103418.

Zhu S, Chao M, Zhang J, Xu X, Song P, Zhang J and Huang Z (2019a) Identification of soybean seed varieties based on hyperspectral imaging technology. *Sensors (Switzerland)* **19**, 5225.

Zhu S, Zhou L, Gao P, Bao Y, He Y and Feng L (2019b) Near-infrared hyperspectral imaging combined with deep learning to identify cotton seed varieties. *Molecules* **24**, 3268.

Zhu S, Zhou L, Zhang C, Bao Y, Wu B, Chu H, Yu Y, He Y and Feng L (2019c) Identification of soybean varieties using hyperspectral imaging coupled with convolutional neural network. *Sensors (Switzerland)* **19**, 4065.

Zhu S, Zhang J, Chao M, Xu X, Song P, Zhang J and Huang Z (2020) A rapid and highly efficient method for the identification of soybean seed varieties: hyperspectral images combined with transfer learning. *Molecules* **25**, 152.

https://doi.org/10.1017/S0960258523000028 Published online by Cambridge University Press