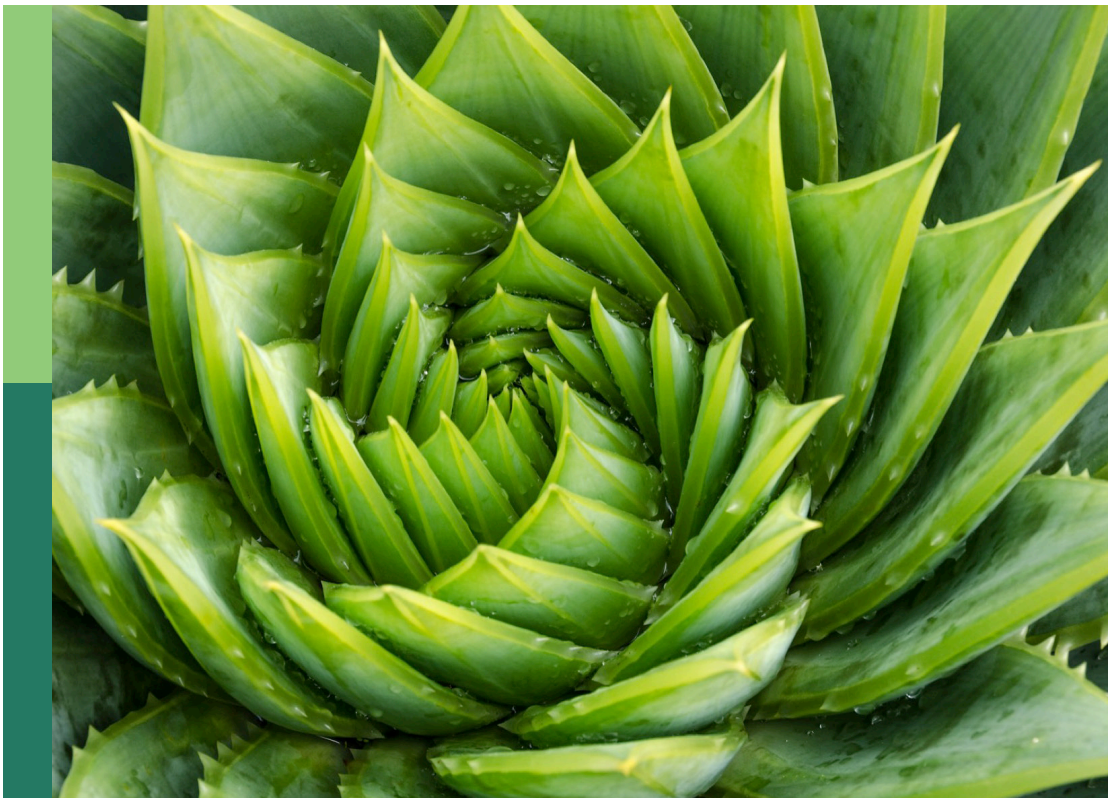
frontiers | Research Topics

Insights in technical advances in plant science 2023

Edited by
Roger Deal

Published in
Frontiers in Plant Science



ISSN 1664-8714
ISBN 978-2-8325-6335-9
DOI 10.3389/978-2-8325-6335-9

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Insights in technical advances in plant science: 2023

Topic editor
Roger Deal — Emory University, United States

Citation
Deal, R., ed. (2025). *Insights in technical advances in plant science: 2023*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-6335-9

Table of contents

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 15 February 2024
DOI 10.3389/fpls.2024.1351958

OPEN ACCESS

EDITED BY
Roger Deal,
Emory University, United States

REVIEWED BY
Satoru Tsuchikawa,
Nagoya University, Japan
Kusumiyati Kusumiyati,
Padjadjaran University, Indonesia

*CORRESPONDENCE
Mario Cunha
mccunha@fc.up.pt

RECEIVED 07 December 2023
ACCEPTED 24 January 2024
PUBLISHED 15 February 2024

CITATION
Tosin R, Cunha M, Monteiro-Silva F, Santos F, Barroso T and Martins R (2024) Bi-directional hyperspectral reconstruction of cherry tomato: diagnosis of internal tissues maturation stage and composition. *Front. Plant Sci.* 15:1351958. doi: 10.3389/fpls.2024.1351958

Bi-directional hyperspectral reconstruction of cherry tomato: diagnosis of internal tissues maturation stage and composition

Renan Tosin[1,2], Mario Cunha[1,2]*, Filipe Monteiro-Silva[2], Filipe Santos[2], Teresa Barroso[2] and Rui Martins[2]

[1]Department of Geosciences, Environment and Spatial Planning, Faculty of Sciences of the University of Porto, Porto, Portugal, [2]INESC TEC - Institute for Systems and Computer Engineering, Technology and Science, Universidade do Porto, Porto, Portugal

Introduction: Precision monitoring maturity in climacteric fruits like tomato is crucial for minimising losses within the food supply chain and enhancing pre- and post-harvest production and utilisation.

Objectives: This paper introduces an approach to analyse the precision maturation of tomato using hyperspectral tomography-like.

Methods: A novel bi-directional spectral reconstruction method is presented, leveraging visible to near-infrared (Vis-NIR) information gathered from tomato spectra and their internal tissues (skin, pulp, and seeds). The study, encompassing 118 tomatoes at various maturation stages, employs a multi-block hierarchical principal component analysis combined with partial least squares for bi-directional reconstruction. The approach involves predicting internal tissue spectra by decomposing the overall tomato spectral information, creating a superset with eight latent variables for each tissue. The reverse process also utilises eight latent variables for reconstructing skin, pulp, and seed spectral data.

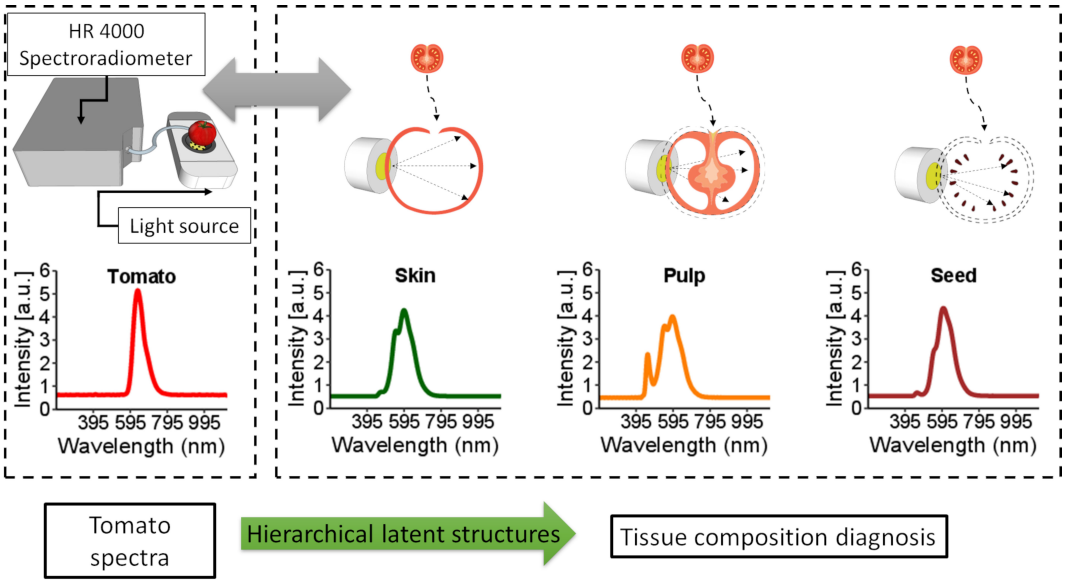
Results: The reconstruction of the tomato spectra presents a mean absolute percentage error of 30.44 % and 5.37 %, 5.25 % and 6.42 % and Pearson's correlation coefficient of 0.85, 0.98, 0.99 and 0.99 for the skin, pulp and seed, respectively. Quality parameters, including soluble solid content (%), chlorophyll (a.u.), lycopene (a.u.), and puncture force (N), were assessed and modelled with PLS with the original and reconstructed datasets, presenting a range of R2 higher than 0.84 in the reconstructed dataset. An empirical demonstration of the tomato maturation in the internal tissues revealed the dynamic of the chlorophyll and lycopene in the different tissues during the maturation process.

10.3389/fpls.2024.1351958

Conclusion: The proposed approach for inner tomato tissue spectral inference is highly reliable, provides early indications and is easy to operate. This study highlights the potential of Vis-NIR devices in precision fruit maturation assessment, surpassing conventional labour-intensive techniques in cost-effectiveness and efficiency. The implications of this advancement extend to various agronomic and food chain applications, promising substantial improvements in monitoring and enhancing fruit quality.

KEYWORDS

fruit maturation, latent structures, precision agriculture, spectral reconstruction, spectroscopy



GRAPHICAL ABSTRACT

1 Introduction

Tomato is a climacteric fresh fruit composed of multiple tissues with diverse physical and biochemical compositions relevant to defining its quality through the food chain. The constitution of these tissues undergoes significant dynamic changes throughout the maturation process and in the post-harvest phase until the consumer such as the levels of antioxidants, lycopene, ascorbic acid, phenols and free radicals (Chandra and Ramalingam, 2011) and bioactive compound [e.g. flavonoids (Tamasi et al., 2019)].

The tomato fruit maturation process is marked by tissue specialisation, which promotes biochemical and physical changes in all tissues (Moco et al., 2007). During the ripening process, the tomato changes colour from green to red, resulting in morphological and biochemical modifications. The cultivar, environmental conditions (e.g., soil, light, temperature) and agronomic practices (e.g., irrigation, fertilisation) are essential factors that contribute to the tomato maturation process. For instance, in diverse regions, variations in antioxidants and phenols have been observed (Chandra et al., 2012). Under salinity conditions, morphological aspects such as size, water content, and colour undergo changes, impacting sugar content and acidity (Pascale et al., 2015). Additionally, genetic factors exhibit divergence even under identical conditions (Toor and Savage, 2005).

Several non-destructive techniques are available in the literature for characterising fruit maturation. However, none of these techniques can provide detailed information about the internal tissues of the fruit while also predicting its biophysical and biochemical characteristics efficiently. Although similar works approached the reconstruction of fruits using other techniques, such as electrical impedance to detect the tomato level of

maturation (Verma et al., 2021), computed tomography in detect bitter pit in apples (Si and Sankaran, 2016) and X-ray in predicting the sugar content in kiwi fruit (Kanno and Kuroyama, 2020) and phenotyping charcteristics of seed in predicting the seed length, width, thickness, and radius of soybean and wheat with an accuracy ranging 80-96% (Liu et al., 2020), some of these techniques are expensive and not expedite for *in situ* measurements. Additionally, these methods are mainly used to identify physical damage, and none of them analysed each tissue individually or presented a qualitative approach (Donis-González et al., 2014). In biomedical science, non-destructive tissue characterisation (human and animal) has been developed through Vis-NIR spectroscopy, particularly for detecting tissue anomalies (Malone et al., 2014; Dahlstrand et al., 2019) and guiding the right incision during surgery (Vo-Dinh et al., 2010; Stelzle et al., 2012), which indicate that similar techniques can be applied to vegetation tissues.

Monitoring tomato precision maturation throughout the food chain is crucial for minimising losses within the food supply chain and improving pre- and post-harvest production and utilisation (Garcia and Barrett, 2006). From the high-tech horticulture point of view, monitoring maturation is essential to improve cultural practices, such as irrigation, fertilisation and canopy management, which are directly related to the pigments, organic acids and sugars in the tomato fruit (Bertin and Génard, 2018). In addition, to ensure that the seeds are fully developed for tomato seed production (Shrestha et al., 2016). They can guarantee that all tissues are well developed and that the gustative parameters favour the final consumer.

Traditional methods like chemical assays and chromatography used to characterise tomato biochemical parameters (e.g., lycopene) in different tissues are, in most cases, destructive, very expensive, non-adapted for small amount of tissues and time-consuming, which hinders the assessment of tomato quality parameters based on the composition of each tissue. Therefore, despite the high cost and time consumption of traditional analysis, alternative non-destructive techniques have been developed to assess tomato maturation and quality parameters, including colourimetric (Gómez et al., 2001), fluorescence (Wu and Wang, 2014; Konagaya et al., 2020), and Vis-NIR techniques (Torres et al., 2015; Zhu et al., 2015).

Vis-NIR devices have demonstrated outstanding potential for estimating fruit's biochemical and biophysical parameters like lycopene and β-carotene (Tilahun et al., 2018), water potential (Tosin et al., 2021) and sugars and acids (Martins et al., 2022). Numerous studies on tomatoes have utilised Vis-NIR techniques to estimate various biochemical and biophysical parameters. These include soluble solid content (SSC) with reported R^2 of 0.87 (Ecarnot et al., 2013), 0.60-0.77 (Torres et al., 2015), and 0.88-0.98 (Ding et al., 2016). The pH levels were determined with a R^2 of 0.80 (Huang et al., 2018a), while colour attributes showed R^2 values ranging from 0.91-0.99 (Ecarnot et al., 2013). Vitamin C content was estimated with a R^2 of 0.67 (Azadshahraki et al., 2018), total acidity with R^2 of 0.66-0.94 (Ding et al., 2016) and R^2 of 0.91-0.98 (Najjar and Abu-Khalaf, 2021), and firmness exhibited R^2 of 0.70-0.72 (Najjar and Abu-Khalaf, 2021). The analysis of β-carotene revealed R^2 values of 0.77-0.88 (Tilahun et al., 2018), phenols showed R^2 values of 0.76-0.98 (Ding et al., 2016), malic acid with R^2 of 0.27-0.42 (Torres et al., 2015), citric acid with R^2 of 0.66-0.94 (Ding et al., 2016) and lycopene with R^2 values ranging from 0.45-0.75 (Clément et al., 2015), 0.73-0.83 (Ciaccheri et al., 2018) and 0.85-0.89 (Tilahun et al., 2018). However, tomato exhibits differences in the structural and biochemical characteristics of different tissues, which leads to significant ramifications in the absorption and scattering of light inside the tomato fruit (Skolik et al., 2019). These complex structures of tomato make it challenging to investigate the inner tissues through Vis-NIR spectroscopy and how they behave during maturation.

Spectral information about the inner tissues of fruits can be acquired using Vis-NIR data (Martins et al., 2023). Martins et al. (2022) demonstrated empirically that different grape tissues influence the whole fruit during the maturation process and that the concentration of pigments changes during maturation in various tissues. Obtaining spectral information of the internal tissues of fruits through Vis-NIR requires appropriate modelling techniques. Principal component analysis (PCA) is one such technique that uses latent variables (LV) to perform an orthogonal transformation of the original dataset onto a reduced subspace that is spanned by the principal components (Dahlstrand et al., 2019). Conversely, the combination of LV creates a superset, presenting a direct relationship with the original information. LV models can deal with significant and correlated variables (Trygg and Wold, 2003). Multi-block hierarchical PCA (HPCA) and hierarchical partial least squares (HPLS) are frequently used in chemometrics to deal with spectral information from different sensors and a batch of data (Mishra et al., 2021; Martins et al., 2023). Therefore, multi-block analysis can create a bi-directional reconstruction of whole tomato fruit from the skin, pulp, and seed spectral data.

Tissue reconstruction is based on hierarchical latent relationships between the spectral patterns of the observed tissues, providing details of the internal plant structures (Martins et al., 2023). This manuscript uses the term "tomography-like" to partly define the capacities of data-driven class reconstruction using hierarchical relationships. The practicality of bidirectionality involves connecting the principal latent space derived from tomato tissues to the entire tomato spectrum and executing the reverse process, which consists of breaking down the tomato fruit's spectra into the spectra of its tissues, namely the skin, pulp, and seed. This methodology was recently applied to grapes, as demonstrated by Tosin et al. (2023), and it is expected to work in tomato, in which the tissue composition is different from grapes. The term tomography-like is debatable in that it only implies the resolution of the tissue image; here, it is aimed to provide a median spectrum of each tissue, given the fruit spectra in a non-destructive way. Once several spectra are taken from different positions on the fruit, a 3D resolved image can be reconstructed by relating the positions [x, y, z] and spectral gradients within internal tissues (Martins et al., 2022; Tosin et al., 2023). In this sense, this work can be considered 'tomography-like'. Data-driven reconstruction does not use the same numerical approaches and solutions as the classical approaches but is entering several research areas due to their computational efficiency (Bar-Sinai et al., 2019; Martins et al., 2022).

Furthermore, spatially resolved tissue fruit composition is yet very complex to be obtained experimentally (Tosin et al., 2023). There are still very significant constraints at the level of analytical chemistry state-of-the-art on the quantity of sample that can be used to quantify parameters considered in today's routine analysis, such as SSC and pigments. Therefore, metabolic or compositional imaging validation is still limited to the laboratory ground truth methods.

This research used a point-of-measurement (POM), where the light enters the fruit and has internal reflections, being only able to return to the spectrometer through a centre fibre optics pinhole, meaning that all light reaching the spectrometer interacts with the inner fruit's tissues, maximising the spectral information on all internal tissues.

In multispectral or hyperspectral imaging, light is generally illuminated outside the fruit. This demonstrates that most light reaching the imaging sensor is reflected, non-absorbed, and carries little information about internal tissue composition.

Therefore, relationships with the recorded spectra using this method are generally limited to covariant information with pigmentation, which has the danger of quantifying through correlation and not due to the causal characteristic features present in the spectra. At present, it is believed that POM devices can be of more practical application to field studies than hyperspectral cameras.

Through multi-block analysis, this research facilitates the reconstruction of hyperspectral data by utilising information from individual spectra of tomato tissues, namely skin, pulp, and seed. This method enables the decomposition of the overall tomato spectrum into its constituent tissues, offering a bi-directional relationship. This study further provides a qualitative analysis of the tomato ripening process, elucidating the maturation levels in the skin, pulp, and seed at various developmental stages. By employing a POM sensing approach that maximises spectral information from internal tissues, the research addresses the limitations of traditional destructive methods, providing a non-destructive alternative for characterising tomato biochemical parameters. This contribution to the high-tech horticulture food supply chain can be used to ensure superior quality produce, reduce waste, enhance market value, and advance agricultural practices.

Therefore, three main objectives have been established in this work: i) to reconstruct the tomato hyperspectral data using information from the skin, pulp, and seed spectra through multi-block analysis; ii) to demonstrate that the spectral information of the entire tomato can be decomposed into the skin, pulp, and seed; and iii) to provide a qualitative analysis of the dynamics of the tomato ripening process, demonstrating the maturation levels in the skin, pulp, and seed, and how these tissues behave at different stages of the maturation process.

2 Materials and methods

2.1 Sampling and tomato properties

A total of 118 cherry tomatoes, freshly picked at several maturation stages, were promptly taken to the laboratory for analysis. Puncture force (N) and SSC (%) were performed after measuring the tomato spectra using a digital penetrometer (model PCE-PTR 200, PCE Group, D-59872 Meschede, Germany), registering the resistance force and maximal force until puncture and a hand refractometer Milwaukee model MR32ATC, with a scale range of SSC from 0 to 32.0%, respectively.

The tomato skin, pulp, and seeds were methodically extracted from each tomato (n = 118) and subjected to individual analysis to obtain their respective spectral records. Then, aliquots measuring approximately 0.5 cm^2 were taken using a lancet and deposited onto a glass microscope slide. The procedure involved peeling the tomato skin, slicing the pulp to a thickness of approximately 3 mm, and directly placing a single well-developed seed, selected from the various seeds present in the tomato, onto the microscope slide. This meticulous process ensured the separation of tomato tissues and allowed the acquisition of specific spectral data for the skin, pulp, and seeds.

The tomato process of maturation progresses from green to red and can be classified into six different colours: i) green, ii) breakers, iii) turning, iv) pink, v) light-red and vi) red (USDA, 1991).

Tomato is a complex fruit regarding internal tissues (Figure 1, Supplementary Figure 1). At the green stage of maturation, the tissues are not well developed, which makes tissue separation hard. Therefore, this work considered the epidermis as the skin, columella, placenta and pericarp as pulp and seeds as seeds (Figure 1). The jelly parenchyma and sepal were not considered in the analysis.

2.2 Spectroscopy

Tomato spectra were recorded with a white LED platform (Supplementary Figure 1). The platform comprises a reflection disk with a power LED (6500K, Philips SpotOn Ultra 69141/31/PH) at the bottom. The spectral range of the LED emits light from 380 nm to 780 nm. Therefore, LED spectra were used as a reference to check measurement and light emission stability. The tomato is placed above the LED, and the measurement is performed by collecting reflectance with a fibre optic probe (Ocean Insite). Skin, pulp, and seeds are placed on the microscope slide centred with the LED, and the reflectance probe also collects light. Spectra were recorded by a high-resolution spectroradiometer (Ocean Insite HR4000), which obtains information from 195.34 nm to 1118.33 nm; the integration time was optimised for each sample to maintain most of the spectra within the linear response.

After collecting all the spectral data, a logarithm multiplicative scattering correction (Martins et al., 2022) was applied to normalise and reduce the noise in the spectral information. The correction is a widely used method that addresses the issue of light scattering, which can distort the spectral signal and lead to inaccuracies in the measurements.

The logarithmic transformation helps to remove this scattering effect and improve the accuracy of the spectral data.

2.3 Hierarchical latent structures reconstruction

Latent structures are spaces obtained by matrix decomposition into their eigenvectors, a new basis where the contained information is projected. Latent structures provide a geometrical

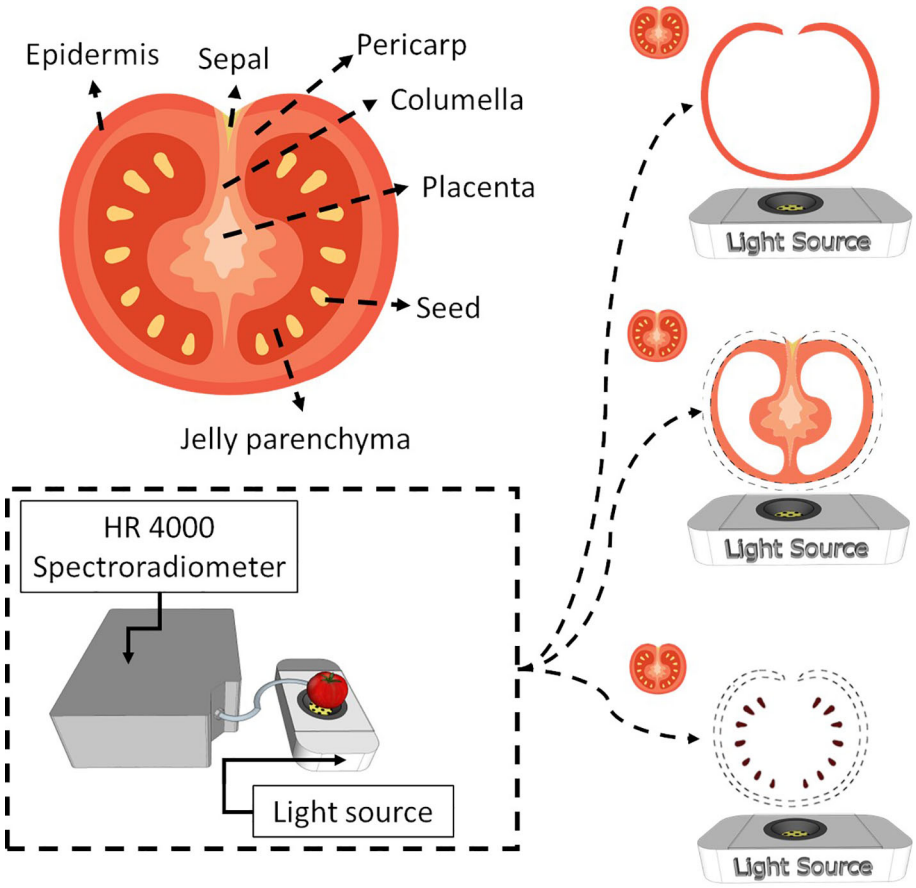


FIGURE 1
Internal tomato tissues and the spectroscopy system used to obtain spectral information from the entire tomato and the respective tissues (skin, pulp and seed).

interpretation of the dataset and its samples by understanding their position on the eigenvector basis. Eigenvectors can be extracted with different properties, but one of the most common decompositions is PCA, where orthogonal eigenvectors are obtained by maximising the dataset variance, allowing to provide the interpretation of relevant variation. PCA can be obtained from the dataset $\mathbf{X}$ by singular value decomposition (SVD) after subtracting the mean of $\mathbf{X}$: $\mathbf{X} = \mathbf{USV}^t$; where $\mathbf{U}$ is the left singular, $\mathbf{S}$ the singular values and $\mathbf{V}^t$ the left singular. In PCA, the scores (aka latent structures) $\mathbf{t}$ are given by $\mathbf{US}$, and the loadings (aka basis/eigenvectors) by $\mathbf{p}^t = \mathbf{V}^t$.

The geometry of information contained in $\mathbf{X}$ can be studied to determine what eigenvectors represent non-random information by performing randomisation tests (Martins et al., 2022), where the spectra dataset reconstruction can be decomposed into $\mathbf{X} = \mathbf{tp}^t + \mathbf{e}$, where $\mathbf{e}$ is random information, irrelevant for spectral reconstruction (Martins et al., 2023).

Let's consider the corresponding database of tissue spectra: skin $\mathbf{X}_1$, pulp $\mathbf{X}_2$ and seeds $\mathbf{X}_3$ and their corresponding relevant PCA decomposition (Algorithm 1):

$$\mathbf{X}_1 = \mathbf{t}_1\mathbf{p}_1^t + \mathbf{e}_1 \quad (1)$$

$$\mathbf{X}_2 = \mathbf{t}_2\mathbf{p}_2^t + \mathbf{e}_2 \quad (2)$$

$$\mathbf{X}_3 = \mathbf{t}_3\mathbf{p}_3^t + \mathbf{e}_3 \quad (3)$$

Where $\mathbf{t}_1$, $\mathbf{t}_2$, $\mathbf{t}_3$ are the relevant latent features that reconstruct the original tissue spectra $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$, and $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ discarded random spectral information. The latent information associated with each tissue can now be fused by determining the relevant common dimensions of their latent variance in geometry along each eigenvector.

Let's take $\mathbf{t}_f = [\mathbf{t}_1^i, \mathbf{t}_2^i, \mathbf{t}_3^i]$ as the concatenation of the i dimension of $\mathbf{t}_1$, $\mathbf{t}_2$, $\mathbf{t}_3$, to be fused into a superset latent space $\mathbf{T}$ by finding the relevant information of each sub-level. The superset latent space can be determined by:

$$\mathbf{t}_f = \mathbf{T}_i\mathbf{P}_i^t \quad (4)$$

Being $\mathbf{T}_i$ the superset latent information of the i dimension of the subsets (tissue spectra), and $\mathbf{P}_i^t$ provide the contribution of each subset to the fused information $\mathbf{T}_i$.

If the information of $[\mathbf{t}_1^i, \mathbf{t}_2^i, \mathbf{t}_3^i]$ has the same direction, $\mathbf{T}_i$ will be described by a single eigenvector $\mathbf{P}_i^t$ or a single dimension; otherwise, further relevant dimensions are added to $\mathbf{T}_i$.

The superset latent structure is constructed for each dimension of $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$, representing the relevant information of the tissue dataset, that is, the relevant characteristics from the skin, pulp, and seeds that relate to the observed tomato spectra.

2.4 Association and bi-directionality

The relevant features extracted from the sub-levels represented in the superset **T** have a similar latent structure to the direct PCA decomposition of the tomato spectra (**Y**). By performing a PCA decomposition to **Y**, it is gotten $\mathbf{Y} = \mathbf{UC}^t$; where a direct association between the latent spaces **T** and **U** is expected ($\mathbf{T} \simeq \mathbf{U}$). One can expect that samples with similar composition and morphological characteristics will generate cluster aggregations in **T** and **U**, reflecting the different skin, pulp, and seed maturation state combinations (Martins et al., 2023).

Therefore, one can establish a direct relationship between neighbouring samples in **T** or **U**, ensuring bi-directionality between the subsets $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$ and **Y** (Figure 2).

Inferring the internal tissues for a given unknown sample is performed by projecting the spectra **Y** into the feature space **U**, by $\mathbf{U} = \mathbf{YC}$, and finding the neighbouring samples (**k**) in this feature space. The **k** can be used to verify its propagation from **T** to the sub-level spaces $\mathbf{t}_1$, $\mathbf{t}_2$ and $\mathbf{t}_3$, by reconstructing $\mathbf{t}_f$ (Algorithm 2, Equation 4).

2.5 Validation

The hierarchical latent structure model was optimised and validated in a two-step approach: i. cross-validation (CV) to optimise the number of principal components (PC) of the sub-spaces $\mathbf{t}_1$, $\mathbf{t}_2$ and $\mathbf{t}_3$ and superspace **T**; and ii. hold-out samples (HO) are used to test predictions and provide quantitative metrics.

CV is a test to the null hypothesis, and HO samples are double-check confirmations of the CV metrics. If the knowledge base is representative, any unknown HO or removed from the dataset, CV should provide statistically similar prediction metrics, proving the null hypothesis. By leaving samples out, CV provides the determination of the optimal error of each tissue reconstruction, $\mathbf{e} = [\mathbf{e}_1 \ \mathbf{e}_2 \ \mathbf{e}_3]$ (Algorithm 3, Equations 1–3). For each sample, the training set, the CV algorithm removes one sample (leave-one-out) for determining the error (**e**) for increasing the number of PCs of sub-space and superset. The optimal number of PCs is considered the one that provides minimal CV errors, preventing over-characterisation of random features at the sub-level passing into the superset. Suppose the training set is representative, CV and HO errors are expected to be similar. In that case, the model can efficiently reproduce the spectral information, and the null hypothesis is verified.

After reconstructing and decomposing the spectral data, a standardisation of the original, reconstructed, and decomposed spectral data was applied to mitigate the effects of signal intensity. Standardisation is a common practice used in data analysis to rescale variables with a mean of zero and a standard deviation of

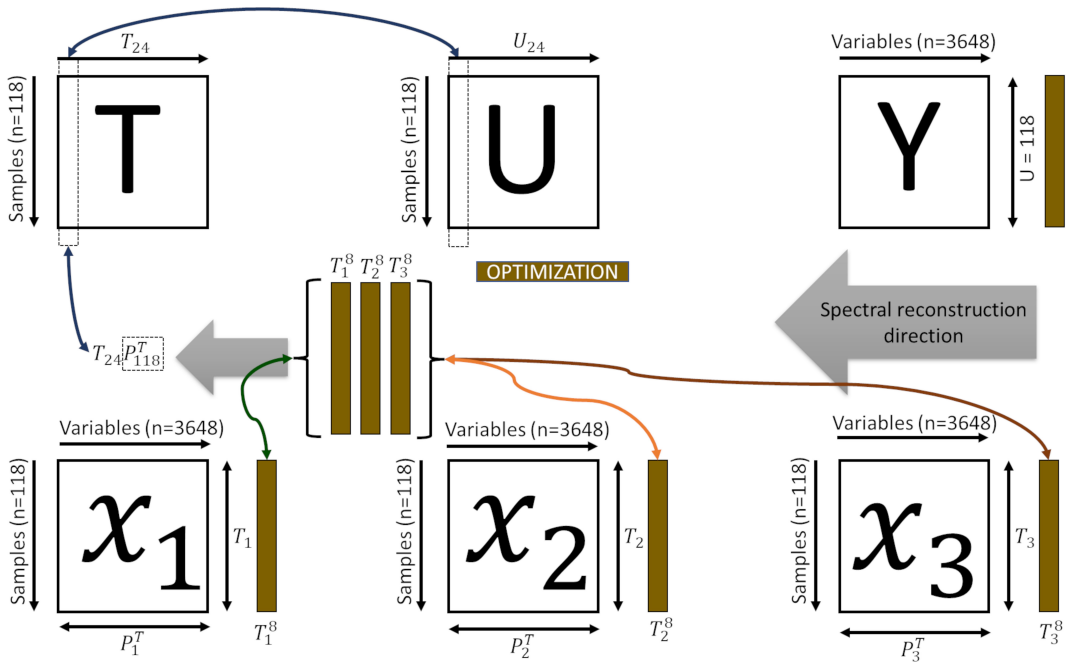


FIGURE 2
Representation of the bi-directional reconstruction process and decomposition of the spectral information. T is the superset latent space; T_{24} the superset latent information of the dimension of the subsets (T_1, T_2, T_3); U is the feature space of each subset; U_8 the superset latent information of the dimension of the subsets Y is the tomato spectra; P_{118}^t is the contribution of each subset (P_1^8, P_2^8, P_3^8); X_1, X_2 and X_3 spectra of the skin, pulp and seed, respectively.

one. In this study, the formula ((**x** – **mean**)/**sd**) was used to standardise the spectral data. This method allows us to compare and analyse the spectral data more accurately, as it removes any differences in signal intensity that could impact the interpretation of the results.

The following metrics were used to benchmark spectral reconstruction: i. mean standard error (MSE) in counts/wavelength (nm), representing the average reconstruction error per wavelength (nm); ii. mean absolute percentage error (MAPE) in % wavelength (nm), representing the average bias per wavelength (nm); iii. Pearson's correlation coefficient and the p-value were extracted to check for significant differences between the original and reconstructed datasets; and iv. Euclidean distances in **T** and **U** of CV and HO samples: This metric measures the knowledge-base representativeness and stability, which is important for evaluating the performance of the spectral reconstruction model.

Require: $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$
Ensure: $j = argmin(\mathbf{X}_i - \mathbf{t}_i\mathbf{p}_i^t)$

$\mathbf{X}_1 = \mathbf{t}_1\mathbf{p}_1^t$

$\mathbf{X}_2 = \mathbf{t}_2\mathbf{p}_2^t$

$\mathbf{X}_3 = \mathbf{t}_3\mathbf{p}_3^t$
While $j > j_{min}$ **do**

$\tau_i \leftarrow [\mathbf{t}_1^i | \mathbf{t}_2^i | \mathbf{t}_3^i]$

$\tau_i = \mathbf{U}_i\mathbf{S}_i\mathbf{V}_i^t$

$\mathbf{T}_i = \mathbf{U}_i\mathbf{S}_i$

$\mathbf{P}_i^t = \mathbf{V}_i^t$
end while
Output: $\mathbf{T} = [\mathbf{T}_1 \dots \mathbf{T}_n]$; $\mathbf{P} = [\mathbf{P}_1 \dots \mathbf{P}_n]$

Algorithm 1. Hierarchical latent structures algorithm.

Require: $\mathbf{Y}$, $\mathbf{T}$
Ensure: $\mathbf{Y} = \mathbf{U}\mathbf{Q}^t$

$\mathbf{U}^k = \mathbf{T}\mathbf{b}$

$\mathbf{T}^k = \mathbf{U}\mathbf{c}$

$\mathbf{Y}^k = (\mathbf{T}\mathbf{b})\mathbf{Q}^t$
Output: $\mathbf{Y}^k$

Algorithm 2. Outer relationships for reconstruction.

Require: $\mathbf{Y}$, $\mathbf{T}$, $\mathbf{P}$, $[\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3]$
Ensure: $j = argmin(\mathbf{X}_i - \mathbf{t}_i\mathbf{p}_i^t)$

$\mathbf{Y} = \mathbf{U}\mathbf{Q}^t$

$\mathbf{T} = \mathbf{U}\mathbf{b}$

$[\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3] = \mathbf{T}_p\mathbf{P}^t$

$\mathbf{X}_1 = \mathbf{t}_1\mathbf{p}_1^t$

$\mathbf{X}_2 = \mathbf{t}_2\mathbf{p}_2^t$

$\mathbf{X}_3 = \mathbf{t}_3\mathbf{p}_3^t$
Output: $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$

Algorithm 3. Latent structures reconstruction algorithm.

2.6 Prediction of tomato quality

As a proof of concept, the organoleptic characteristics of tomatoes were predicted by comparing quantification results obtained from real spectral datasets with those obtained from reconstructed spectra using latent hierarchical structures. The visible range of the spectral data, which falls between 400-700 nm, contains valuable information related to pigments, specifically chlorophyll and lycopene content. Based on the findings of Ciaccheri et al. (2018) and Moco et al. (2007), the inferred chlorophyll content used the ratio of green (520-570 nm) to red (571-700 nm) spectral bands, and lycopene content used the ratio of red to green spectral bands. The results demonstrate that the reconstructed tissue spectra provide a good relationship to the estimates obtained with the fruit spectra, thus serving as proof of the principle of internal tissue quantification. Also, it presents other parameters obtained with the fruit, such as puncture force and SSC, which correlate with the reconstructed tissue spectra, further demonstrating the feasibility of the approach. Although there is a state-of-the-art, optimised ground truth method for measuring fruit composition at the fruit level for small fruits such as grapes (Martins et al., 2022; Tosin et al., 2023), recording that data would not provide significant advantages as it does not allow better tissue resolution quantification than the one presented in this work.

This investigation employed a computational approach to assess tomato pigment content in tissue reconstruction, driven by the need for a time-efficient evaluation of lycopene and chlorophyll. Wet lab analyses commonly used for larger tissue samples were unsuitable due to the small tissue quantities (around 0.5 cm^2) involved in the spectral analysis (Tosin et al., 2023). Routine methods for these analytes require larger tissue quantities, hindering direct comparison with results obtained through computational methods (Clément et al., 2008; Tilahun et al., 2018). Sophisticated analytical methods for small tissue quantities are costly and impractical for evaluating a system at a low technology readiness level (TRL). Therefore, validating the results using expedited and cost-effective methods suitable for assessing the mentioned pigments and potentially other analytes is advisable. Adopting expedited approaches and avoiding expensive wet lab

methods can validate the proof of concept without incurring high costs, facilitating a smoother transition for further system development and refinement.

The study employed a partial least squares (PLS) approach to predict the quality parameters of tomatoes. The dataset comprised 118 samples, divided into two sets: 70% (n=82) for training and 30% (n=36) for validation. This division of the dataset into training and validation sets ensures the utilisation of a significant portion of the data for model development while still allowing for robust evaluation and assessment of the model's performance.

A robust validation technique, leave-one-out cross-validation (LOOCV), was employed to evaluate the model's performance. This approach involved systematically excluding one sample at a time during the evaluation process, allowing for an accurate estimation of the model's predictive ability and mitigating the risk of overfitting.

The determination of the optimal number of LV in PLS model was carried out through an assessment of root mean square error (RMSE) values. This integral step in PLS modelling aimed to minimise the RMSE, underlining its fundamental role in refining the model for superior precision and effectiveness in predicting outcomes. The selection of the ideal number of LV was strategically driven by the overarching goal of achieving the most accurate and reliable results, a chase evident in the search of minimised RMSE values.

Within the data-driven analysis, representativeness and hypothesis-testing principles serve as foundational pillars. Representativeness ensures that the dataset employed for training and validation accurately represents the entire population of interest. Hypothesis testing facilitates the formulation and evaluation of statistical hypotheses, guaranteeing the results' reliability and significance.

By adhering to these principles, it is possible to construct robust models and generate reliable predictions in data-driven analysis.

Benchmarks were performed using the following modelling approaches: i. Similarity(Sim)-Euclidean distance as a metric of the spectral and compositional similarity between neighbouring samples in the feature space (e.g., FaChada et al., 2014); ii. Principal component regression (PCR) - where the latent structures of the sub-level spectra $[\mathbf{t}_1\ \mathbf{t}_2\ \mathbf{t}_3]$, superset $\mathbf{T}$ and tomato spectra $\mathbf{U}$; PLS maximises the covariance between the spectra $\mathbf{X}$ and tomato composition $\mathbf{Y}$ by determining the eigenvectors of $\mathbf{X}^t\mathbf{Y}$ (Martins et al., 2023). This method forces the latent structures of spectra and composition (PLS scores - $\mathbf{U}$) to be equal (NIPALS algorithm) (Ergon, 2009) for the determination of each correspondent basis $\mathbf{U}^t$ and $\mathbf{Q}^t$ (Geladi and Kowalski, 1986). It proceeds with deflation and sequential orthogonal eigenvectors of the remaining information in $\mathbf{X}^t\mathbf{Y}$ (Phatak and De Jong, 1997). The number of deflation or LV is optimised by cross-validation/hold-out samples with minimal predicted sum of squares (PRESS) (Krstajic et al., 2014). PLS uses an oblique projection to determine the $\mathbf{b}_{pls}$ coefficients in $\mathbf{Y} = \mathbf{X}\mathbf{b}_{pls}$ (Phatak and De Jong, 1997; Ergon, 2009).

3 Results

3.1 Tomato tissue reconstruction

Figure 3 presents an application of PCA to investigate the spectral data of tomatoes, including the entire tomato and its internal tissues (skin, pulp, and seeds). Each data point represents a unique spectral measurement obtained from an individual

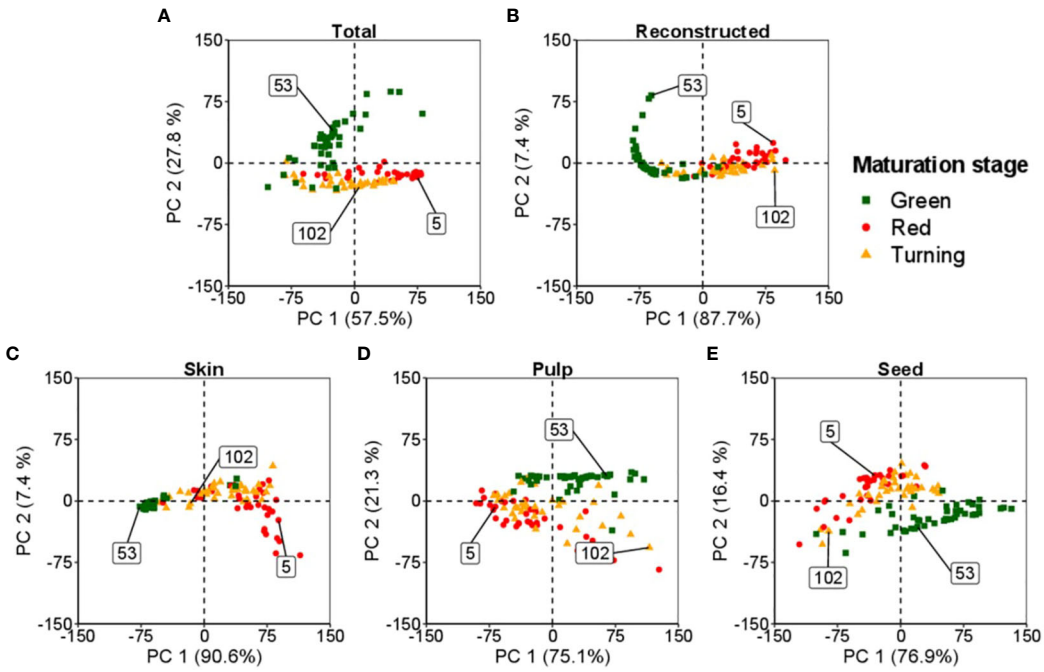


FIGURE 3
Spectral bi-directionality in reconstructing tomato components: skin, pulp, and seeds, at different maturation levels. The numbers in the box indicate the observations from the dataset for each maturation stage: ■ (53) for green, ▲ (102) for turning, and • (5) for red. **(A, B)** represent the feature space for the total and total reconstructed spectra, respectively. **(C–E)** define the feature space of the skin, pulp and seed, respectively.

tomato. The primary objective of PCA is to identify a lower-dimensional representation of the data that captures the most important information.

In this study, the analysis classified the tomatoes into three maturation stages: i) green, ii) turning, and iii) red. A random sample from each maturation stage was selected to represent the feature space of these samples.

Figure 3A illustrates the comprehensive space occupied by the assessed tomatoes. Each tissue of the tomato, namely the skin, pulp, and seeds, possesses its distinct feature space, as illustrated in Figures 3C–E, respectively. PCA enables the decomposition of the space, resulting in individual spaces corresponding to each tomato fraction.

In Figure 3B, the plot demonstrates the reconstructed spectral data of the tomato obtained by combining the information from the skin, pulp, and seeds. The reconstructed data provides a condensed version of the original spectral data while preserving essential characteristics. It is also possible to reverse the process, enabling the data decomposition into the individual feature spaces of the tomato's skin, pulp, and seeds.

The application of PCA in this study enhances the understanding of the tomato's spectral data and internal structures. In addition, it provides insights into the relationships between different tomato fractions and their corresponding spectral properties, shedding light on the distinct characteristics of each tissue within the tomato.

Figure 4 shows the spectral signature of tomato tissues at different stages of maturation: green, middle stage (turning), and ripened (red). The bi-directional spectral reconstruction (Figure 2) works better for internal tomato tissues than for the entire tomato (Figure 4, Table 1). In addition, during the experiments, it was observed that the light had more difficulty passing through the green tomatoes (Figure 4), which is a limitation in obtaining complete information about the internal tissues. It is suggested that at this stage of maturation, the tissues are not fully developed and contain a high concentration of pectin, which interferes with the optical properties of the tomato.

The dynamics of maturation in tomato is shown in Figure 4. Based on these spectral signatures, different band peaks in the spectral signature are suggested to be related to tissue pigments (e.g., chlorophyll and lycopene). For example, green tomato presents higher signal intensity in the bands 460-500 nm and 660-700 nm, suggesting a correlation with chlorophyll content. Likewise, the 500-550 nm range of bands is more related to the carotene group, probably related to lycopene content in the more advanced stages of maturation.

Table 1 presents a benchmark for spectral reconstruction and shows that the spectral reconstruction did not present significative differences (p-value< 0.001) over the original spectral data. However, compared with the respective tissues, the total spectral of tomato shows a higher MSE (0.30) and MAPE (30.4%) and a lower Pearson's correlation coefficient (r=0.85; p-value< 0.001). Three leading causes can explain the low accuracy of the whole tomato: i) the green stage of maturation seems to be a hindrance for the light going through the internal tissues; ii) during the green stage of maturation, the tissues are not fully developed; and iii) the acquisition of the spectral information with an optical probe aimed at finding the best position to obtain the light signal through the internal tissues.

On the other hand, the decomposition of the whole tomato to predict the internal tissues worked better (Figure 4, Table 1). These results suggest that the spectral data of the tomato presented sufficient information on the internal tissues studied in this work.

The lower accuracy in the reconstruction of tomato is discussed in the next section.

3.2 Quality parameters evaluation

The original and reconstructed spectral information were used to predict the quality parameters of the tomato and their respective tissues. Overall analysis showed that the original and reconstructed spectral data were consistent and robust for predicting the quality parameters analysed (Table 2). Additionally, the different tissues of tomato could be used to predict the quality parameters assessed in this experiment. It is important to highlight that for SSC and puncture force, the entire tomato was considered for the measurement, and the different tissues of the tomato could predict these values for the whole tomato fruit. The results presented for chlorophyll and lycopene used an empirical dry lab method based on the spectral information of each tissue and the whole tomato, which helped to infer the pigment concentration in each tissue individually and in the entire tomato. Regression plots of the SSC (%), chlorophyll (a.u.), lycopene (a.u.) and puncture force for the skin, pulp and seed are presented in the Supplementary Materials (Supplementary Figures 2–5).

Figure 5 presents the changes in chlorophyll and lycopene concentrations in different tomato tissues during ripening. As the tomato ripens, chlorophyll concentration decreases while lycopene increases, as demonstrated by the spectral information and dynamic concentration data in Figure 5.

During the early maturation stage (Figure 5A), the tomato skin has a higher concentration of chlorophyll and a lower concentration of lycopene. Likewise, the pulp has a higher concentration of chlorophyll and a lower concentration of lycopene than other tissues of the tomato. This distribution of chlorophyll and lycopene is also reflected in the peaks observed in the tomato spectra. Specifically, peaks in the 460-500 nm and 670-700 nm range are associated with chlorophyll, while those in the 530-560 nm range are associated with carotenes, including lycopene. These wavelength assignments are drawn from the findings of Ciaccheri et al. (2018) and Moco et al. (2007).

During the ripening process, the peaks associated with chlorophyll decrease while those associated with lycopene increase, as shown in Figure 5 for the tomato skin, pulp, and seed. These changes in pigment concentrations are responsible for the observed colour changes in the tomato from green to red as it ripens.

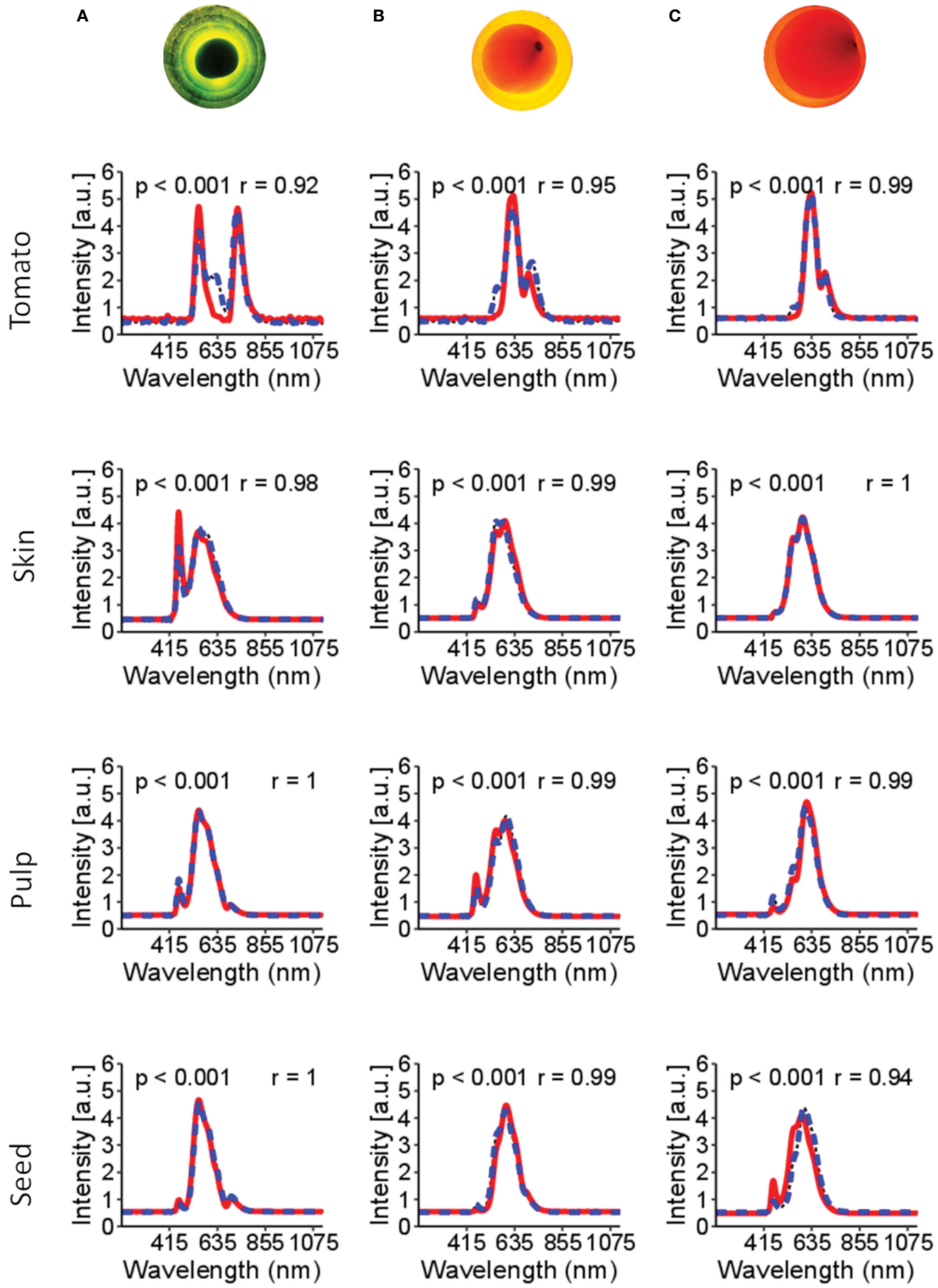


FIGURE 4
Spectral signatures of the tomato and the internal tissues. In the red line (—) is the original spectral signature, and in the blue dashed line (– –) is the reconstructed spectral. Figures **(A–C)** are tomatoes subjected to LED light. Figure **(A)** is a tomato in the green stages of maturation; **(B)** tomato in the turning stages of maturation; **(C)** tomato in the red stage of maturation. The p-value< 0.001 indicates no significant difference between the original and reconstructed spectra. r is Pearson's correlation.

4 Discussion

This paper presents a disruptive methodology for the bi-directional reconstruction of whole tomato hyperspectral data and internal tissues (skin, pulp, and seeds). This non-destructive method can explain how different tomato tissues behave at various stages of the ripening process. As described in Figures 1–3, this work aims to bi-directionally reconstruct the spectral information of the tomato from the data of the skin, pulp, and seed and to decompose the information of the tomato

TABLE 1 Mean square error (MSE), mean absolute percentage error (MAPE), p-value and Pearson's correlation coefficient (r) for the reconstruction of tomato spectra and decomposition of the entire tomato spectra into skin, pulp and seeds spectra.

Reconstruction	MSE(Count/wave-length (nm))	MAPE (%)	p-value	r
Tomato Spectra	0.30	30.44	< 0.001	0.85
Skin Spectra	0.04	5.37	< 0.001	0.98
Pulp Spectra	0.02	5.25	< 0.001	0.99
Seeds Spectra	0.03	6.42	< 0.001	0.99

Mean square error (MSE), mean absolute error (MAPE), and p-value< 0.001 indicate that there is no significant difference between the original and reconstructed spectral matrices.

TABLE 2 Reconstruction benchmark of the tomato, skin, pulp and seed in predicting the soluble solid content (SSC), chlorophyll, lycopene and puncture force.

Property	Metric	Real dataset			
		Tomato	Skin	Pulp	Seeds
SSC (%)	r	0.99	–	–	–
	R^2	0.98	–	–	–
	MSE (%)	0.85	–	–	–
	MAPE (%)	10.78	–	–	–
	LV	3	–	–	–
Puncture force (N)	r	0.98	–	–	–
	R^2	0.97	–	–	–
	MSE (N)	2.94	–	–	–
	MAPE (%)	15.87	–	–	–
	LV	3	–	–	–
Reconstructed Dataset					
SSC (%)	r	0.99	–	–	–
	R^2	0.98	–	–	–
	MSE (%)	0.86	–	–	–
	MAPE (%)	11.12	–	–	–
	LV	3	–	–	–
Chlorophylls (a.u.)*	r	0.95	0.99	0.99	0.99
	R^2	0.90	0.99	0.99	0.99
	MSE (a.u.)	0.49	0.64	0.53	0.51
	MAPE (%)	51.32	6.01	10.06	13.51
	LV	2	2	3	2
Lycopene (a.u.)*	r	0.92	0.99	0.98	0.99
	R^2	0.84	0.99	0.96	0.98
	MSE (a.u.)	0.62	0.64	0.73	0.57
	MAPE (%)	37.68	6.44	11.09	12.91
	LV	2	2	2	2
Puncture force (N)	r	0.90	–	–	–
	R^2	0.95	–	–	–

(Continued)

TABLE 2 Continued

Property	Metric	Real dataset			
		Tomato	Skin	Pulp	Seeds
	MSE (N)	3.84	–	–	–
	MAPE (%)	21.83	–	–	–
	LV	2	–	–	–

*Values computed with the spectral information (nm); the number of latent variables (LV) used in the partial least square (PLS). SSC and puncture force measurements were exclusively conducted for the entire tomato. Prediction based on individual tissues such as skin, pulp, and seeds is deemed impractical, given the integrated nature of these tissues. As a result, a dash (–) is denoted to signify the exclusion of these components in the predictive analysis.
The original dataset exclusively predicted SSC and puncture force for the entire tomato. Chlorophyll and lycopene content were predicted solely in the reconstructed dataset, as these pigments were estimated using real data, rendering their prediction in the original dataset nonsensical.

spectra into the internal tissues. Although each tomato tissue presented a particular space (Figure 3), creating a superset with the scores of each tomato fraction made it possible to reproduce the entire spectral tomato (Figure 4). The most important LV of each fraction formed the superset used to reconstruct the tomato spectra. Through hierarchical PLS, the tissues could predict the entire tomato spectral data. The decomposition of the whole tomato data and the same number of LV combined in the hierarchical PLS can predict the tomato tissues. The literature reports that the tomography-like approach (Martins et al., 2023) successfully worked in grapes (Tosin et al., 2023), and the results of this paper support that it can be applied to aqueous fruits like tomato. Due to the complex nature of tomato maturation and the diverse biochemical compositions of its internal tissues (Skolik et al., 2019), encompassing skin, pulp, and seed, this work presents a technique for the bi-directional spectral reconstruction of tomatoes using Vis-NIR data.

The literature offers several methodologies (Mishra et al., 2021) that could be used for spectral reconstruction. For instance, O2-PLS (Trygg and Wold, 2003) and OnPLS (Lofstedt et al., 2013) utilise spectral data's local and global joints, bioheat models (Alzahrani and Abbas, 2019; Marin et al., 2021) could be adapted to predict the internal tomato tissues and adaptive neuro-fuzzy inference system (Abdullahi et al., 2021; Abdullahi et al., 2022), a hybrid computational model that combines the adaptive capabilities of neural networks with the interpretability of a mathematical framework that deals with uncertainty and imprecision in decision-making. However, these methods are not hierarchical and do not allow for convolution and fusion of information in a superset or deconvolution in a reverse way. Similarly, advanced approaches such as deep learning (DL) can deal with complex data (Mishra et al., 2022) and reconstruct the whole tomato with spectral information from the different tissues. Nevertheless, the decomposition of the tomato spectra into the spectral data of its

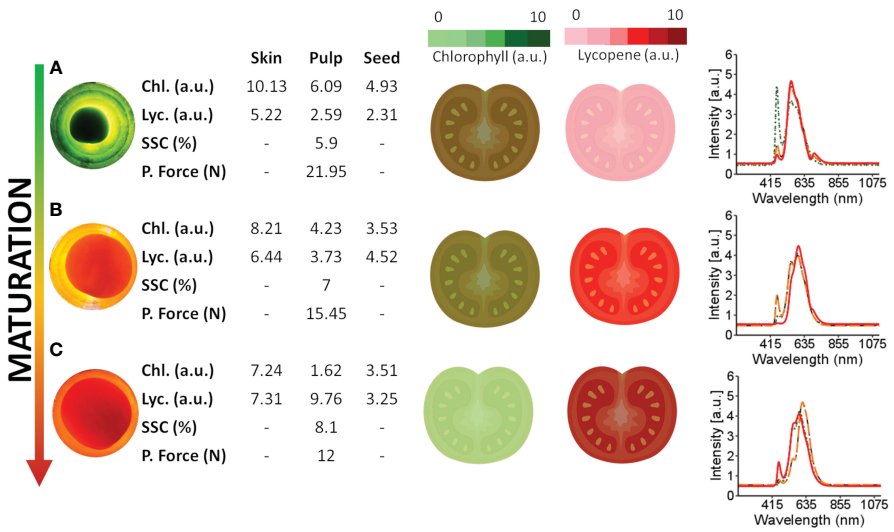


FIGURE 5
Spectral information obtained from tomatoes at different maturation stages and internal tissues and the dynamics of lycopene and chlorophyll during maturation. Chlorophyll content was empirically quantified, with light green representing lower levels and dark green representing higher levels. Lycopene levels were also quantified, with dark red representing lower levels and light red representing higher levels. **(A–C)** are red, turning and green tomatoes over the light source, respectively. The green spectra signature (·····) represents the spectral signature of the skin, the orange spectral signature (– –) represents the spectral signature of the pulp, and the red spectra signature (—) represents the spectral signature of the seed.

tissues becomes even more complex and challenging. Nonetheless, the results presented in Table 1 show that hierarchical PCA combined with PLS can effectively perform bi-directional modelling with the directions $\mathbf{X} - \mathbf{Y}$ and $\mathbf{Y} - \mathbf{X}$, and remove orthogonal data between $\mathbf{Y}$ and $\mathbf{X}$, facilitating the reconstruction of the tomato and its internal tissues.

Owing to the multiple internal tissues that compose the tomato fruit, it faced a challenge in reconstructing the complete spectral information of the tomato (Figure 4). Furthermore, separating and identifying different tissues in green tomato are difficult because green tomato presents more fibre concentration (Chandra and Ramalingam, 2011) and pectin (Moco et al., 2007; Huang et al., 2018b). As a solution, all internal tissues (except the jelly parenchyma) were considered the pulp. Nevertheless, green tomatoes have less interest when compared with more advanced stages of maturation.

Spectral data acquisition encountered a few challenges related to the position of the tomato in the system used to obtain spectral data. First, green tomato is opaque, limiting light's ability to pass through (Figure 4). This limitation requires the optical fibre probe to be well-positioned to obtain more light signals. However, searching for light using the probe may not obtain a signal from all internal structures, which limits tissue reconstruction. A similar effect was observed in matured tomato. Depending on the probe position, some internal tissues may not be assessed, or less information may be obtained. Second, the tissues considered as the pulp will affect the reconstruction of the entire tomato.

During the data acquisition of the entire tomato, almost all the internal tissues are expected to be evaluated using a fibre probe. However, to provide detailed information in the superset utilised to reconstruct the entire tomato, it may be necessary to individually assess each tissue considered as pulp and examine their specific details. Therefore, the errors observed in Table 1 and Figure 4 indicate less accuracy in the reconstruction of the entire green tomato and less in the other tomato fractions of the matured tomato. Nevertheless, this method can reconstruct the Vis-NIR spectra of tomato and internal tissues and be used for different purposes.

This work assessed the SSC, chlorophyll, lycopene, and puncture force of the tomato using the original and reconstructed spectra for each fraction of the tomato (Table 2). The SSC and puncture force were assessed in the entire tomato, and the spectral data of each fraction were considered in the modelling. Chlorophyll and lycopene provided spectral information for each tissue and empirically demonstrated the concentration of each pigment in the tissues. The original and reconstructed spectra results were very similar for reconstructing the entire tomato and the reconstructed spectra (Table 2).

Among the quality parameters assessed, the SSC results were the most stable and robust for the original and reconstructed spectra. Nevertheless, it is essential to highlight that the different tomato tissues present distinct SSC concentrations, and the individual tissues could have been assessed to predict the concentration of SSC in each fraction. For chlorophyll and lycopene, the ratio of the zones of the spectrum empirically demonstrated the different concentrations of pigments in the distinct tissues (Supplementary Figures 2–5). Considering all the tomato tissues, the skin is the part that is indicated to have more concentration of chlorophylls. This study reveals that the skin has the highest concentrations of chlorophyll and lycopene, except during the red maturation stage. These results align with existing literature, particularly Chandra et al. (2012), highlighting the skin's elevated lycopene levels compared to pulp and seeds. However, in terms of chlorophyll content, the skin ranks as the second lowest tissue during maturation, as observed by Moco et al. (2007). The qualitative approach to classifying internal tomato tissues throughout maturation may contribute to these variations compared to the quantitative methods in the cited literature.

Consistent with prior research (Moco et al., 2007; Chandra et al., 2012), this paper reveals that the skin exhibits a higher concentration of lycopene, as shown in Supplementary Figure 3. Puncture force analysis indicates elevated values in green tomatoes, probably attributable to heightened fibre and pectin content. This coincides with lower levels of SSC and lycopene, alongside increased chlorophyll concentrations (Moco et al., 2007; Huang et al., 2018b).

The temporal dynamics of maturation across distinct tomato tissues are demonstrated in Figure 5, where chlorophyll concentrations, particularly in the skin, are higher in green tomatoes (Figure 5A). Conversely, matured tomatoes (Figure 5C) tend to exhibit increased lycopene concentrations, mainly in the skin. Considering the spectral signatures of the tomato fractions (Figure 5), further studies can be conducted to determine the type of information that can be extracted. Empirically assessing the full tomato spectrum (Figure 5), two bands peak near 500 nm and 690 nm in green tomato, probably related to chlorophyll (Ecarnot et al., 2013; Huang et al., 2018b). In the spectral skin signature, a similar peak (near 490 nm) in the green tomato may be related to chlorophyll a. When the pulp and seed were analysed, the same peak (near 490 nm) was present but with less intensity, suggesting a lower chlorophyll concentration in those tissues. The peaks near 550 nm are related to the carotene group (Ciaccheri et al., 2018), especially the lycopene concentration. For matured tomato, these picks present more intensity in the full tomato spectra, skin, and pulp, suggesting a higher concentration of lycopene when compared with less mature tomatoes.

Vis-NIR data can enhance the efficiency and quality of crop production by providing valuable information for optimising various agricultural practices, such as irrigation, fertilisation, and pruning (Xia et al., 2021). Furthermore, by leveraging this data, crop growers can gain precise insights into the maturation process of fruits, which is influenced by a range of biotic and abiotic factors, including diseases, water availability, temperature, and light intensity.

The methodology presented in this paper has the main advantage of providing more accurate and detailed information about the internal structure of the fruit as a tomography-like system when compared to the whole-fruit measurement by using hyperspectral or multispectral data (e.g., Mishra and Woltering, 2023; Mishra et al., 2023) that obtain majority external information. The tomography-like presented in this paper allows for the assessment of individual tissues, facilitating the acquisition of more accurate and detailed information about the internal structure of the fruit (Martins et al., 2023). Figure 3 represents the feature space of the entire tomato and the skin, pulp and seed,

where the different maturation stages occupy distinct feature spaces. In contrast, traditional whole-fruit measurements often lack precision in providing insights into the inner tissue of the fruit. Also, the methodology presented in this paper can lead to more accurate predictions of the internal tissue properties and better quality control. It can also enable a more specific and targeted analysis of internal tissue properties by offering comprehensive and high-dimensional data. These data provide more detailed information about the fruit's internal structure and can also be utilised to determine additional components beyond those presented, particularly in supporting metabolomic studies. This knowledge can help to fine-tune agricultural practices and mitigate potential risks, ultimately leading to improved crop yields and higher-quality produce.

This paper presents a novel technique for determining the quality parameters SSC (%), chlorophyll (a.u.), lycopene (a.u.) and puncture force (N) of fruits using visible and near-infrared (Vis-NIR) spectroscopy of their skin, pulp, and seed. The approach builds upon a growing body of research demonstrating the ability of hyperspectral sensors to measure a wide range of quality parameters non-destructively and accurately in crops (Martins et al., 2022; Tosin et al., 2022; Tosin et al., 2023). Furthermore, by leveraging the high-dimensional data obtained for each tissue, the method showcased in this study has the potential to unlock a multitude of additional quality parameters during fruit maturation. This capacity for rapid and precise determination could enhance fruit production's efficiency and effectiveness while elevating the final product's overall quality. Finally, it is worth noting that the extensive dimensionality of the data obtained for each tissue opens possibilities for identifying and characterising other components beyond those currently presented, particularly in supporting metabolomic studies.

5 Conclusion

This paper proposes a tomography-like system that can predict the Vis-NIR information of the internal tissue. Applying multi-block hierarchical component analysis in conjunction with PLS enables the bi-directional reconstruction of spectral information, facilitating the prediction of internal tissue spectra (skin, pulp, and seed) and the decomposition of the overall tomato spectral information into its constituent tissues.

This novel approach allows assessing tomato maturation dynamics by analysing internal tissue characteristics, offering pertinent information for precision agricultural practices. Moreover, the method can identify physiological issues related to abiotic (e.g., water stress, high temperature) and biotic (e.g., bacterial infection). These identified stressors can be integrated into multifaceted omics techniques to understand the plant's physiological responses.

Building on successful testing in grapes this technique, demonstrates its efficacy in the complex tissue structure of tomato. Thus, the same approach could be applied to other aqueous fruits, such as blueberries. However, further work is necessary to test the applicability of this technique in other fruits, to study the dynamic of the Vis-NIR information with the internal tissues during the maturation process, and to incorporate additional analytical data for validation.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

RT: Investigation, Methodology, Writing – original draft, Writing – review & editing. MC: Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing. FM-S: Methodology, Writing – review & editing. FS: Investigation, Writing – review & editing. TB: Methodology, Writing – review & editing. RM: Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work is financed by National Funds through the FCT – Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project OmicBots – OmicBots: High-Throughput Integrative Omic-Robots Platform for a Next Generation Physiology-based Precision Viticulture, with reference PTDC/ASP-HOR/1338/2021.

Acknowledgments

RT and FM-S acknowledge Fundação para a Ciência e Tecnologia (FCT) PhD research grants Ref. SFRH/BD/145182/2019 and SFRD/BD/09136/2020. RM acknowledges Fundação para a Ciência e Tecnologia (FCT) research contract grant (CEEIND/017801/2018).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2024.1351958/full#supplementary-material

References

Abdullahi, S. B., Khunpanuk, C., Bature, Z. A., Chiroma, H., Pakkaranang, N., Abubakar, A. B., et al. (2022). Biometric information recognition using artificial intelligence algorithms: A performance comparison. *IEEE Access* 10, 49167–49183. doi: 10.1109/access.2022.3171850

Abdullahi, S. B., Muangchoo, K., Abubakar, A. B., Ibrahim, A. H., and Aremu, K. O. (2021). Data-driven AI-based parameters tuning using grid partition algorithm for predicting climatic effect on epidemic diseases. *IEEE Access* 9, 55388–55412. doi: 10.1109/access.2021.3068215

Alzahrani, F. S., and Abbas, I. A. (2019). Analytical estimations of temperature in a living tissue generated by laser irradiation using experimental data. *J. Therm Biol.* 85, 102421. doi: 10.1016/j.jtherbio.2019.102421

Azadshahraki, F., Jamshidi, B., and Sharabiani, V. R. (2018). Non-destructive determination of vitamin C and lycopene contents of intact cv. Newton tomatoes using NIR spectroscopy. *Yuzuncu Yıl Univ. J. Agric. Sci.* 28, 389–397. doi: 10.29133/yyutbd.423458

Bar-Sinai, Y., Hoyer, S., Hickey, J., and Brenner, M. P. (2019). Learning data-driven discretizations for partial differential equations. *Proc. Natl. Acad. Sci. U.S.A.* 116, 15344–15349. doi: 10.1073/pnas.1814058116

Bertin, N., and Génard, M. (2018). Tomato quality as influenced by preharvest factors. *Scientia Hortic.* 233, 264–276. doi: 10.1016/j.scienta.2018.01.056

Chandra, H. M., and Ramalingam, S. (2011). Antioxidant potentials of skin, pulp, and seed fractions of commercially important tomato cultivars. *Food Sci. Biotechnol.* 20, 15–21. doi: 10.1007/s10068-011-0003-z

Chandra, H. M., Shanmugaraj, B. M., Srinivasan, B., and Ramalingam, S. (2012). Influence of genotypic variations on antioxidant properties in different fractions of tomato. *J. Food Sci.* 77, C1174–C1178. doi: 10.1111/j.1750-3841.2012.02962.x

Ciaccheri, L., Tuccio, L., Mencaglia, A. A., Mignani, A. G., Hallmann, E., Sikorska-Zimny, K., et al. (2018). Directional versus total reflectance spectroscopy for the in *situ* determination of lycopene in tomato fruits. *J. Food Composition Anal.* 71, 65–71. doi: 10.1016/j.jfca.2018.01.023

Clément, A., Bacon, R., Sirois, S., and Dorais, M. (2015). Mature-ripe tomato spectral classification according to lycopene content and fruit type by visible, NIR reflectance and intrinsic fluorescence. *Qual. Assur. Saf. Crops Foods* 7, 747–756. doi: 10.3920/qas2014.0521

Clément, A., Dorais, M., and Vernon, M. (2008). Nondestructive measurement of fresh tomato lycopene content and other physicochemical characteristics using visible–NIR spectroscopy. *J. Agric. Food Chem.* 56, 9813–9818. doi: 10.1021/jf801299r

Dahlstrand, U., Sheikh, R., Dybelius Ansson, C., Memarzadeh, K., Reistad, N., and Malmsjo, M. (2019). Extended-wavelength diffuse reflectance spectroscopy with a machine-learning method for in *vivo* tissue classification. *PloS One* 14, e0223682. doi: 10.1371/journal.pone.0223682

Ding, X., Guo, Y., Ni, Y., and Kokot, S. (2016). A novel NIR spectroscopic method for rapid analyses of lycopene, total acid, sugar, phenols and antioxidant activity in dehydrated tomato samples. *Vibration. Spectrosc.* 82, 1–9. doi: 10.1016/j.vibspec.2015.10.004

Donis-González, I. R., Guyer, D. E., Pease, A., and Barthel, F. (2014). Internal characterisation of fresh agricultural products using traditional and ultrafast electron beam X-ray computed tomography imaging. *Biosyst. Eng.* 117, 104–113. doi: 10.1016/j.biosystemseng.2013.07.002

Ecarnot, M., Baczyk, P., Tessarotto, L., and Chervin, C. (2013). Rapid phenotyping of the tomato fruit model, Micro-Tom, with a portable VIS-NIR spectrometer. *Plant Physiol. Biochem.* 70, 159–163. doi: 10.1016/j.plaphy.2013.05.019

Ergon, R. (2009). Re-interpretation of NIPALS results solves PLSR inconsistency problem. *J. Chemo.* 23, 72–75. doi: 10.1002/cem.1180

FaChada, N., Figueiredo, M., Lopes, V. V., Martins, R. C., and Rosa, A. C. (2014). Spectrometric differentiation of yeast strains using minimum volume increase and minimum direction change clustering criteria. *Pattern Recogn. Lett.* 45, 55–61. doi: 10.1016/j.patrec.2014.03.008

Garcia, E., and Barrett, D. M. (2006). Evaluation of processing tomatoes from two consecutive growing seasons: quality attributes, peelability and yield. *J. Food Process. Preserv.* 30, 20–36. doi: 10.1111/j.1745-4549.2005.00044.x

Geladi, P., and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Anal. Chimica Acta* 185, 1–17. doi: 10.1016/0003-2670(86)80028-9

Gómez, R., Costa, J., Amo, M., Alvarruiz, A., Picazo, M., and Pardo, J. E. (2001). Physicochemical and colorimetric evaluation of local varieties of tomato grown in SE Spain. *J. Sci. Food Agric.* 81, 1101–1105. doi: 10.1002/jsfa.915

Huang, Y., Lu, R., and Chen, K. (2018a). Assessment of tomato soluble solids content and pH by spatially-resolved and conventional Vis/NIR spectroscopy. *J. Food Eng.* 236, 19–28. doi: 10.1016/j.jfoodeng.2018.05.008

Huang, Y., Lu, R., Hu, D., and Chen, K. (2018b). Quality assessment of tomato fruit by optical absorption and scattering properties. *Postharvest Biol. Technol.* 143, 78–85. doi: 10.1016/j.postharvbio.2018.04.016

Kanno, I., and Kuroyama, T. (2020). Estimation of the sugar content of fruit by energy-resolved computed tomography using a material decomposition method. *J. Nucl. Sci. Technol.* 58, 533–541. doi: 10.1080/00223131.2020.1845836

Konagaya, K., Al Riza, D. F., Nie, S., Yoneda, M., Hirata, T., Takahashi, N., et al. (2020). Monitoring mature tomato (red stage) quality during storage using ultraviolet-induced visible fluorescence image. *Postharvest Biol. Technol.* 160. doi: 10.1016/j.postharvbio.2019.111031

Krstajic, D., Buturovic, L. J., Leahy, D. E., and Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminform.* 6, 10. doi: 10.1186/1758-2946-6-10

Liu, W., Liu, C., Jin, J., Li, D., Fu, Y., and Yuan, X. (2020). High-throughput phenotyping of morphological seed and fruit characteristics using X-ray computed tomography. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.601475

Lofstedt, T., Hoffman, D., and Trygg, J. (2013). Global, local and unique decompositions in OnPLS for multiblock data analysis. *Anal. Chim. Acta* 791, 13–24. doi: 10.1016/j.aca.2013.06.026

Malone, E., Sato Dos Santos, G., Holder, D., and Arridge, S. (2014). Multifrequency electrical impedance tomography using spectral constraints. *IEEE Trans. Med. Imaging* 33, 340–350. doi: 10.1109/TMI.2013.2284966

Marin, M., Hobiny, A., and Abbas, I. (2021). Finite element analysis of nonlinear bioheat model in skin tissue due to external thermal sources. *Mathematics* 9. doi: 10.3390/math9131459

Martins, R. C., Barroso, T. G., Jorge, P., Cunha, M., and Santos, F. (2022). Unscrambling spectral interference and matrix effects in Vitis vinifera Vis-NIR spectroscopy: Towards analytical grade '*in vivo*' sugars and acids quantification. *Comput. Electron. Agric.* 194. doi: 10.1016/j.compag.2022.106710

Martins, R. C., Santos, F., Cunha, M., Monteiro-Silva, F., Tosin, R., Magalhães, S., et al. (2023) *Method and device for non-invasive tomographic characterisation of a sample comprising a plurality of differentiated tissues*. Available at: https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2023126532&_cid=P12-LK2EKH-19952-1.

Mishra, P., Passos, D., Marini, F., Xu, J., Amigo, J. M., Gowen, A. A., et al. (2022). Deep learning for near-infrared spectral data modelling: Hypes and benefits. *TrAC Trends Anal. Chem.* 157. doi: 10.1016/j.trac.2022.116804

Mishra, P., Roger, J.-M., Jouan-Rimbaud-Bouveresse, D., Biancolillo, A., Marini, F., Nordon, A., et al. (2021). Recent trends in multi-block data analysis in chemometrics for multi-source data integration. *TrAC Trends Anal. Chem.* 137. doi: 10.1016/j.trac.2021.116206

Mishra, P., Verschoor, J., Vries, M. N.-D., Polder, G., and Boer, M. P. (2023). Portable near-infrared spectral imaging combining deep learning and chemometrics for dry matter and soluble solids prediction in intact kiwifruit. *Infrared Phys. Technol.* 131. doi: 10.1016/j.infrared.2023.104677

Mishra, P., and Woltering, E. (2023). Semi-supervised robust models for predicting dry matter in mango fruit with near-infrared spectroscopy. *Postharvest Biol. Technol.* 200. doi: 10.1016/j.postharvbio.2023.112335

Moco, S., Capanoglu, E., Tikunov, Y., Bino, R. J., Boyacioglu, D., Hall, R. D., et al. (2007). Tissue specialization at the metabolite level is perceived during the development of tomato fruit. *J. Exp. Bot.* 58, 4131–4146. doi: 10.1093/jxb/erm271

Najjar, K., and Abu-Khalaf, N. (2021). Non-destructive quality measurement for three varieties of tomato using VIS/NIR spectroscopy. *Sustainability* 13. doi: 10.3390/su131910747

Pascale, S. D., Maggio, A., Fogliano, V., Ambrosino, P., and Ritieni, A. (2015). Irrigation with saline water improves carotenoids content and antioxidant activity of tomato. *J. Hortic. Sci. Biotechnol.* 76, 447–453. doi: 10.1080/14620316.2001.11511392

Phatak, A., and De Jong, S. (1997). The geometry of partial least squares. *J. Chemo.* 11, 311–338. doi: 10.1002/(sici)1099-128x(199707)11:4<311::Aid-cem478>3.0.Co;2-4

Shrestha, S., Knapič, M., Žibrat, U., Deleuran, L. C., and Gislum, R. (2016). Single seed near-infrared hyperspectral imaging in determining tomato (Solanum lycopersicum L.) seed quality in association with multivariate data analysis. *Sensors Actuators B: Chem.* 237, 1027–1034. doi: 10.1016/j.snb.2016.08.170

Si, Y., and Sankaran, S. (2016). Computed tomography imaging-based bitter pit evaluation in apples. *Biosyst. Eng.* 151, 9–16. doi: 10.1016/j.biosystemseng.2016.08.008

Skolik, P., Morais, C. L. M., Martin, F. L., and Mcainsh, M. R. (2019). Determination of developmental and ripening stages of whole tomato fruit using portable infrared spectroscopy and Chemometrics. *BMC Plant Biol.* 19, 236. doi: 10.1186/s12870-019-1852-5

Stelzle, F., Adler, W., Zam, A., Tangermann-Gerk, K., Knipfer, C., Douplik, A., et al. (2012). *In vivo* optical tissue differentiation by diffuse reflectance spectroscopy: preliminary results for tissue-specific laser surgery. *Surg. Innov.* 19, 385–393. doi: 10.1177/1553350611429692

Tamasi, G., Pardini, A., Bonechi, C., Donati, A., Pessina, F., Marcolongo, P., et al. (2019). Characterization of nutraceutical components in tomato pulp, skin and locular gel. *Eur. Food Res. Technol.* 245, 907–918. doi: 10.1007/s00217-019-03235-x

Tilahun, S., Park, D. S., Seo, M. H., Hwang, I. G., Kim, S. H., Choi, H. R., et al. (2018). Prediction of lycopene and β-carotene in tomatoes by portable chroma-meter and VIS/NIR spectra. *Postharvest Biol. Technol.* 136, 50–56. doi: 10.1016/j.postharvbio.2017.10.007

Toor, R. K., and Savage, G. P. (2005). Antioxidant activity in different fractions of tomatoes. *Food Res. Int.* 38, 487–494. doi: 10.1016/j.foodres.2004.10.016

Torres, I., Pérez-Marín, D., Haba, M.-J.D.L., and Sánchez, M.-T. (2015). Fast and accurate quality assessment of Raf tomatoes using NIRS technology. *Postharvest Biol. Technol.* 107, 9–15. doi: 10.1016/j.postharvbio.2015.04.004

Tosin, R., Martins, R., Pôças, I., and Cunha, M. (2022). Canopy VIS-NIR spectroscopy and self-learning artificial intelligence for a generalised model of predawn leaf water potential in Vitis vinifera. *Biosyst. Eng.* 219, 235–258. doi: 10.1016/j.biosystemseng.2022.05.007

Tosin, R., Monteiro-Silva, F., Martins, R., and Cunha, M. (2023). Precision maturation assessment of grape tissues: Hyperspectral bi-directional reconstruction using tomography-like based on multi-block hierarchical principal component analysis. *Biosyst. Eng.* 236, 147–159. doi: 10.1016/j.biosystemseng.2023.10.011

Tosin, R., Pôças, I., Novo, H., Teixeira, J., Fontes, N., Graça, A., et al. (2021). Assessing predawn leaf water potential based on hyperspectral data and pigment's concentration of Vitis vinifera L. in the Douro Wine Region. *Scientia Hortic.* 278. doi: 10.1016/j.scienta.2020.109860

Trygg, J., and Wold, S. (2003). O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. *J. Chemo.* 17, 53–64. doi: 10.1002/cem.775

USDA. (1991). "United States standards for grades of fresh tomatoes". United States Department of Agriculture. *Agric. Market. Service*.1, 4.

Verma, M., Gharpure, D. C., and Wagh, V. G. (2021). Non-destructive testing of fruits using electrical impedance tomography: A preliminary study. *AIP Conf. Proc.* 2335, 100003. doi: 10.1063/5.0043734

Vo-Dinh, T., Zam, A., Grundfest, W. S., Stelzle, F., Tangermann-Gerk, K., Mahadevan-Jansen, A., et al. (2010). "Tissue differentiation by diffuse reflectance spectroscopy for automated oral and maxillofacial laser surgery: ex vivo pilot study," in *Advanced Biomedical and Clinical Diagnostic Systems VIII*. (San Francisco, California, USA: SPIE - International Society for Optics and Photonics).

Wu, G., and Wang, C. (2014). Investigating the effects of simulated transport vibration on tomato tissue damage based on vis/NIR spectroscopy. *Postharvest Biol. Technol.* 98, 41–47. doi: 10.1016/j.postharvbio.2014.06.016

Xia, J. A., Zhang, W., Zhang, W., Yang, Y., Hu, G., Ge, D., et al. (2021). A cloud computing-based approach using the visible near-infrared spectrum to classify greenhouse tomato plants under water stress. *Comput. Electron. Agric.* 181. doi: 10.1016/j.compag.2020.105966

Zhu, Q., He, C., Lu, R., Mendoza, F., and Cen, H. (2015). Ripeness evaluation of 'Sun Bright' tomato using optical absorption and scattering properties. *Postharvest Biol. Technol.* 103, 27–34. doi: 10.1016/j.postharvbio.2015.02.007

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 13 March 2024
DOI 10.3389/fpls.2024.1360925

OPEN ACCESS

EDITED BY
Roger Deal,
Emory University, United States

REVIEWED BY
Jamilur Rahman,
Sher-e-Bangla Agricultural University, Bangladesh
Changtian Pan,
Zhejiang University, China

*CORRESPONDENCE
Daisuke Miki
daisukemiki@cemps.ac.cn

RECEIVED 24 December 2023
ACCEPTED 21 February 2024
PUBLISHED 13 March 2024

CITATION
Cheng Y, Zhang L, Li J, Dang X, Zhu J-K, Shimada H and Miki D (2024) Simple promotion of Cas9 and Cas12a expression improves gene targeting via an all-in-one strategy.
Front. Plant Sci. 15:1360925.
doi: 10.3389/fpls.2024.1360925

Simple promotion of Cas9 and Cas12a expression improves gene targeting via an all-in-one strategy

Yiqiu Cheng[1,2], Lei Zhang[1,2], Jing Li[1,2], Xiaofei Dang[1], Jian-Kang Zhu[3,4], Hiroaki Shimada[5] and Daisuke Miki[1]*

[1]Shanghai Center for Plant Stress Biology, CAS Center for Excellence in Molecular Plant Sciences, Chinese Academy of Sciences, Shanghai, China, [2]University of Chinese Academy of Sciences, Beijing, China, [3]Institute of Advanced Biotechnology and School of Life Sciences, Southern University of Science and Technology, Shenzhen, China, [4]Center for Advanced Bioindustry Technologies, Chinese Academy of Agricultural Sciences, Beijing, China, [5]Department of Biological Science and Technology, Tokyo University of Science, Tokyo, Japan

Gene targeting (GT) is a promising tool for precise manipulation of genome sequences, however, GT in seed plants remains a challenging task. The simple and direct way to improve the efficiency of GT via homology-directed repair (HDR) is to increase the frequency of double-strand breaks (DSBs) at target sites in plants. Here we report an all-in-one approach of GT in Arabidopsis by combining a transcriptional and a translational enhancer for the Cas expression. We find that facilitating the expression of Cas9 and Cas12a variant by using enhancers can improve DSB and subsequent knock-in efficiency in the Arabidopsis genome. These results indicate that simply increasing Cas protein expression at specific timings - egg cells and early embryos - can improve the establishment of heritable GTs. This simple approach allows for routine genome engineering in plants.

KEYWORDS

genome engineering, CRISPR/Cas9, CRISPR/ttCas12a, *RPS5a*, all-in-one, gene targeting (GT), transcriptional and translational enhancers

1 Introduction

Gene targeting (GT), such as precise sequence knock-ins (KIs) and substitutions, is a valuable tool for precision genome engineering. Homology-directed repair (HDR)-mediated GT has been used in a variety of organisms, but the extremely low frequency of HDR in seed plants makes GT still a challenging technology (Fauser et al., 2012; Miki et al., 2021a). We recently reported sequential transformation strategies for efficient CRISPR/Cas9-mediated GT in Arabidopsis and rice (Miki et al., 2018; Zhang et al., 2022, 2023; Li et al., 2024). Briefly, in Arabidopsis, the constructs bearing the donor and sgRNA are transformed into parental lines that stably express Cas9 in egg cells and early embryos by the DD45 promoter. Although

the efficiency of GT with sequential transformation strategy is higher than with the all-in-one method, the requirement for stable Cas9 transgenic parental lines hinders the broad application of GT in plants, e.g., in cases of different genetic backgrounds. The efficiency of double-strand breaks (DSBs) by sequence-specific nucleases (SSNs) such as Cas9 is one of the most important critical determinants for the efficient establishment of GTs (Zhang et al., 2022; Li et al., 2024). Therefore, the simplest and straight-forward approach to improve the efficiency of GT establishment is to increase the DSB frequency of SSNs. Various approaches have been examined to increase the frequency of DSBs and subsequent GT efficiency. One of these attempts applied the omega translational enhancer from tobacco mosaic virus (TMV) to promote Cas9 translation and successfully improve GT with an all-in-one strategy in Arabidopsis (Peng et al., 2020).

Recently, other transcriptional and translational enhancers have been applied to Cas9 expression to increase the efficiency of mutagenesis in plants. Using the first intron of Arabidopsis *Ubiquitin 10* (AtUbq10) as a transcriptional enhancer, Cas9-mediated heritable mutants were generated at high frequency in barley (Gasparis et al., 2018). Furthermore, dMac3, a highly efficient translational enhancer of the rice *OsMac3* gene, increased the efficiency of targeted mutagenesis by Cas9 and TALEN in rice and potato (Kusano et al., 2018; Onodera et al., 2018; Takeuchi et al., 2021; Kusano et al., 2023; Ohnuma et al., 2023). These enhancers have not been applied much to mutagenesis or GT with the CRISPR/Cas systems, and combinations of enhancers have not yet been reported. To test the utility of the enhancers in establishing GT, the present study used the AtUbq10 transcriptional enhancer and the dMac3 translational enhancer simultaneously, which are expected to drastically improve GT efficiency.

Another difficulty is that the Cas9 sgRNA design limits the target sites of GT. Any coding region or promoter sequence can be targeted if the purpose is to disrupt gene function. On the other hand, pinpoint targeting by SSNs is necessary for GTs such as KI or substitution of nucleotide sequences. The most commonly used *Streptococcus pyogenes* Cas9 (SpCas9: hereafter Cas9) recognizes the NGG (N: A/G/C/T) protospacer adjacent motif (PAM) sequence. If the target sequence of interest is AT-rich or an appropriate sgRNA sequence cannot be designed, this PAM sequence will hinder the broad application of Cas9-mediated GT. Therefore, Cas9 and another popular CRISPR/Cas system, *Lachnospiraceae bacterium ND2006* Cas12a (LbCas12a: hereafter Cas12a), which recognizes TTTV (V: A/G/C) PAM sequences, were applied to Arabidopsis with the aim of establishing GTs at a wider range of target sites. In this study, the temperature tolerant LbCas12a (ttCas12a) variant, which exhibits higher double strand break (DSB) activity under normal growth conditions (22°C) (Schindele and Puchta, 2020), was employed for GT via an all-in-one strategy in Arabidopsis.

In the present study, we investigated a way to improve precise and heritable GT efficiency in an all-in-one method using Arabidopsis as a model. The results show that simply promoting Cas protein expression improves double-strand break (DSB) efficiency, which in turn enhances GT mediated by both Cas9 and ttCas12a in plants.

2 Materials and methods

2.1 Gene accession numbers

RPS5A, At3g11940; *AtUbq10*, At4g05320.

2.2 Plant materials and growth condition

The Arabidopsis (*Arabidopsis thaliana*) accession Col-0 was used for all experiments. All plants were grown at 22°C on half Murashige and Skoog (MS) medium or in soil with a 16 h light/8 h dark photoperiod.

2.3 Plasmid construction

GT constructs for the all-in-one strategy followed the publications (Miki et al., 2018, 2021a, 2021b). Briefly, a human codon-optimized *Streptococcus pyogenes* Cas9 was used. In addition, ttCas12a was generated by using the human codon-optimized *Lachnospiraceae bacterium ND2006* Cas12a (previously known as LbCpf1) (Wang et al., 2018) and introducing a D156R amino acid substitution (Schindele and Puchta, 2020). For mutation analysis, four constructs with an AtU6-26 promoter-driven sgRNA (or crRNA) cassette and the DD45 promoter and enhancer upstream of Cas proteins were created in pCambia1300. And the *RPS5A-Bar* KI donor sequence was cloned into the above plasmids for all-in-one GT. All primers used in this study are listed in Supplementary Table 1.

2.4 Arabidopsis plant transformation

The generated constructs were transferred to Agrobacterium (*Agrobacterium tumefaciens*) GV3101 competent cells by heat shock method and spread on LB solid medium containing kanamycin and rifampicin, incubated in the dark at 28°C for 2 days to obtain positive transformants. The transformed Agrobacterium is pre-cultured in the 5 mL liquid LB, then grew in a large culture with 150ml LB, and then collected by centrifugation at 4000 rpm for 20 m. The collected Agrobacterium was resuspended in an infection solution containing 5% (w/v) sucrose, 0.22% (w/v) MS and 0.05% (v/v) Silwet-77. Cut off all fruit pods and white flowers from the plants the night before or on the day of transformation. Soak the plant buds in the infection solution for 45 s, remove and shake gently, then wrap in plastic wrap to maintain humidity. The plants were placed in darkness at 22°C for 20 h. The wrapping was removed and the plants were transferred to normal growth conditions.

T1 seeds produced by the flower dipping method were sown on half MS plates containing 50 mg/L hygromycin. Hygromycin-resistant plants were transplanted to soil and screened with three times sprays of Basta at 0.2% (v/v) concentration every three days.

2.5 DNA analysis

Genomic DNA was extracted from leaf tissue by the cethyltrimethyl ammonium bromide (CTAB) method for individual plant analysis. Leaf tissues were ground to a fine powder in liquid nitrogen using the ShakeMaster AUTO (Bio Medical Science Inc., Tokyo, Japan). The extracted DNA was used for PCR analysis of GT events. Primers were designed for genotyping and sequencing (Supplementary Table 1). The PCR system used 2x Taq Plus Master Mix II (Vazyme, Nanjing, China), according to the instructions. The PCR products were separated by electrophoresis on 1.5% (w/v) agarose gel and visualized by Image Lab software (Bio-Rad Laboratories, Hercules, USA).

The TIDE website (https://tide.nki.nl) was used to determine the mutation frequency of the target *RPS5A* locus (Brinkman et al., 2014). Total DNA was extracted from a pool of three to five independent T1 transgenic plants, and PCR was performed using the extracted total DNA as template. PCR amplicons of the target sites were subjected to Sanger sequencing. For each construct, mutation frequency compared to the Col-0 control was determined. Student's *t*-test and one-way ANOVA test were performed to compare mutation frequencies between constructs.

To examine the correlation between mutation frequency and GT ratio, the coefficient of determination (R^2) was calculated.

3 Results

3.1 Mutation efficiency by Cas9 and ttCas12a using enhancers

To examine the contribution of enhancers to DSB, a series of constructs combining the DD45 promoter-driven enhancer Cas and an sgRNA (or crRNA) expression cassette were made. The AtUbq10 first intron was coupled to dMac3 and then linked to Cas9 (*UdCas9*) or ttCas12a (*UdttCas12a*) (Figure 1A). The combined use of these two enhancers could be expected to provide much greater improvement than if each enhancer were used alone. The TMV omega translational enhancer was employed as a control (*eCas9*, *ettCas12a*) (Figure 1A) because the TMV omega enhancer has been reported to improve Cas9-mediated GT efficiency (Peng et al., 2020). The combination of the two enhancers in this study was expected to highly improve both DSB frequency and GT efficiency over the TMV omega enhancer. The sgRNA and crRNA were designed at the 3' UTR sequence of *Ribosomal Protein S5 A* (*RPS5A*) (Supplementary Figure 1), which is highly expressed in all developmental stages (Tsutsui and Higashiyama, 2017). To assess DSB efficiency, mutation frequencies via the CRISPR/Cas systems were measured. Mutation frequencies were examined in a total of 37 to 70 independent T1 transgenic plants. Total DNA was extracted from a pool of 3 to 5 plants. The target region was amplified using specific primers and subjected to Sanger sequencing. Mutation rates were calculated and determined by the TIDE website. The results showed that the combination of the AtUbq10 and the dMac3 significantly increased the mutation frequency in both Cas9 and ttCas12a compared to the use of the TMV omega enhancer (Figure 1B). And the results indicate that the combination of the AtUbq10 enhancer and the dMac3 enhancer improves DSB frequency. Therefore, it was hypothesized that the efficiency of GT would be improved when the enhancers were used in combination for the all-in-one strategy.

3.2 Gene targeting via Cas9 and ttCas12a with all-in-one strategy

Four all-in-one KI constructs were designed to determine if the combination of enhancers would improve GT efficiency (Figure 2A). In this study, the *RPS5A* gene was chosen as the target of the Basta resistance gene *Bar* KI, and the same sgRNA and crRNA as in the mutation analysis were applied. This is because the *RPS5A* gene is highly constitutively expressed in all developmental and vegetative stages (Weijers et al., 2001; Tsutsui and Higashiyama, 2017). The KI donor constructs consist of the 2A peptide, which functions as a translation initiator for polycistronic mRNA, and the *Bar* gene, flanked by a 1Kbp homology arms (Figure 2A, Supplementary Figure 1). Thus, it is likely that the precise GT plants will exhibit a

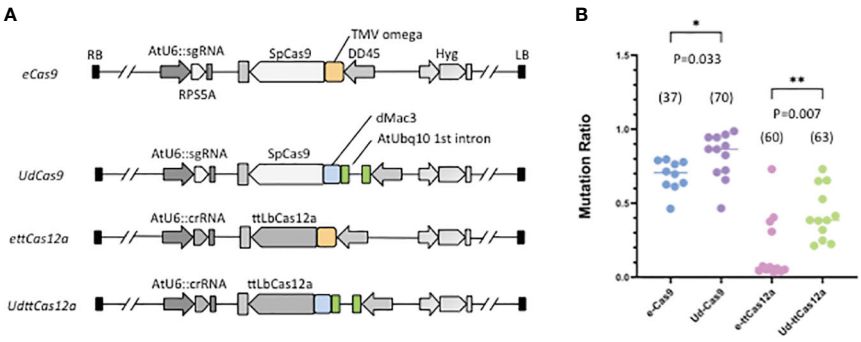


FIGURE 1
Mutation rates at the target *RPS5A* locus by Cas9 and ttCas12a. **(A)** Schematic diagram of the mutagenesis constructs at the *RPS5A* locus. Pale orange boxes indicate the TMV omega translational enhancer, green boxes represent the AtUbq10 first intron transcriptional enhancer, and blue boxes represent the dMac3 translational enhancer, respectively. All T1 transgenic plants were screened by resistance to hygromycin, followed by PCR and TIDE analysis to determine mutation rates. **(B)** Mutation rates at target *RPS5A* locus in T1 transgenic plants. The numbers in parentheses represent the number of total independent T1 transgenic plants analyzed. The standard deviation of Student's *t*-test was determined (*P<0.05, **P<0.01).

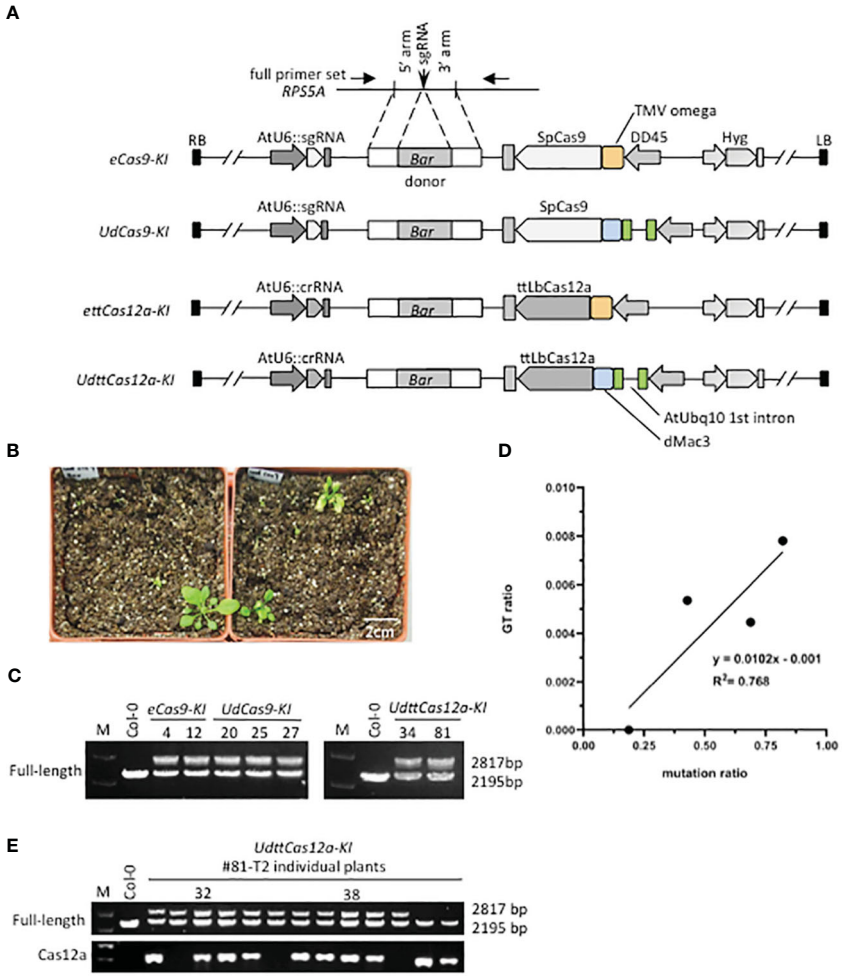


FIGURE 2
RPS5A-Bar knock-in gene targeting by Cas9 and ttCas12a. **(A)** Schematic representation of the *RPS5A-Bar* KI all-in-one constructs. The KI donor sequences of the 2A and *Bar* gene flanked by 1 Kbp homology arms were cloned into the construct shown in Figure 1A. Detailed information is provided in Supplementary Figure 1. The full-length primer set used for detecting precise and heritable GT events is indicated by arrows. **(B)** Plant growth after Basta screening. First, Arabidopsis T1 transgenic plants were screened with hygromycin on 1/2 MS plates. Hygromycin-resistant T1 transgenic plants were transplanted to soil and then sprayed with Basta. The white scale bar indicates 2 cm. **(C)** Genotyping *RPS5A-Bar* KI in individual T1 plants. The full-length primers were used to detect the precise and heritable GT events. The PCR products with a size of 2195 bp represent the endogenous *RPS5A* and the PCR products with a size of 2817 bp represent the *RPS5A-Bar* KI allele. **(D)** Statistical analysis of the relationship between mutation and GT ratio for all four constructs. For statistical analysis, the coefficient of determination (R^2) between mutation rate and GT efficiency at the *RPS5A* allele in T1 plants was calculated. **(E)** Inheritance of *RPS5A-Bar* KI in *UdttCas12a* T2 generation. Precise and heritable GT events were detected by the full-length primer set. The ttCas12a-specific primer set was used to test for the presence of T-DNA.

strong herbicide Basta resistance phenotype. Two biological repeats were performed for each construct. Transformants were screened with hygromycin on half MS plates, transplanted to soil, and sprayed with Basta (Figure 2B). All obtained Basta-resistant plants were subjected to PCR-based genotyping to detect precise and heritable GT events at the target *RPS5A* locus. From 3 to 25 independent Basta resistant T1 transformants were obtained (Table 1). Genotyping with full-length primer sets, in which both the precise GT and the endogenous alleles were detectable (Figure 2A, Supplementary Figure 1), revealed precise GT events in all constructs except *ettCas12a-KI* (Figure 2C, Table 1). GT events detected by full-length primer sets have been reported to accurately incorporate both homologous arms via HDR and to be inherited by progenies (Zhang et al., 2022, 2023; Li et al., 2024). The GT ratio of both Cas9 and ttCas12a was increased with the combination of the enhancers in comparison to the TMV omega enhancer alone (Figure 2C, Table 1). Statistical analysis revealed a significant positive correlation (R^2 = 0.768) between mutation and GT ratio (Figure 2D). This indicates that DSB frequency by the CRISPR/Cas systems is a crucial factor for the efficiency of GT, a result consistent with the previous reports (Zhang et al., 2022; Li et al., 2024).

3.3 Inheritance of gene targeting

All precise *Bar*-KI GT events detected by the full-length primer set in the T1 generation were heterozygous and inherited to the next generation as in previous reports (Miki et al., 2018; Zhang et al., 2022; Li et al., 2024) (Figure 2E, Supplementary Figure 2A). Surprisingly, no homozygous *Bar*-KI GT plants were obtained in the progeny

TABLE 1 GT efficiencies for RPS5A-Bar KI.

Construct	Enhancer	Hygromycin screening	Basta screening	GT positive	Hygromycin		Basta	
					GT frequency	Average	GT frequency	Average
eCas9-KI	TMV omega	36	3	0	0%	0.45%	0%	7.14%
		224	14	2	0.89%		14.29%	
UdCas9-KI	Ubq10 & dMac3	53	4	0	0%	0.78%	0%	9.38%
		192	16	3	1.56%		18.75%	
ettCas12a-KI	TMV omega	224	25	0	0%	0%	0%	0%
		160	22	0	0%		0%	
UdttCas12a-KI	Ubq10 & dMac3	224	23	1	0.45%	0.54%	4.35%	6.02%
		160	13	1	0.63%		7.69%	

GT efficiency was calculated based on the number of individual T1 transformants examined.

(Figure 2E, Supplementary Figure 2A). This could be due to lethality caused by RPS5A dysfunction (Weijers et al., 2001). The expression levels of all *RPS5A* mRNAs in *RPS5A-Bar* heterozygous T2 plants were similar to those in Col-0 WT plants (Supplementary Figure 2B). In this study, the 2A peptide sequence was used to generate two distinct translation products, RPS5A and BAR, from a polycistronic transcript, which often results in a single fusion protein (Barakate et al., 2020). The BAR fusion would have likely interfered with the function of the RPS5A protein. Observations of various semi-dominant phenotypes have been reported in *RPS5A* heterozygous mutants (Weijers et al., 2001). In contrast, *RPS5A-Bar* heterozygous GT plants did not show any visible morphological phenotypes. This would be due to the quantity of functional RPS5A protein in the plants. YFP KI at the C-terminal end of the *AFL1* gene has been reported to interfere with the accumulation of AFL1-YFP protein and proper subcellular localization of AFL1 to the membrane (Longkumer et al., 2024). These results suggest that sequence KI to endogenous loci can sometimes hinder their function. Conversely, it is not clear whether RPS5A fusion affects BAR function, which should be investigated in the future.

In addition, *RPS5A-Bar* heterozygous T2 plants without transgenes (Cas9 and ttCas12a) were obtained by self-pollination (Figure 2E, Supplementary Figure 2A). Since the sequential transformation strategy required two backcrosses to remove all transgenes (Zhang et al., 2022), the ability to easily obtain GT plants free of transgenes by self-pollination would be a major advantage of the all-in-one strategy. Furthermore, it has been reported that GT frequency increases in Arabidopsis and barley when donor transgenes are incorporated near endogenous target sites (Fauser et al., 2012; Lawrenson et al., 2021). However, our results suggest that the GT locus and the randomly integrated donor transgenes are not tightly linked in the chromosome (Zhang et al., 2022, 2023).

4 Discussion

The sequential transformation strategy, as previously reported, provides a higher efficiency of GT, but requires the use of parental lines, which limits its broad application (Miki et al., 2018, 2021a). Therefore, the establishment of a highly efficient all-in-one GT technology is urgently needed. Here, we demonstrated that the combined placement of enhancers increases the efficiency of GT through both Cas9 and ttCas12a with the all-in-one strategy. Although ttCas12a has been reported to show higher GT efficiency than unmodified Cas12a (Merker et al., 2020), this study demonstrates that GT can be obtained with even higher efficiency by employing enhancers. These results strongly indicate that high levels of Cas protein in egg cells and early embryos efficiently generate DSBs that facilitate homology-directed repair (HDR)-mediated heritable GT establishment in Arabidopsis.

Establishing precise and heritable GTs in seed plants remains difficult due to the extremely low efficiency of homologous recombination (Paszkowski et al., 1988; Fauser et al., 2012). The development of engineered sequence-specific nucleases (SSNs) has facilitated the establishment of GTs in many organisms, but their efficiency in plants is not yet high enough for routine use by universal users (Miki et al., 2021a). Many approaches have been attempted to improve the GT efficiency of plants. The simple and effective way to improve the efficiency of HDR-mediated GT is to promote the efficiency of the DSB. Examples include the use of the highly efficient CRISPR/Cas system (Merker et al., 2020) and the use of enhancers to promote Cas protein expression (Peng et al., 2020). Our sequential transformation method is also one way to promote DSB efficiency. This is because the use of highly efficient parental lines allows maintaining a higher level of Cas9 activity (Miki et al., 2018; Zhang et al., 2022). Transcriptional and translational enhancers have been applied for mutagenesis purposes (Gasparis et al., 2018; Kusano et al., 2018; Onodera et al., 2018; Takeuchi et al., 2021; Kusano et al., 2023; Ohnuma et al., 2023), but rarely for HDR-mediated GT (Peng et al., 2020). During the preparation of this manuscript, it has been reported that the intron-containing version of ttCas12a showed higher GT efficiency than the unmodified ttCas12a (Schindele et al., 2023). These introns may function in the same way as the AtUbq10 first intron in this study, facilitating the transport of mature mRNA into the cytoplasm by splicing and increasing translation efficiency

(Mascarenhas et al., 1990; Köhler and Hurt, 2007; Rose et al., 2008). The use of enhancers and introns can increase Cas expression and DSB frequency, resulting in increased GT efficiency. Furthermore, this study showed a strong and statistically significant positive correlation between the DSB ratio and GT efficiency. Taken together, these results indicate that DSB is one of the most important factors determining HDR-mediated GT efficiency.

The objective of this study is to establish a more efficient GT method than previously reported using an all-in-one strategy. With this motivation, the TMV omega enhancer was used as a control in this study to achieve higher GT efficiency. A 3-fold increase in GT efficiency has been reported when using the TMV omega enhancer in the all-in-one strategy (Peng et al., 2020). In this study, a small but significant differences were detected in the enhancer combination, but drastic improvements must be obtained if a version without enhancers is used as a control. Based on previous reports (Gasparis et al., 2018; Kusano et al., 2018; Onodera et al., 2018; Takeuchi et al., 2021; Kusano et al., 2023; Ohnuma et al., 2023), the AtUbq10 first intron and dMac3 is presumed to be an ideal option for improving efficiency, but further analysis is needed, including analysis of individual enhancers separately, as it is also speculated that the combination of enhancers may have some negative effects.

Here, we chose the *RPS5A* gene as a model case for efficient GT establishment. GT plants of *Bar*-KI were expected to exhibit a strong resistance phenotype to Basta herbicide treatments because of the strong and constitutive expression of the target *RPS5A* gene. However, no precise GT events were detected in the majority of Basta-resistant plants. The false antibiotic-positive phenotype of resistance gene KI plants is consistent with previous reports (Wright et al., 2005; Cai et al., 2009; Begemann et al., 2017; Permyakova et al., 2021; Vu et al., 2021). The most likely explanation for the Basta-resistant phenotype in GT-negative plants is the unwanted expression of the *Bar* gene from the randomly incorporated transgenes, even though the 5' donor homology arm does not contain a promoter sequence. Another possibility is the GT events in which only one homology arm is precisely integrated. A number of GT events have been reported in which one arm is correctly incorporated by HDR and the other arm is T-DNA integrated by NHEJ (Wolter et al., 2018; Wolter and Puchta, 2019; Gao et al., 2020; Huang et al., 2021; Peterson et al., 2021; Zhang et al., 2022; Li et al., 2024). In this study, if the 5' homology arm is correctly incorporated by HDR into the *PRS5A* locus, both *RPS5A* and *Bar* would be expressed, resulting in a Basta-resistant phenotype. Such GT events in which one arm was precise and the other arm was imprecise were detected fairly frequently (Zhang et al., 2022; Li et al., 2024), but we did not attempt to detect such GT events in the present study. Hence, it is hypothesized that these imprecise GT events contribute to the Basta resistance phenotype.

In this study, Cas9 showed higher DSB ratio and GT efficiency than ttCas12a even though they recognize almost identical target sequence. We consider that these DSB activities are mainly dependent on the design of sgRNAs (crRNAs) and that comparisons of DSB ratio and GT efficiency between Cas9 and Cas12a are irrelevant. Proper design of sgRNAs (crRNAs) and high expression systems of Cas proteins, and high expression of sgRNA also (Li et al., 2024), are crucial to obtain highly efficient and precise GTs. Although only one endogenous target locus was examined in this study, the effects of improvements in the CRISPR/Cas system are usually universal for other loci (Schindele et al., 2023). Therefore, the findings of this study can be widely applied to other plant species as well.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

YC: Data curation, Formal Analysis, Investigation, Writing – original draft. LZ: Data curation, Investigation, Writing – review & editing. JL: Data curation, Investigation, Writing – review & editing. XD: Data curation, Investigation, Writing – review & editing. JZ: Funding acquisition, Supervision, Writing – review & editing. HS: Funding acquisition, Resources, Writing – review & editing. DM: Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Shanghai Science and Technology Innovation Plan (20ZR1467000 and 23WZ2500800), the Foreign Expert Project (G202201355L), the Chinese Academy of Sciences to DM, by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) (No. 21570050) to HS, and by the National Key R&D Program of China (2021YFA1300404), National Natural Science Foundation of China (32188102) to JZ.

Acknowledgments

We would like to thank Professor Holger Puchta of the Karlsruhe Institute of Technology for providing information on ttCas12a. We would like to thank all lab members of Epigenetics and Genome engineering group and the Shanghai Center for Plant Stress Biology, CAS Center for Excellence in Molecular Plant Sciences, Chinese Academy of Sciences for assistance.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2024.1360925/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Detailed design of sgRNA and crRNA, *RPS5A-Bar* GT constructs and primers. Schematic of the *RPS5A-Bar* GT donor construct and the endogenous *RPS5A* locus, showing the full-length primer set used to detect GT events. Red letters for sgRNA and blue letters for crRNA represent PAM sequences, respectively. Green square indicates stop codon. The full-length primers are designed to anneal upstream and downstream of the homology arms and can amplify endogenous and precise knock-in alleles.

SUPPLEMENTARY FIGURE 2
RPS5A-Bar KI-GT genotyping and expression in T2. **(A)** Genotyping *RPS5A-Bar* KI in *eCas9-KI* and *UdCas9-KI* T2 individual plants. Precise and heritable GT events were detected by the full-length primer set. The Cas9-specific primer set was used to test for the presence of T-DNA. **(B)** qRT-PCR analysis for *RPS5A* expression in T2. For each construct, two heterozygous *RPS5A-Bar* GT plants were examined. The primer set for qRT-PCR is designed to anneal upstream the 5′ homology arm and can amplify endogenous and precise knock-in alleles (Supplementary Figure 1). The error bars indicate standard deviation of Student's *t*-test (n=3).

SUPPLEMENTARY FIGURE 3
Raw data of electrophoresis. Pictures show unprocessed electrophoresis gel images of Figures 2A, C, and Supplementary Figure 2A.

References

Barakate, A., Keir, E., Oakey, H., and Halpin, C. (2020). Stimulation of homologous recombination in plants expressing heterologous recombinases. *BMC Plant Biol.* 20, 336. doi: 10.1186/s12870-020-02545-7

Begemann, M. B., Gray, B. N., January, E., Gordon, G. C., He, Y., Liu, H., et al. (2017). Precise insertion and guided editing of higher plant genomes using Cpf1 CRISPR nucleases. *Sci. Rep.* 7, 11606. doi: 10.1038/s41598-017-11760-6

Brinkman, E. K., Chen, T., Amendola, M., and van Steensel, B. (2014). Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res.* 42, e168–e168. doi: 10.1093/nar/gku936

Cai, C. Q., Doyon, Y., Ainley, W. M., Miller, J. C., Dekelver, R. C., Moehle, E. A., et al. (2009). Targeted transgene integration in plant cells using designed zinc finger nucleases. *Plant Mol. Biol.* 69, 699–709. doi: 10.1007/s11103-008-9449-7

Fauser, F., Roth, N., Pacher, M., Ilg, G., Sanchez-Fernandez, R., Biesgen, C., et al. (2012). In planta gene targeting. *Proc. Natl. Acad. Sci. U.S.A.* 109, 7535–7540. doi: 10.1073/pnas.1202191109

Gao, H., Mutti, J., Young, J. K., Yang, M., Schroder, M., Lenderts, B., et al. (2020). Complex trait loci in maize enabled by CRISPR-Cas9 mediated gene insertion. *Front. Plant Sci.* 11, 535. doi: 10.3389/fpls.2020.00535

Gasparis, S., Kała, M., Przyborowski, M., Łyżnik, L. A., Orczyk, W., and Nadolska-Orczyk, A. (2018). A simple and efficient CRISPR/Cas9 platform for induction of single and multiple, heritable mutations in barley (Hordeum vulgare L.). *Plant Methods* 14, 111. doi: 10.1186/s13007-018-0382-8

Huang, T.-K., Armstrong, B., Schindele, P., and Puchta, H. (2021). Efficient gene targeting in Nicotiana tabacum using CRISPR/SaCas9 and temperature tolerant LbCas12a. *Plant Biotechnol. J.* 19, 1314–1324. doi: 10.1111/pbi.13546

Köhler, A., and Hurt, E. (2007). Exporting RNA from the nucleus to the cytoplasm. *Nat. Rev. Mol. Cell Biol.* 8, 761–773. doi: 10.1038/nrm2255

Kusano, H., Ohnuma, M., Mutsuro-Aoki, H., Asahi, T., Ichinosawa, D., Onodera, H., et al. (2018). Establishment of a modified CRISPR/Cas9 system with increased mutagenesis frequency using the translational enhancer dMac3 and multiple guide RNAs in potato. *Sci. Rep.* 8, 13753. doi: 10.1038/s41598-018-32049-2

Kusano, H., Takeuchi, A., and Shimada, H. (2023). Efficiency of potato genome editing: Targeted mutation on the genes involved in starch biosynthesis using the CRISPR/dMac3-Cas9 system. *Plant Biotechnol.* 40, 201–209. doi: 10.5511/plantbiotechnology.23.0611a

Lawrenson, T., Hinchliffe, A., Clarke, M., Morgan, Y., and Harwood, W. (2021). In-planta gene targeting in barley using Cas9 with and without geminiviral replicons. *Front. Genome editing* 3, 663380–663380. doi: 10.3389/fgeed.2021.663380

Li, J., Kong, D., Ke, Y., Zeng, W., and Miki, D. (2024). Application of multiple sgRNAs boosts efficiency of CRISPR/Cas9-mediated gene targeting in Arabidopsis. *BMC Biol.* 22, 6. doi: 10.1186/s12915-024-01810-7

Longkumer, T., Grillet, L., Chen, C.-Y., Putra, H., Schmidt, W., and Verslues, P. E. (2024). Insertion of YFP at P5CS1 and AFL1 shows the potential, and potential complications, of gene tagging for functional analyses of stress-related proteins. *Plant Cell Environ.* doi: 10.1111/pce.14861

Mascarenhas, D., Mettler, I. J., Pierce, D. A., and Lowe, H. W. (1990). Intron-mediated enhancement of heterologous gene expression in maize. *Plant Mol. Biol.* 15, 913–920. doi: 10.1007/BF00039430

Merker, L., Schindele, P., Huang, T.-K., Wolter, F., and Puchta, H. (2020). Enhancing in planta gene targeting efficiencies in Arabidopsis using temperature-tolerant CRISPR/LbCas12a. *Plant Biotechnol. J.* 18, 2382–2384. doi: 10.1111/pbi.13426

Miki, D., Wang, R., Li, J., Kong, D., Zhang, L., and Zhu, J.-K. (2021a). Gene targeting facilitated by engineered sequence-specific nucleases: potential applications for crop improvement. *Plant Cell Physiol.* 62, 752–765. doi: 10.1093/pcp/pcab034

Miki, D., Zhang, W., Zeng, W., Feng, Z., and Zhu, J.-K. (2018). CRISPR/Cas9-mediated gene targeting in Arabidopsis using sequential transformation. *Nat. Commun.* 9, 1967. doi: 10.1038/s41467-018-04416-0

Miki, D., Zinta, G., Zhang, W., Peng, F., Feng, Z., and Zhu, J.-K. (2021b). "CRISPR/Cas9-based genome editing toolbox for Arabidopsis thaliana," in *Arabidopsis Protocols Fourth Edition, Methods in Molecular Biology*. Eds. J. J. Sanchez-Serrano and J. Salinas (New York, NY: Springer US), 121–146.

Ohnuma, M., Ito, K., Hamada, K., Takeuchi, A., Asano, K., Noda, T., et al. (2023). Peculiar properties of tuber starch in a potato mutant lacking the α-glucan water dikinase 1 gene GWD1 created by targeted mutagenesis using the CRISPR/dMac3-Cas9 system. *Plant Biotechnol.* 40 (3), 219–227. doi: 10.5511/plantbiotechnology.23.0823a

Onodera, H., Shingu, S., Ohnuma, M., Horie, T., Kihira, M., Kusano, H., et al. (2018). Establishment of a conditional TALEN system using the translational enhancer dMac3 and an inducible promoter activated by glucocorticoid treatment to increase the frequency of targeted mutagenesis in plants. *PLoS One* 13, e0208959. doi: 10.1371/journal.pone.0208959

Paszkowski, J., Baur, M., Bogucki, A., and Potrykus, I. (1988). Gene targeting in plants. *EMBO J.* 7, 4021–4026. doi: 10.1002/embj.1988.7.issue-13

Peng, F., Zhang, W., Zeng, W., Zhu, J.-K., and Miki, D. (2020). Gene targeting in Arabidopsis via an all-in-one strategy that uses a translational enhancer to aid Cas9 expression. *Plant Biotechnol. J.* 18, 892–894. doi: 10.1111/pbi.13265

Permyakova, N. V., Marenkova, T. V., Belavin, P. A., Zagorskaya, A. A., Sidorchuk, Y. V., Uvarova, E. A., et al. (2021). Assessment of the Level of Accumulation of the dIFN Protein Integrated by the Knock-In Method into the Region of the Histone H3.3 Gene of Arabidopsis thaliana. *cells* 10, 2137. doi: 10.3390/cells10082137

Peterson, D., Barone, P., Lenderts, B., Schwartz, C., Feigenbutz, L., St. Clair, G., et al. (2021). Advances in Agrobacterium transformation and vector design result in high-frequency targeted gene insertion in maize. *Plant Biotechnol. J.* 19, 2000–2010. doi: 10.1111/pbi.13613

Rose, A. B., Elfersi, T., Parra, G., and Korf, I. (2008). Promoter-proximal introns in Arabidopsis thaliana are enriched in dispersed signals that elevate gene expression. *Plant Cell* 20, 543–551. doi: 10.1105/tpc.107.057190

Schindele, P., Merker, L., Schreiber, T., Prange, A., Tissier, A., and Puchta, H. (2023). Enhancing gene editing and gene targeting efficiencies in Arabidopsis thaliana by using an intron-containing version of ttLbCas12a. *Plant Biotechnol. J.* 21, 457–459. doi: 10.1111/pbi.13964

Schindele, P., and Puchta, H. (2020). Engineering CRISPR/LbCas12a for highly efficient, temperature-tolerant plant gene editing. *Plant Biotechnol. J.* 18, 1118–1120. doi: 10.1111/pbi.13275

Takeuchi, A., Ohnuma, M., Teramura, H., Asano, K., Noda, T., Kusano, H., et al. (2021). Creation of a potato mutant lacking the starch branching enzyme gene StSBE3 that was generated by genome editing using the CRISPR/dMac3-Cas9 system. *Plant Biotechnol. (Tokyo)* 38, 345–353. doi: 10.5511/plantbiotechnology.21.0727a

Tsutsui, H., and Higashiyama, T. (2017). pKAMA-ITACHI vectors for highly efficient CRISPR/Cas9-mediated gene knockout in Arabidopsis thaliana. *Plant Cell Physiol.* 58, 46–56. doi: 10.1093/pcp/pcw191

Vu, T. V., Doan, D. T. H., Tran, M. T., Sung, Y. W., Song, Y. J., and Kim, J.-Y. (2021). Improvement of the LbCas12a-crRNA system for efficient gene targeting in tomato. *Front. Plant Sci.* 12, 722552–722552. doi: 10.3389/fpls.2021.722552

Wang, M., Mao, Y., Lu, Y., Wang, Z., Tao, X., and Zhu, J.-K. (2018). Multiplex gene editing in rice with simplified CRISPR-Cpf1 and CRISPR-Cas9 systems. *J. Integr. Plant Biol.* 60, 626–631. doi: 10.1111/jipb.12667

Weijers, D., Franke-van Dijk, M., Vencken, R.-J., Quint, A., Hooykaas, P., and Offringa, R. (2001). An Arabidopsis Minute-like phenotype caused by a semi-dominant mutation in a RIBOSOMAL PROTEIN S5 gene. *Development* 128, 4289–4299. doi: 10.1242/dev.128.21.4289

Wolter, F., Klemm, J., and Puchta, H. (2018). Efficient in planta gene targeting in Arabidopsis using egg cell-specific expression of the Cas9 nuclease of Staphylococcus aureus. *Plant J.* 94, 735–746. doi: 10.1111/tpj.13893

Wolter, F., and Puchta, H. (2019). In planta gene targeting can be enhanced by the use of CRISPR/Cas12a. *Plant J.* 100, 1083–1094. doi: 10.1111/tpj.14488

Wright, D. A., Townsend, J. A., Winfrey, R. J. Jr., Irwin, P. A., Rajagopal, J., Lonosky, P. M., et al. (2005). High-frequency homologous recombination in plants mediated by zinc-finger nucleases. *Plant J.* 44, 693–705. doi: 10.1111/j.1365-313X.2005.02551.x

Zhang, W., Wang, R., Kong, D., Peng, F., Chen, M., Zeng, W., et al. (2023). Precise and heritable gene targeting in rice using a sequential transformation strategy. *Cell Rep. Methods* 3, 100389. doi: 10.1016/j.crmeth.2022.100389

Zhang, Z., Zeng, W., Zhang, W., Li, J., Kong, D., Zhang, L., et al. (2022). Insights into the molecular mechanisms of CRISPR/Cas9-mediated gene targeting at multiple loci in Arabidopsis. *Plant Physiol.* 190, 2203–2216. doi: 10.1093/plphys/kiac431

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 05 April 2024
DOI 10.3389/fpls.2024.1323296

OPEN ACCESS

EDITED BY
Roger Deal,
Emory University, United States

REVIEWED BY
Teerakiat Kerdcharoen,
Mahidol University, Thailand
Shicheng Yan,
Lanzhou University, China

*CORRESPONDENCE
Paulo Sergio De Paula Herrmann
paulo.herrmann@embrapa.br

RECEIVED 17 October 2023
ACCEPTED 18 March 2024
PUBLISHED 05 April 2024

CITATION
Herrmann PSP, Santos Luccas M, Ferreira EJ and Torre Neto A (2024) Application of electronic nose and machine learning used to detect soybean gases under water stress and variability throughout the daytime.
Front. Plant Sci. 15:1323296.
doi: 10.3389/fpls.2024.1323296

Application of electronic nose and machine learning used to detect soybean gases under water stress and variability throughout the daytime

Paulo Sergio De Paula Herrmann[1]*, Matheus dos Santos Luccas[2], Ednaldo José Ferreira[1] and André Torre Neto[1]

[1]Embrapa Instrumentation, São Carlos, Brazil, [2]Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, Brazil

The development of non-invasive methods and accessible tools for application to plant phenotyping is considered a breakthrough. This work presents the preliminary results using an electronic nose (E-Nose) and machine learning (ML) as affordable tools. An E-Nose is an electronic system used for smell global analysis, which emulates the human nose structure. The soybean (Glycine Max) was used to conduct this experiment under water stress. Commercial E-Nose was used, and a chamber was designed and built to conduct the measurement of the gas sample from the soybean. This experiment was conducted for 22 days, observing the stages of plant growth during this period. This chamber is embedded with relative humidity [RH (%)], temperature (°C), and CO_2 concentration (ppm) sensors, as well as the natural light intensity, which was monitored. These systems allowed intermittent monitoring of each parameter to create a database. The soil used was the red-yellow dystrophic type and was covered to avoid evapotranspiration effects. The measurement with the electronic nose was done daily, during the morning and afternoon, and in two phenological situations of the plant (with the healthful soy irrigated with deionized water and underwater stress) until the growth V5 stage to obtain the plant gases emissions. Data mining techniques were used, through the software "Weka™" and the decision tree strategy. From the evaluation of the sensors database, a dynamic variation of plant respiration pattern was observed, with the two distinct behaviors observed in the morning (~9:30 am) and afternoon (3:30 pm). With the initial results obtained with the E-Nose signals and ML, it was possible to distinguish the two situations, i.e., the irrigated plant standard and underwater stress, the influence of the two periods of daylight, and influence of temporal variability of the weather. As a result of this investigation, a classifier was developed that, through a non-invasive analysis of gas samples, can accurately determine the absence of water in soybean plants with a rate of 94.4% accuracy. Future investigations should be carried out under controlled conditions that enable early detection of the stress level.

KEYWORDS

E-nose, water stress, non-invasive phenotyping, artificial intelligence, data mining, soybean

1 Introduction

Abiotic stress is a term used to describe a range of environmental stresses that can affect crops, such as elevated temperature, chilling, excessive light, drought, waterlogging, wounding, exposure to ozone, UV-B irradiation, osmotic shock, and salinity. According to Bray et al. (2000), abiotic stress can lead to a potential yield loss of 51-82% in annual crops.

Zhao and collaborators in their investigation predict that significant crop yields, such as wheat, rice, corn, and soybeans, will decrease by an average of 6.0%, 3.2%, 7.4%, and 3.1%, respectively, for each degree Celsius increase in the global average temperature (Zhao et al., 2017).

The present moment demands careful consideration to improve the knowledge about biotic and abiotic stress to sustainable agriculture, food security, population growth, and the efficient use of natural resources, necessitating multidisciplinary and interdisciplinary research. As a result, collaboration among various fields, such as engineering, physics, geosciences, plant sciences, ecophysiology, computer science, and instrumentation, is crucial to developing effective non-invasive plant phenotyping techniques and methods. In agriculture, the key to practical applications lies in affordable, lightweight, and adaptable devices, instruments, sensors, and biosensors. The current trend in phenotyping research favors non-invasive techniques (Fiorani and Schurr, 2013).

Land vegetation accounts for 90% of global VOC emissions (Kiendler-Scharr et al., 2009). Plants emit volatile organic compounds (VOCs) when they suffer from disease, making them an ideal measure of phenotypic dynamics with promising results (Niederbacher et al., 2015).

Affordable plant gas detection methods could soon include electronic nose (E-Nose) and A.I. applications. The concept of electronic nose was first introduced in 1982 by Persaud and Dodd at the University of Warwick (Persaud and Dodd, 1982). Gardner and Bartlett (1994) provided the most accepted definition, defining the system as "an instrument comprised of an array of electronic sensors with specific recognition capabilities and a standard recognition system that can detect olfactory substances ranging from simple to complex" (Schaller et al., 1998). The olfactory system is more complex than other sensory systems like vision and hearing, with hundreds of different biological sensors involved in olfaction. Each olfactory receptor cell has only one type of odor receptor, which can detect only a limited number of substances (Lozano et al., 2005). Figure 1 illustrates a block diagram of the E-Nose concept. E-noses have been widely used and studied by large companies in industries such as food, cosmetics, packaging, pharmaceuticals, chemicals, petrochemicals, and agriculture (Manzoli et al., 2011; Steffens et al., 2014). This technology is a fast, simple, low-cost, and non-destructive tool for quality control and decision-making. In medicine, it has been used to detect chemicals in lung cancer patients and monitor the fertile period of cows in livestock (de Vries et al., 2019). In agriculture, monitoring insects and pests with current techniques is time-consuming and often yields variable results, making it challenging for producers and consultants to make reliable and accessible decisions. Electronic noses can use distinct types of sensors, including conductive polymers, "Carbon Black," and carbon nanotubes (Manzoli et al., 2019; Garcia-Berrios et al., 2013; Chatterjee et al., 2013).

Hazarika and collaborators presented in their investigation showed a technique to detect a pathogen called *Citrus Tristeza Virus* (CTV) in Khasi mandarin plants, where the biological process of smell was mimicked by electronic nose (E-Nose). They used invasive and destructive methods, where leaf samples were cut with scissors into square pieces measuring approximately 1 cm by 1 cm and placed in the sample holder. To evaluate the signal from the E-Nose was used classifier models such as bagging k-nearest neighbors (KNN Bag), adaptive boosting (AdaBoost) decision tree (ABDT) (Hazarika et al., 2020) and deep neural network (DNN) (Sharma et al., 2023).

Water stress triggers various physiological and biochemical responses in plants, such as stomatal closure, growth and photosynthesis repression, and respiration activation (Hale and Orcutt, 1987).

Environmental factors that affect transpiration change the water vapor gradient between the leaf surface and surrounding air: the energy balance between the sun and the leaf, air humidity and temperature, wind, and soil water availability (Angelocci et al., 2004). Therefore, transpiration intensifies with decreasing relative humidity and increasing air temperature.

Plant transpiration is vital because it goes beyond eliminating excess water and accelerates the transport of raw sap. The sap is a nutrient-rich substance, from the root to the leaves. It is transformed into an elaborate sap related to the plant's production. Transpiration assessment was used as a direct quantitative relationship of water status in vines, and this parameter was used as an indicator to organize the plant's irrigation schedule (Patakas et al., 2005).

Transpiration is the evaporation of water from plant leaves. Transpiration involves vaporizing liquid water in plant tissues and removing the vapor into the atmosphere. Crops lose water through stomata.

The relationship between transpiration and water stress can be measured using a variety of methods. One common method is to measure the rate of water loss from a plant leaf using a potometer (Pasqualotto et al., 2019). Another method is to measure the leaf water potential, which is a measure of the amount of water that is available to the plant (Ratzmann et al., 2019), and others, as for example the lysimeter. They all have limitations, disadvantages, high-cost and can sometimes produce incorrect results. However,

Abbreviations: ABDT, Adaptive Boosting Decision Tree; BR, Brazil; CMOS, Complementary Metal-Oxide- Semiconductor; CTV, Citrus Tristeza Virus; DAP, Days After Planting; DAS, Days After Sowing; DIN, Deutsches Institut für Normung; DNN, Deep Neural Network; DT, Decision Tree; GDP, Gross Domestic Product; IQR, Interquartile Range; ISO, International Organization for Standardization; KNN, K-Nearest Neighbors; LVAd, dystrophic Red-Yellow Latosol; ML, Machine Learning; PMMA, Poly methyl methacrylate; PTFE, Polytetrafluoroethylene; RH (%), relative humidity (%); R_0, Initial electrical resistance (Ω); R, Electrical resistance varying over time (Ω); SP, São Paulo; TR, Transpiration; S(%), Sensitivity (%); UV-B, Ultraviolet-B; VOC, Volatile Organic Compound; VPD, Vapor Pressure Deficit.

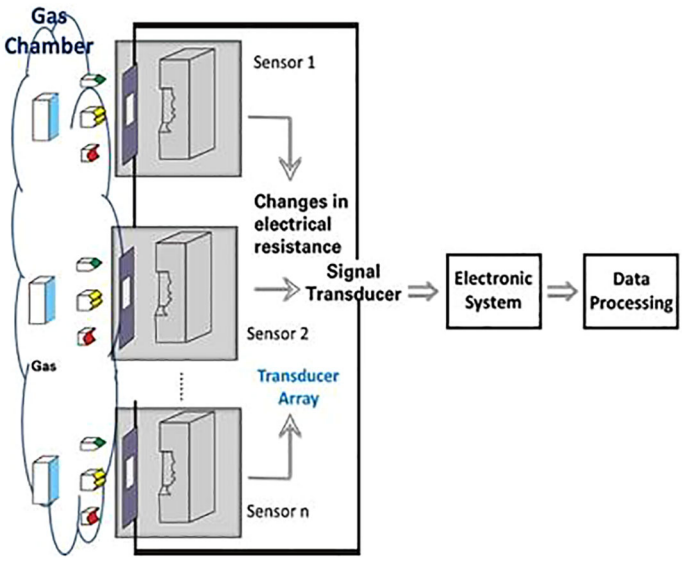


FIGURE 1
The block diagram of an E-nose and its components, including sensors, signal transducer, electronic system, and data processing.

the potential of electronic nose applications can overcome these difficulties and present a new exploration technique and method.

Sinclair et al. (2010) showed in their work the relations of vapor pressure deficit (VPD) and how it affected the sensitivity of the transpiration (TR): the time hours of Low VPD are between 7 – 11:00 a.m. and High VPD are between 11:00 a.m. – 3:00 p.m.

Studies have shown that water stress can significantly reduce the rate of transpiration in soybean plants, which can impact the plants' overall health and productivity (Lambers et al., 2008).

An increase in atmospheric CO_2, in terms of transpiration, or water use by the plant, means that the stomata, or the leaf pores that exchange gases between the leaf and the atmosphere, do not need to open as much. The effects can occur at the level of abiotic and biotic stress. Work by Sun and collaborators showed that with an increase in CO_2, the leaf transpiration rate (mmol H_2O m-2 s-1) decreased and the work showed the influence on the infestation of pea aphid (*Acyrthosiphon pisum*) in Medicago truncatula (Sun et al., 2015).

Soybean crops are highly vulnerable to the detrimental effects of drought, an abiotic stress that can cause severe damage to the plant's growth and development. This stress is particularly impactful during certain stages of the soybean's life cycle, leading to substantial yield losses. Studies have shown that soybean's sensitivity to drought is relatively high, with annual losses of up to 40% attributed to this kind of stress (Basal and Szabó, 2020).

The precise effects of water stress on soybean physiology and biochemistry remain unclear. Insufficient soil moisture triggers a range of plant adaptations, including morphological, physiological, and biochemical processes that can inhibit growth, lower photosynthesis, and transpiration rates, diminish chlorophyll levels, and modify protein structures. Given the complex nature of photosynthesis and gas exchange, these processes serve as valuable indicators of soybean response to soil moisture stress during the vegetative phase (Wijewardana et al., 2019).

The process of measuring gas exchange in leaves often involves interfering with their natural physiology, as it requires direct contact with the leaves.

Machine learning is a combination of data science and statistics that is based on the probability of occurrence of events, patterns, and behaviors in the provided database (Tan et al., 2006; Han et al., 2011). For this project, data mining techniques were employed to enable the machine to study a relevant database and detect stress levels.

Those techniques (E-Nose and Machine Learning) could be a valuable tool for assessing water stress levels in soybeans, serving as a new method of phenotyping plants that can be applied in precision agriculture. It is an affordable device that can be used for global gas analysis. It allows the application of machine learning – the basis of artificial intelligence – to examine data generated due to abiotic plant stress.

Soybeans are of great economic importance to Brazil, the second-largest producer of this crop in the world, and this crop's success directly impacts the national GDP. In 2018/2019, soybeans occupied an area of 44,062 million hectares, producing 154.566 million tons – resulting in productivity rate of 3,508 kg/ha (Embrapa Soja, 2023).

The electronic nose, through global gas analysis, is being used as a new tool for plant phenotyping, aiming to investigate water stress and the influence of sample acquisition during two separate times of gas acquisition [morning (9:30 a.m.) and (3:30 p.m.)] as well as the use of machine learning to detect the absence of irrigation. The system will provide easy handling, quick response, and a flexible tool for pattern recognition through machine learning and artificial intelligence techniques for severity observations and non-destructive measurements with portability. These differential factors highlight the advantages researchers, producers, and consultants can use in decision-making in favor of crop management.

This project work aims to investigate the use of an electronic nose and machine learning techniques to obtain non-invasive values of transpiration in soybeans (Glycine Max), evaluate the water stress and investigate de temporal variability.

The study conducted experiments proposing innovations in phenotyping, presenting, and enhancing an affordable system that employs E-Nose technology. This system allows for non-invasive and non-destructive studies using automated instrumentation in data collection. It is particularly useful in investigating abiotic effects such as water stress and examining the seasonal influence of gas emissions throughout the day. Additionally, it utilizes data mining and machine learning techniques to extract meaningful information.

2 Materials and methods

2.1 Electronic Nose Alpha FOX 2000

An Electronic Nose, model Alpha FOX 2000 was used, which also came with several tools in data processing and analysis software, helping in the proposal of joining the device with the Data Mining technology.

The equipment is built with six (06) n-type tin oxide complementary metal oxide semiconductors (CMOS) sensors. The quantity of catalytic metals (platinum palladium) in the tin oxide will be influenced by their selectivity (Vernat-Rossi et al., 1996). Table 1 lists all sensors and their main applications. They detect the variation of the electrical resistance due to the interaction of the gases with the semiconductor surface (FOX Analyzer - Hardware User's Guide, 2000).

2.1.1 E-nose measurements

The E-Nose FOX 2000 model was configured to acquire data from the variation in electrical resistance (Ω) of every six sensors over time, using the following parameters:

Acquisition duration (s): 240; Acquisition period (s): 1; Acquisition time (s): 300; Flow rate of 150 (ml/min); Injection Volume (µl): 500; Injection speed (µl/s): 500. The internal chamber of the E-Nose, that there are located the sensors of the equipment with an internal temperature adjusted to 64°C and the 0.0 (%) of the relative humidity.

TABLE 1 The sensors installed in the E-Nose are (Wei et al., 2017).

No.	Sensor	Sensitivity property	Reference Materials
1	T30/1	Organic compounds	Organic compounds
2	P10/1	Combustible gas	hydrocarbon
3	P10/2	Inflammable gas	methane
4	P40/1	Oxidizing gas	fluorine
5	T70/2	Aromatic compounds	methylbenyene, xylene
6	PA/2	Organic compounds and toxic gas	Ammonia, amines, ethyl alcohol

The gas sample, from the chamber, were collected using a Syringe for Headspace 2,500 (µl) H 0,72 (G22) d 51 PTFE seal. The precision:< ± 1% of the volume. The volume used to extract the samples was 500 (µl) to each measurement.

2.1.1.1 Calculation used on the response of the sensors

The sensitivity S (%) for each sensor was calculated using the following Equation (1):

$$S(\%) = \left(\frac{R - R_0}{R_0}\right) x100 \qquad (\%) \qquad (1)$$

R_0– Initial electrical resistance (Ω);

R – Electrical resistance varying over time (Ω).

To analyze the data that were acquired from the E-Nose, has been used the radar chart and the area radar from the peak of the sensitivity [S (%)] of each six sensors were used (S1: T30/1; S2: P10/1; S3: P10/2; S4: P40/1; S5: T70/2 and S6: PA/2). An area radar chart is a type of radar chart that uses the area enclosed by the lines connecting the data points to represent the values. In Figure 2, there is a representation of the radar chart and radar area to the peak of the sensitivity [S (%)]. The negative sign of S (%) shows that the electrical resistance (Ω) of each sensor is decreasing relative to its baseline. The information shows that the sensor is more conductive. A higher numerical value in S (%) indicates that the sensor is more sensitive to the gas sample it is detecting.

An area radar chart is a type of radar chart that uses the area enclosed by the lines connecting the data points to represent the values. In Figure 3, there is a representation of the radar chart and radar area to the peak of the sensitivity [S (%)]. This can be useful for comparing the overall performance of distinct data groups. Liu and collaborators have used the method that uses radar charts to visualize multi-dimensional data. Radar charts are a type of chart that can be used to represent multiple variables at once (Liu et al., 2008).

A radar chart is a graphical representation that effectively illustrates multidimensional data by expressing the values of each

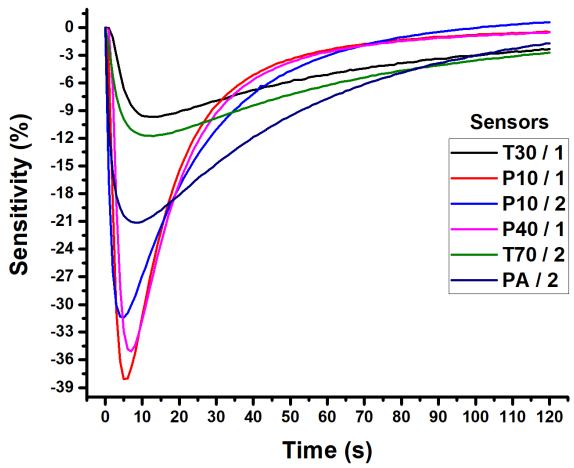


FIGURE 2
The variation in sensitivity of each of the six sensors in relation to time, depending on the gas sampled and measured in the E-Nose.

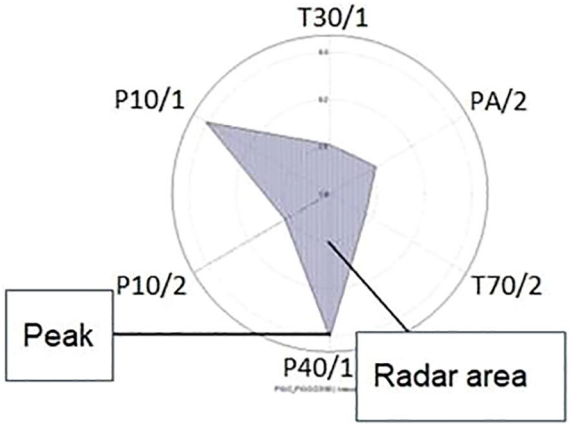


FIGURE 3
The radar chart and radar area from the sensitivity (%) peak to the six sensors (S1: T30/1; S2: P10/1; S3: P10/2; S4: P40/1; S5: T70/2 and S6: PA/2) from the E-nose.

attribute in a clear and concise manner. Its 2D visualization provides a comprehensive view of the data, making it easier to analyze and understand its various dimensions (Peng, 2022).

The method of radar chart for Multidimensional Data:

$X = \{X_1, X_2X_j, \cdots X_n\}$ is a multi-dimensional data set, and X_i $\{x_{i1}, x_{i2}, x_{i3}x_{iN}\}$ is a N-dimensional vector. Use the radar chart when N⩾3 (Liu et al., 2008).

A method for evaluating the accessibility of a facility location using the area of a radar chart was provided by Takenaka and collaborators (Takenaka et al., 2018). The authors argue that the area of a radar chart is a more stable measure of accessibility than other measures.

The Area of the Radar (A_n) was calculated with the Equation (2) where $X_i = S_i$ {S1(%), S2(%), S3(%), S4(%), S5(%), S6(%)}.

$$A_n \equiv \frac{1}{2} \sin \frac{2\pi}{n} \sum_{i=1}^{n} x_{i-1} * x_i \qquad (\%^2) \qquad (2)$$

2.2 Instrumented chamber

The instrumented chamber was specially designed to collect gaseous samples while soybeans were growing. The chamber was equipped with sensors to measure the temperature (T in °C), relative humidity (RH in %), and CO_2 concentration (CO_2 in ppm). A computer fan was also installed inside the chamber to simulate wind (flux wind = 1 cubic feet per minute or 28.32 l/min). This chamber was also designed to administer irrigation without compromising insulation, with a valve connected directly to the ground. The pot containing the soil was covered with aluminum foil to avoid gas exchange between this medium and the chamber.

The chamber is positioned in an open and isolated area with solar illumination, and it is externally and internally instrumented for monitoring.

The technique for obtaining the gas was headspace.

Chamber indoor humidity was controlled with dry air (99% purity).

The monitoring of Temperature (°C), R.H. (%), and CO_2 (ppm), both parameters measuring inside and outside of the chamber, was performed. The sensors used were an internal Vaisala CO_2 Probe GMP252 sensor for measuring CO_2 levels and an external Vaisala CO_2 Probe GMP343 model, both operating in the range of 0 – 2,000 (ppm), an internal and an external digital thermometer with a resolution of 0.1°C, an indoor and an outdoor relative humidity sensor with a resolution of 0.5 (%). This experiment's luminosity was natural and measured through a lux meter ranging from 0.001 (lumen/m²) to 19.9 K (lumen/m²).

Each internal sensor's data in the chamber was obtained every five minutes and fed to a database. Acquisition software was developed, allowing the storage and reading of sensors in real time.

The temperature to experiment was ambient (monitored internally in the chamber and externally). Irrigation control occurred by calculating the desired amount of water, based on the volumetric moisture value of the wilting soybean point, concerning the soil used and the absence of water during the plant's vegetative growth.

Figure 4 illustrates the developed chamber used to allocate the plant and extract the emanating gas to be monitored in the experiment. The homemade chamber was built with a Poly (methyl methacrylate) (PMMA) tube, also known as acrylic or Plexiglas, with the Transmittance (DIN 5036, Part 3): ca. 92% (<0.05 (%) absorption in the visible range), Refractive index (ISO 489): 1.491, Max. permanent service temperature: 70 (°C), Material density (ISO 1183): 1.19 (g cm^{-3}), Permeability coefficient (P0) @ 25 (°C) of oxygen 5.8 – 6.7 [(cm^3 *·mm)/(m^2 *·d·* atm)] and water vapor 1.7 [(cm^3 *·mm)/(m^2 *·d·* atm)] (Keller and Kouzes, 2017).

The baseline was obtained with an empty chamber before starting the soybean experiment. The Temp. (°C), R.H. (%), and CO_2 (ppm) values were measured during 03 days before including the plant. Three measurements were performed with the E-Nose for the volume of gas samples of 500 (µl). Measurements were always performed at the same time as the experiment. The temperature inside the chamber fluctuated by 4.0°C, ranging from 23.0°C to 27.0°C. The relative humidity inside the chamber displayed a variation of 9.0%, ranging between 16% and 25%. Additionally, the concentration of CO_2 inside the chamber showed a variation of 20.0 ppm_v, ranging from 250 ppm_v to 270 ppm_v.

2.3 Plant used in the experiment: soybean (*Glycine Max*)

The Brazilian soybean cultivar (*Glycine max L.* Merrill) BR-16 was used, treated, and subjected to drought under controlled conditions. The BR-16 soybean plants were irrigated during the growth phase and then subjected to drought for nine consecutive days.

The experiments were carried out during plant growth until the V5 stage of their vegetative cycle. Plants in V5 are approximately 25 to 30 (cm) tall and have six nodes, in which the leaves have unfolded leaflets.

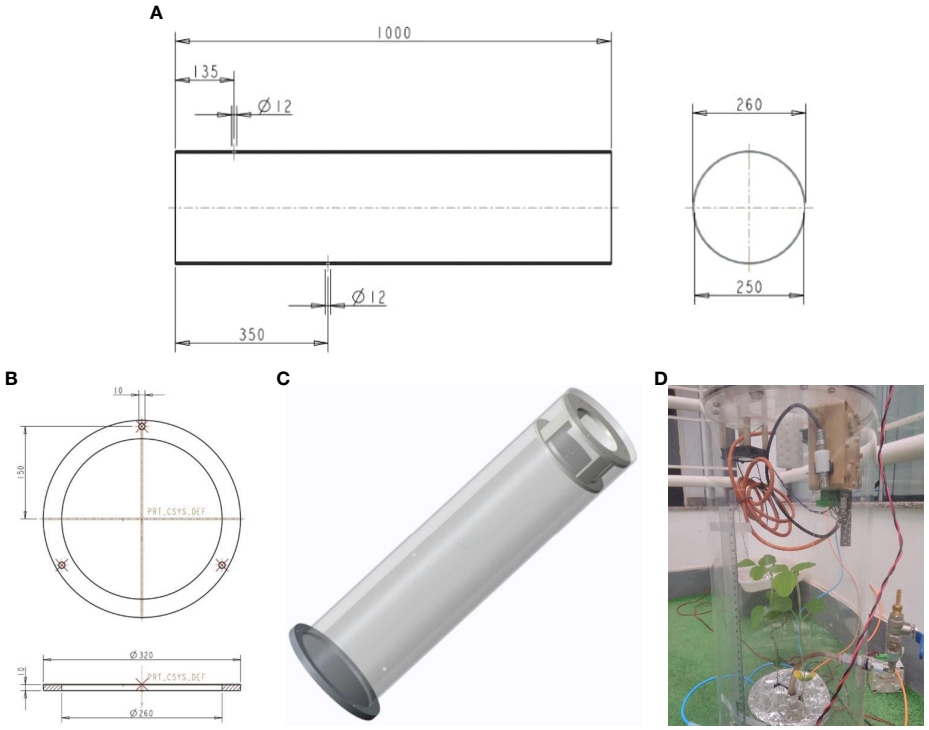


FIGURE 4
The illustration of dimension and configuration of the instrumented chamber developed and used in the experiment, **(A)** Diameter ext (d_{ext}) = 260 (mm), Diameter int (d_{int}) = 250 (mm), Thickness (Th) = 10 (mm), y = 570 (mm), Total volume = 27.93 (l); **(B)** details of the base to the chamber; **(C)** the draw of chamber assembled and **(D)** details of chamber that was developed with two plants of soybean growing inside. In this picture is possible to see the sensors, the fan installed, as well the inlet and outlet of the carrier gas and valve to control the water.

Soybean specimens were studied in a laboratory environment, under controlled conditions, and with irrigation to verify water stress. The soil moisture for soybean emergence was between 15% and 20%, and the pot, with the plant, was included in the experiment chamber as was prepared.

The experiment was carried out with the soybean for 21 days (Days After Sowing (DAS) 11 – 32). Following the steps: DAS Irrigate: 11 – 20; DAS Not Irrigated: 21 – 32. The gas samples were obtained in the daylight hours [morning (09:00 – 10:00 a.m.) and afternoon (03:30 – 04:30 p.m.)].

2.4 Irrigation procedure

Irrigation was performed with Milli-Q deionized water (~12 MΩ*cm), through the adapted valve of the chamber.

The volume used to feed the plant was 100 (cm^3) of Milli-Q water every two days.

The irrigation described was maintained for ten days, after which irrigation was completely stopped.

2.5 Soil

The soil used for this experiment was a dystrophic Red-Yellow Latosol (LVAd) with the following granulometry: Clay: 369 g/kg; Silt: 54 g/kg; Sand: 577 g/kg, with humidity at field capacity (considered at water tension of 10 KPa) of 0.295 cm^3/cm^3 and humidity at permanent wilting point (considered at water tension of -1,500 kPa) of 0.134 cm^3/cm^3.

The soil sample was obtained from Embrapa National Laboratory for Precision Agriculture (LANAPRE) at geographic coordinates 21°57'14" S and 47°51'08" W, 860 (m).

A study carried out by Ferreira et al. (2015) would have tested the covering of vessels with varied materials considering soybean and demonstrated that the isolation is effective in causing water losses to be the result only of transpiration, which was crucial to this experiment. Therefore, the soil was isolated with aluminum foil to reduce gas exchange between the medium of interest and the rhizosphere, and irrigation was administered directly into the ground. The volume of the constructed pot is V_{pot} = 8,100 (cm^3). The dimensions of the pot used is the height (h_{pot})= 24.5 (cm), diameter (D_{pot})= 14.5 (cm), and the empty pot weight 682.9 (g).

The pot with the soil and the soybean, to conduct the experiment, was prepared with the following characteristics: dry soil weight (p_{ds}) = 4,758 (g), the dry soil volume (V_{ds}) = 8,090 (cm^3) and the soil density (ρ_{ds}) = 0.59 (g/cm^3).

2.6 Data mining

The Weka™ (The University of Waikato Webpage, 2021; WEKA WIKI, 2021) tool was used for this work, and it was possible to apply several classification algorithms. K nearest

neighbor (KNN) and the decision tree were used to evaluate the results from the database. A total of 500 gas samples from 12 soybeans were used to obtain the measured values with the E-Nose. The data was used to feed Weka database.

Data Mining is a subfield of machine learning that focuses on seeking patterns and behaviors within a database (Feyyad, 1996).

Several classic data mining strategies were considered and tested on the obtained database—for example, association algorithms, k-means clustering, k-nearest neighbors, logistic regression, and decision trees. The latter strategies returned more efficient and consistent results.

Decision trees represent a classification strategy based on a tree's construction, where each node represents a logical test that separates a sample into different classes through parallel cuts in hyperplane space (Han et al., 2011; Loh, 2011).

After obtaining a well-structured and efficient tree, classifying a sample is a relatively simple task. This is one of the significant advantages of using this method when good results are achieved (Han et al., 2011; Wang et al., 2023).

3 Results

Experiment was conducted over a period of three years (from November 2017 to March 2020) and involved different soybeans subjected to water stress. This paper considers the results obtained from 22 days or roughly three weeks of experimentation. Specifically, the experiment was conducted between days after sowing (DAS) 10 to 32 for the plant. Sensors were placed both inside and outside the chamber to record the temperature (in °C), CO_2 (ppm), R.H. (%), and LUX (external values only) during both morning and afternoon periods. Figures 4–7 depict these results.

The x-axis on each figure shows the number of DAS from 11 to 32 DAS, while the y-axis displays the internal temperature in the chamber during both 9:30 a.m. (morning) and 3:30 p.m. (afternoon). The correlation between the x and y axis is represented by a trend line visible in red (afternoon) and black (morning), with a full red circle (afternoon) and a black square (morning) denoting the relationship. The graphs (3-6) depict two distinct segments of the experiment, with irrigation spanning from the 11th to the 20th day and no irrigation from the 21st to the 32nd. On the right-hand side of the y-axis, the temperature of the laboratory environment in which the plant chamber is situated is illustrated. The deep blue full circle depicts the external temperature correlation in the afternoon, while the light green inverted square represents it in the morning. Throughout the 31-day experiment period after sowing (DAS), gas samples were collected at 9:30 a.m. and 3:30 p.m. for the E-Nose (Figure 7).

The relationship between temperature, relative humidity, CO_2 concentration, and soybean growth in a closed chamber while experiencing water stress is complex and multifaceted (Smith et al., 2010). In a closed chamber with water stress, achieving

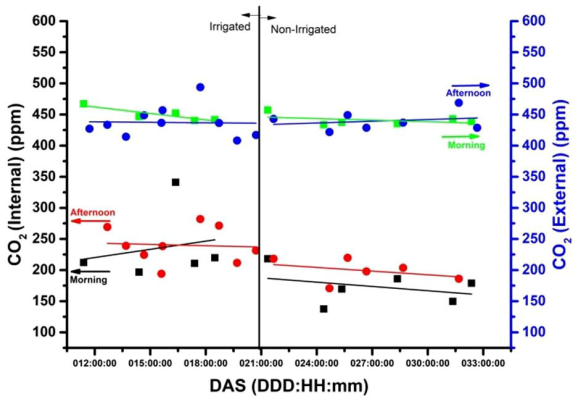


FIGURE 6
The external CO_2 concentrations (ppm) measurement on the blue y-axis and the internal measurement (irrigated and non-irrigated) on the black y-axis, during the morning (9:30 am) and afternoon (3:30 pm).

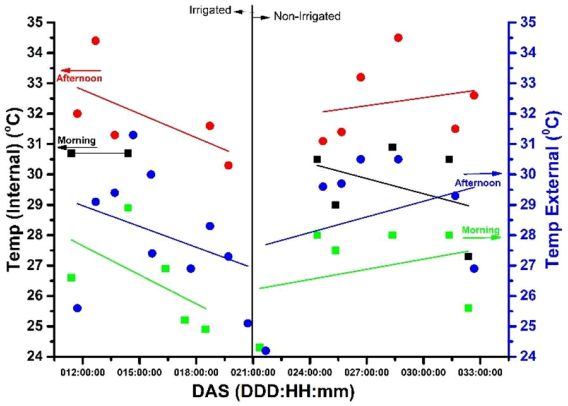


FIGURE 5
The temperature measurements, both external (blue y-axis) and internal (black y-axis), in degrees Celsius for both irrigated and non-irrigated conditions, in the morning (9:30 a.m.) and afternoon (3:30 p.m.).

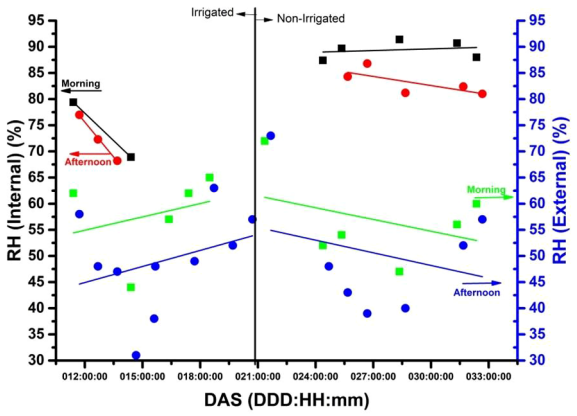


FIGURE 7
The internal relative humidity [R.H. (%)] in a plant chamber with and without irrigation for 31 days (about one month) after emergence (DAS), while also showing the external R.H. (%) in the laboratory during the morning (9:30 a.m.) and afternoon (3:30 p.m.).

optimal soybean growth requires a delicate balance between temperature, relative humidity, and CO_2 concentration.

3.1 Internal and external temperature (°C) of chamber versus days after sowing of soybean

Temperature is perhaps the most crucial variable influencing the soybean's metabolic rate and energy allocation. High temperatures can cause heat stress, reducing photosynthesis rates, impaired carbon fixation, and decreased yield (Yang et al., 2023). On the other hand, low temperatures can slow the soybean's growth rate and delay its development. Figure 5 shows us that there is clearly an increase in temperature, when compared to the temperature in the laboratory environment and the temperature inside the camera, with the plant inside, in irrigated and non-irrigated conditions. It is observed that the temperature is higher in the afternoon than in the morning.

Figure 5, presents temperature (°C) readings taken at 9:30 a.m. and 3:30 p.m. throughout the experiment, highlighting the temperature variance between the chamber and the experimental environment.

3.2 Internal and external CO_2 levels (ppm) of chamber versus days after sowing of soybean

CO_2 concentration (ppm) is a crucial variable for soybean growth as it affects photosynthesis. High CO_2 concentrations (ppm) can enhance the soybean's growth rate and yield, while low concentrations can reduce photosynthesis and growth. In Figure 6, the CO_2 levels (ppm) inside and outside the chamber were measured using an internal and an external sensor respectively, providing insights into the experimental environment. Figure 6 shows the level of CO_2 concentration with the plant being irrigated and not irrigated. It is observed that the CO_2 level, internal to the chamber, with the plant inside, through the trend line is below (~ 200 ppm) the CO2 concentration in the laboratory. It is verified for the non-irrigated period of the plant that there is a decline during the measurements carried out in the afternoons and mornings while the external CO_2 remains practically constant.

3.3 Internal and external relative humidity [RH (%)] of chamber versus days after sowing of soybean

Relative humidity is also critical in soybean growth, affecting the plant's water balance and transpiration rates. High humidity can increase the risk of disease and fungal infections, while low humidity can cause water stress and reduce the soybean's growth rate.

To examine the impact of irrigation on the internal relative humidity [R.H. (%)] within a plant chamber, data was collected over a 31-day period following emergence (DAS). Two sets of data were analyzed, one with irrigation and one without. The resulting information is presented in Figure 7 alongside the external R.H. (%) recorded within the laboratory. The graph in Figure 7 shows the state of relative humidity [RH (%)] external and internal to the chamber. The behavior of RH (%), depending on the irrigated and non-irrigated plant stage, can be seen. Relative humidity is lower in the laboratory environment, while internally, there is a more significant variation for this stage, which was evident than in the non-irrigated condition. In this case, there is an increase in relative humidity, particularly in the morning, while in the afternoon, along the trend line (red), there is a brief decrease.

3.4 External luminosity [LUX (lumen/m^2)] versus days after planting of soybean

On Figure 8, is shows the relationship between LUX (lumen/m^2) and days after planting (DAP) for plots that were irrigated and those that were not. The x-axis represents the number of DAP from 11 to 32, while the y-axis represents the LUX (lumen/m^2). The correlation between the two axes is represented by a blue circle (afternoon) and a green square (morning), both accompanied by a trend line in blue (afternoon) and green (morning). LUX (lumen/m^2) values were obtained at 9:30 in the morning and 3:30 in the afternoon. The figure is divided into two parts, with irrigation taking place between days 11 and 20 and no irrigation from days 21 to 32.

3.5 Radar area (%^{2}) measure with electronic nose versus days after sowing of soybean

During the 31-day experiment period following the Day After Sowing (DAS), the Radar Area (U.A.) from the Electronic Nose (E-Nose) was recorded and is displayed in Figure 9. The x-axis of the graph shows the number of DAS from 11 to 32. At the same time, the y-axis displays the Area Radar (U.A.), based on the value of the intensity of electrical resistance (ohms), measured by six sensors,

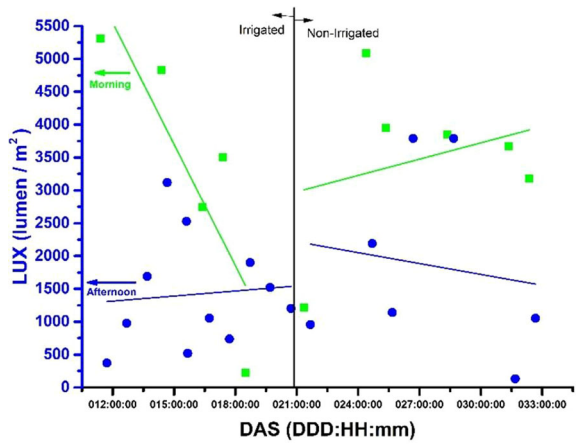


FIGURE 8
The graph shows the environmental luminosity measured in LUX (lumen/m2) for both irrigated and non-irrigated areas, during the morning (9:30 am) and afternoon (3:30 pm).

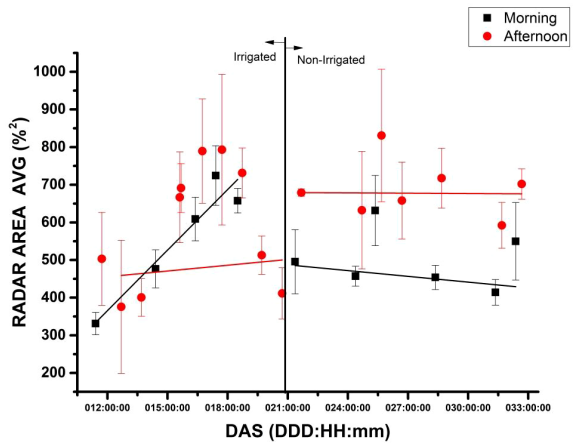


FIGURE 9
E-nose measurements of gas samples taken from a chamber containing soybeans during the DAS, using the average of radar area and standard deviation (n=3). The measurements are presented based on the time of day, either in the morning (9:30 a.m.) denoted by red circles or in the afternoon (3:30 p.m.) denoted by black squares. Moreover, the measurements are obtained from both irrigated and non-irrigated plants. For each DAS, gas samples are measured three times in both periods, i.e., the morning and afternoon to obtain the area radar measurement.

with the gas sample extracted from the chamber during both morning (9:30 a.m.) and afternoon (3:30 p.m.). The correlation between the x and y axis is demonstrated with a full red circle (afternoon) and a black square (morning), with the trend line visible in red (afternoon) and black (morning). Figure 9 is segmented into two parts, with irrigation occurring between the 11th and 20th days and no irrigation from the 21st to the 32nd.

In Figures 10–12 there are graphs obtained from the local climatological station (São Carlos (SP), BRAZIL) showing the box plot of the parameter's temperature (°C), relative humidity (%) and luminosity respectively, for the period (Summer 2020), where the experiments were carried out.

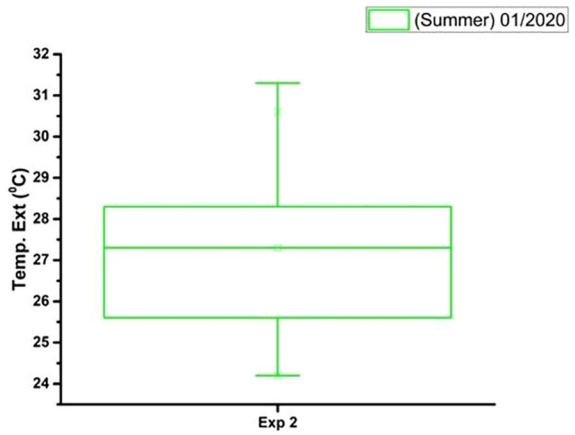


FIGURE 10
The box plot analysis of the external temperature (^{0}C) during the period (Summer 01/2020) and total time of the experiment.

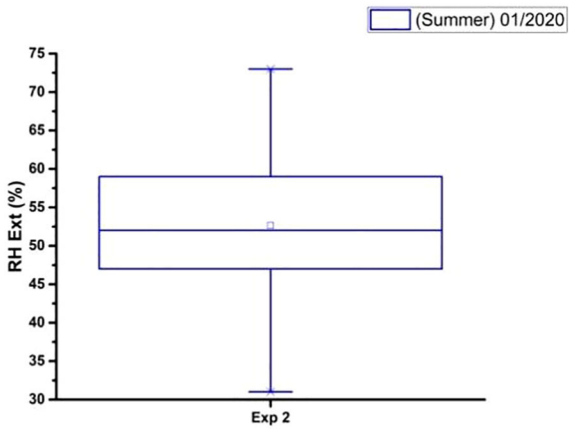


FIGURE 11
The box plot analysis of the external relative humidity (%) during the period (Summer 01/2020) and total time of the experiment.

The decision tree (DT) from the data mining and machine learning (ML) was used to visualize and explicate represent decision and decision making to the gas emanate from the plant in the state irrigated and not irrigated. Figure 13 is showing the model to DT.

A series of experiments on twelve soybean specimens was performed and derived a highly efficient detection device that can accurately identify irrigation malfunctions in a staggering 94.4(%) of cases, as confirmed by a separate efficiency test database. The device solely relies on the data obtained from E-Nose readings, which are acquired by sampling the gas concentrated by soybeans.

Table 2 shows the best learning results from sample classes (rows) and machine classifications (columns).

4 Discussion

Reviewer 2: 2-3 key points for detailed discussion. Discuss the main findings rather than simply list the literature. The discussion

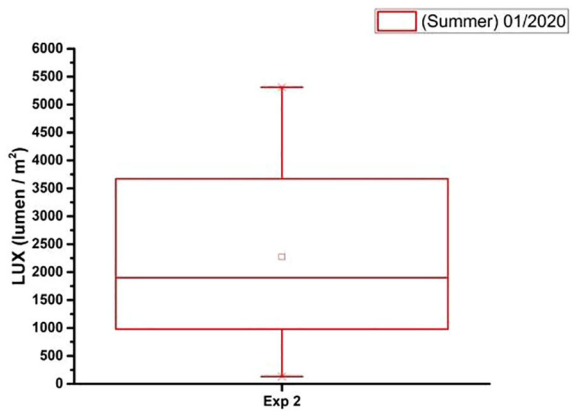


FIGURE 12
The box plot analysis of the laboratory luminosity (lumen/m^2) obtained during the period (Summer 01/2020) and total time of the experiment.

10.3389/fpls.2024.1323296

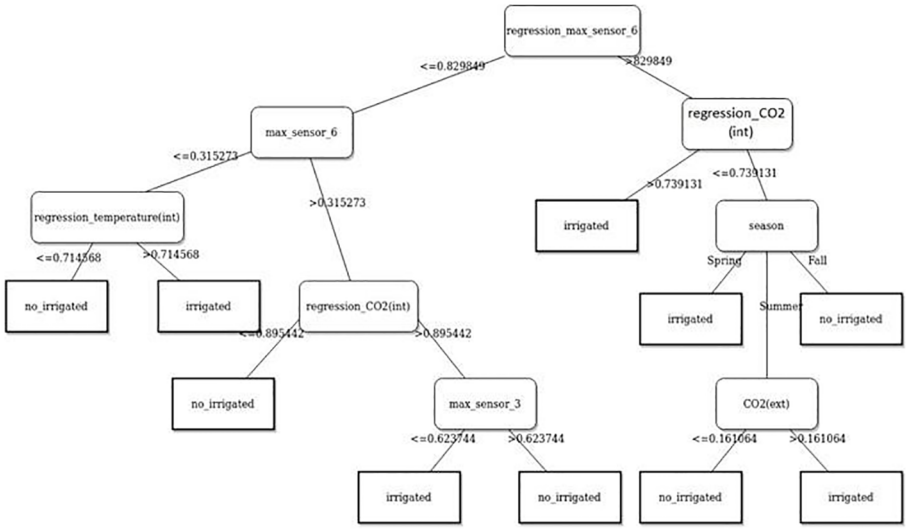


FIGURE 13
Decision tree learning extracted from test to plant 5.

part mainly focuses on the experimental results and the verification with mechanism analysis. Foi feito!

4.1 Instrumented chamber and analysis of the data obtained

The displayed temperature data in Figure 5 exhibits the temperature in both the plant inside of the chamber and laboratory environment. This figure is separated into two sections, one showcasing irrigation between the 11th to 20th day and another without irrigation from the 21st to the 32nd day.

Throughout the observation period, the temperature in the chamber is consistently higher than that of the laboratory environment. The temperature disparity is more noticeable in the afternoon than in the morning. The chamber's temperature reaches its highest point on the 11th day and its lowest on the 32nd day. Conversely, the laboratory environment's temperature peaks on the 15th day and hits its nadir on the 25th day.

It is worth noting that irrigation has a potentially minor cooling impact on the chamber's temperature by adding moisture to its surroundings.

According to the graph provided, it can be observed that the temperature within the chamber is controlled and differs from the temperature in the laboratory surroundings. This variation is primarily caused by water stress which leads to the closure of stomata, ultimately resulting in an increase in the leaf temperature. Mano et al. (2023) found in their works that water deficit led to reductions in stomata size and density in both maize and soybean leaves. These findings collectively support the idea that water stress-induced stomata closure contributes to an increase in soybean leaf temperature.

Based on the findings presented in Figure 6, it appears that the enclosed chamber experienced lower CO_2 concentrations compared to the surrounding environment. This is likely due to the active photosynthesis process of the plants, which absorb carbon dioxide from the air. During the afternoon, when the plants were more actively engaged in photosynthesis due to intense sunlight, CO_2 concentrations within the chamber were higher. However, after irrigation was stopped, the CO_2 concentrations in the chamber decreased, which may have impacted the plant's ability to photosynthesize effectively. It's worth noting that CO_2 concentrations outside the chamber remained stable throughout the day, possibly due to the experiment being carried out in laboratory conditions, where environmental factors that could affect measurements are strongly reduced.

The experiment clearly shows a difference in the amount of CO_2 concentrated between the enclosed chamber and the external environment. This discrepancy is likely due to the plants in the chamber actively undergoing photosynthesis. When irrigation stopped, CO_2 concentrations decreased because the plants were water-stressed and couldn't photosynthesize as efficiently.

The influence of plant respiration on CO_2 concentrations is distinct. While CO_2 levels in the lab environment remained constant at around 450 ppm throughout the day, there was a noticeable change when comparing irrigated and non-irrigated plant conditions. The trend lines for CO_2 concentrations demonstrate a variation between morning and afternoon during the irrigated phase.

In the absence of irrigation, there is a noticeable decline in CO_2 levels during both morning and afternoon periods, indicating a

TABLE 2 Result of the best machine obtained.

	Decision Tree	
	Irrigated	Non irrigated
Irrigated	92.7%	7.3%
Non irrigated	5.7%	94.4%

higher release of carbon dioxide in the afternoon. However, the difference between the two situations is around 50 ppm. The external sensor, which serves as a reference, shows the impact of water stress on the plant's behavior. Low CO_2 concentrations suggest that the plants are actively absorbing CO_2 from the surrounding environment rather than releasing it (Farquhar et al., 2001).

Higher CO_2 concentrations reduce stomata opening, resulting in decreased transpiration. This is because plants can photosynthesize more effectively in an elevated CO_2 environment, reducing the need for stomata opening to obtain the required CO_2. Pallas in his investigation observed that increasing carbon dioxide content caused stomata closure and reduced transpiration rate in various plant species, including soybean (Pallas, 1965).

According to the findings illustrated in Figure 7, it was noted that the relative humidity [R.H. (%)] inside the chamber was greater than the R.H. (%) outside, especially in the morning. Gomes et al. (1987) discovered that elevated R.H. (%) leads to increased stoma opening in *Theobroma cacao* seedlings, implying that the internal R.H. (%) in the chamber may have surpassed the external R.H. (%). In the same way Arve and Torre also support this, showing that high R.H. promotes stomatal opening in tomato leaves (Arve and Torre, 2015). This can be attributed to the plant's transpiration process, which introduced moisture to the air inside the chamber. Additionally, the R.H. (%) inside was higher during the irrigation phase compared to the no-irrigation period, as watering the soil and plants raised the amount of moisture in the air.

After analyzing the data presented in Figure 8 pertaining to LUX (lumen/m^2) values for 32 DAS, some noteworthy observations can be made. The highest LUX value recorded was 5,500 (lumen/m^2), which was observed on DAS 11 in the afternoon under irrigated conditions. On the other hand, the lowest recorded LUX value was 1,000 (lumen/m^2), which was observed on DAS 32 in the morning under non-irrigated conditions. Across all treatments, the average LUX value was 3,750 (lumen/m^2). For irrigated conditions, the average LUX value was 4,000 (lumen/m^2), whereas for non-irrigated conditions, it was 3,500 (lumen/m^2). Figure 7 shows the variability in terms of luminosity during the experiment period, related to the mornings and afternoons in which the measurements were taken.

According to the data from Figure 9, the Radar chart area (%^{2}) values are higher in the afternoon compared to the morning. Daylight time can influence the gas emissions by soybean plants, particularly in terms of photosynthesis and respiration. During the irrigation period from 11 to 21 DAS, the Radar chart area (%^{2}) values of irrigated and non-irrigated plants are similar. However, after the irrigation period ends on the 22nd day, the Radar chart area (%^{2}) values for non-irrigated plants start to decrease. The increased values in the afternoon may be due to plant activity and transpiration during the day. Additionally, the slight increase in values over time may be due to plant growth and development. Similar values during the irrigation period could be attributed to adequate water supply in the soil for both groups of plants. The different behavior of plants in irrigated and non-irrigated conditions during the morning and afternoon periods, in non-irrigated conditions, the plants emit gasses as they grow. However, when the plants are under stress, the emission of gasses remains different in the morning and afternoon. The literature review suggests that as the plant grows from the vegetative stage (Vc) to the V5 stage, the number of leaves and stomatal density increases, leading to more significant gas exchanges. The study examines the non-irrigated portion and tracks environmental factors like temperature, humidity, internal and external CO_2 levels, and luminosity. The findings indicate that the E-Nose detected a stage shift that aligns with previously cited research (Silva et al., 2020) that suggests a continual increase in stomatal conductance until soybeans reach the V3 or V4 stage.

4.2 Variation of radar area (%^{2}) using the electronic-nose to monitor the whole plant under stress

When irrigation stops, non-irrigated plants may experience water stress which can result in a decrease in Radar chart area values.

This increase should continue if there are no interruptions in the water supply. It is important to note that the response to the absence of irrigation after the tenth day can vary depending on several physiological and environmental factors. The health and condition of the plant are crucial in this relationship, according to a study by Rodrigues and collaborators (Rodrigues et al., 2015).

All six sensors of the E-Nose detected the gas emitted by the plants, each with varying levels of sensitivity. Of these sensors, P10/1 (sensor 2) and P40/1 (sensor 4) showed the highest sensitivity (%). Table 1 reveals that P10/1 is sensitive to combustible gasses, with hydrocarbons as the reference material, while P40/1 is sensitive to oxidizing gasses, with fluorine as the reference material. During irrigated conditions, from DAS 11 to 21, the peak sensitivity (%) to sensor P10/1 was -27.97 (%) ± 4.36 (%) and from non-irrigated conditions, from DAS 22 to 32, was -28.62 (%) ± 3.26 (%). Similarly, for sensor P40/1, during DAS from 11 to 21, the peak sensitivity (%) was -28.30 (%) ± 4.87 (%) and from non-irrigated conditions, during DAS 22 to 32, was -28.88 (%) ± 3.59 (%).

Figure 9 displays the standard deviation of the values from the radar area for irrigated and non-irrigated soybean samples. Notably, the variance between morning and afternoon measurements is significant, with the largest standard deviation occurring in the afternoon time. The causes of this disparity could be attributed to various factors, including the plant's physiological state, the environmental conditions during sample extraction, and the specific growth stage of the plant or errors in the syringe headspace. The most significant standard deviation occurred in the afternoon. On the 22nd day of the experiment, during the afternoon measurements, there were weather conditions that included closed weather, rain, and rainy and cloudy conditions. The average luminosity (lumen/m^2) measure during this time was 1,350 with a standard deviation of 1,050 (n=13), around 77% variation, much more than in the morning. In the morning the average luminosity (lumen/m^2) measure was 3,461 with a standard deviation of 1,342 (n=11), around 39% variation. Soybeans are classified as a C3 plant, which means they use the Calvin cycle to

photosynthesize. Abrupt variations in light intensity can stress soybean plants, especially affected by different light intensity treatments (Feng et al., 2019).

In tropical countries, afternoons, compared to mornings, tend to show a high temperature gradient, being significantly hotter. The photosynthetic rate of plants in general (which includes soybean plants) tends to decrease and gas exchange (respiration) increases in higher temperatures. The greatest variances observed in the afternoons (area radar graphs) are likely linked to greater exchange of gases.

The outcomes from the study suggest that the E-Nose has the potential to effectively monitor plant water stress.

The data on the climatic conditions of São Carlos (SP), BRAZIL to Figures 10–12, were provided by the conventional meteorological station of Embrapa Southeastern Livestock, located 21°57'42" S, 47° 50'28" W, 860m.

Figure 10 shows that the external temperature in January 2020 was relatively warm. The average temperature was above 25°C, and the standard deviation was relatively low. This means that the temperatures were generally consistent throughout the month. However, there were a few days with temperatures above 34°C in both experiments, and a few days with temperatures below 24°C.

From Figure 11 the average relative humidity was 70% with a standard deviation of 5%. The median relative humidity was 70%, and the interquartile range (IQR) was 10%. This means that 50% of the relative humidity values were between 60% and 80%. The box plot analysis shows that the external relative humidity in January 2020 was relatively high.

With the analysis of Figure 12 is possible to see that the external luminosity in January 2020 was relatively high. The average luminosity was 6,000 (lumen/m²) with a standard deviation of 500 (lumen/m²). The median luminosity was 5,500 (lumen/m²). This means that 50% of the luminosity values were between 4,500 (lumen/m²) and 7,500 (lumen/m²). The upper whisker extends to 8,500 (lumen/m²) and the lower whisker extends to 3,500 (lumen/m²). This means that there were a few days with luminosity values above 8,500 (lumen/m²) and a few days with luminosity values below 3,500 (lumen/m²).

The environmental conditions that were described by de data presented from the figures in Figures 5–8, 10–12, in the DAS from 21 to 32, in which there was a lack of water and in its vegetative growth would have several impacts on the physiology of soybeans, particularly in terms of water stress. At the end of the experiment, the amount of moisture in the soil {measured as gravimetric soil moisture in percentage [θw (%)]} was determined. The sample of the dystrophic Red red-yellow latosol (LVAd) used in this investigation weighed 127.25 g. It was placed in an oven for 24 hours and regulated to a temperature of 102°C. After 24 hours, the dry weight of the soil was found to be 118.81 g, and the gravimetric soil moisture content was 7.1%.

Water stress is a significant factor affecting the physiology of soybeans. The absence of water for 10 days would likely cause significant stress to the soybean plants. According to the literature (Jumrani and Bhatia, 2019) it can lead to a decrease in photochemical quenching and electron transport rate, both of which are crucial for photosynthesis. According to a study, the photochemical quenching and electron transport rate in soybeans were significantly affected by temperature and water stress. The average photochemical quenching and electron transport rate values declined progressively as the growing temperatures increased.

The decrease in CO_2 would also affect photosynthesis, as CO_2 is a crucial component in the photosynthesis process. A lower concentration of CO_2 can limit the rate of photosynthesis, potentially leading to reduced growth and yield.

The increase in relative humidity might help the soybean plants cope with the lack of water to some extent. Higher humidity can reduce the transpiration rate (water loss from plant leaves), which may help the plants conserve water. However, it's also important to note that high humidity can create a conducive environment for certain plant diseases.

The decrease in ambient light intensity would likely impact photosynthesis as well. Light is another key component of photosynthesis and a decrease in light intensity can lead to a decrease in the rate of photosynthesis (Feng et al., 2019).

In response to these environmental conditions, soybeans would likely exhibit several physiological and biochemical adaptations. For instance, under water stress, the soluble sugar content in soybeans increases, presumably to reduce water-deficit-induced damage (Wang et al., 2022).

4.3 Machine learning technique, using decision tree, from evaluate the stress

The decision tree (DT) model shown in Figure 13 is used to visualize decision making on gas emanate from plant in irrigated and not irrigated state. The DT is a hierarchical structure that starts with a root node and has branches that lead to child nodes. Each child node represents a decision point, and the branches leading away from the child node represent the possible outcomes of that decision. The DT terminates at leaf nodes, which represent the final decisions that can be made.

The DT in the image starts with the root node, which asks the question "Is regression_max_sensor 6<= 0.829849?" If the answer is yes, then the DT goes to the left child node, which asks the question "Is max_sensor 6<= 0.315273?" If the answer is yes, then the DT goes to the left child node, which predicts that the plant is irrigated. If the answer is no, then the DT goes to the right child node, which predicts that the plant is not irrigated.

If the answer to the root node question is no, then the DT goes to the right child node, which asks the question "Is regression_temperature(int)<= 0.714568?" If the answer is yes, then the DT goes to the left child node, which predicts that the plant is irrigated. If the answer is no, then the DT goes to the right child node, which predicts that the plant is not irrigated.

Using advanced technology such as the E-Nose and home built chamber, we were able to gather detailed data on plant irrigation methods. This data was carefully analyzed using Machine Learning algorithms, which ultimately resulted in a highly accurate detection rate of 94.4% for identifying inefficient irrigation practices in plants. This cutting-edge technology is revolutionizing the way we

approach plant cultivation and ensuring that our crops are grown in the most efficient and sustainable ways possible.

5 Conclusion

After developing techniques (E-Nose and ML), methods, and data analysis evaluation, a distinctive water stress pattern was identified in soybean plants. The electronic nose signals, variations in mean and standard deviation, and machine learning proved highly effective in distinguishing plant physiology parameters in the whole plant, including: (i) - The growing plant; (ii) - Two scenarios: watered plants and water-stressed plants; (iii) - Gas collected from the chamber during DAS varied depending on the time of day (morning to 9:30 a.m. and afternoon to 3:30 p.m.); (iv) - The standard deviation of the radar area in each DAS, which suggested the influence of luminosity intensity due to soybean characteristics and variations in environmental conditions; (v) - The potential of using machine learning and decision trees to classify water stress status. These findings suggest that irrigation positively impacts the Area Radar (U.A.) values of the E-Nose. Therefore, it can be used as a non-invasive method to observe the impact of irrigation on whole plants.

Using machine learning and decision tree to detect the absence of irrigation with a 94.4% accuracy rate. The most common error identified was the misclassification of irrigated samples as non-irrigated. This type of error is considered less detrimental than overlooking a sample experiencing water stress. This allowed for the early identification of stress levels, which is a crucial factor in ensuring the healthy growth of plants. Furthermore, the preliminary outcomes acquired from the E-Nose signals and machine learning enabled the researchers to differentiate between the irrigated plant control and the water stress scenario and the impact of the two daylight periods.

The classifier has demonstrated stability when tested across various scenarios, even with different soybeans subjected to varying treatments used as the testing base. As a decision tree, it has the potential to integrate E-Nose and chamber data effectively to determine water stress. The practical development, implementation, and automation of these machines can be easily achieved.

Overall, these findings have significant implications for the field of plant science and could pave the way for more efficient and effective affordable techniques and methods of plant monitoring and care.

The effects observed in the two periods, mainly in the afternoon, really demand extensive research to reach an assertive conclusion. However, it is also understood that some insights (for future work) are useful and must be refined in the light of respiratory process and/or the photosynthetic rate.

Further studies should be carried out with controlled luminosity, aiming to investigate the effect of varying luminosity in a controlled manner, as well as carrying out studies on the use of E-Nose and machine learning with drought-tolerant wheat, one of the main diseases of wheat, caused by *Fusarium graminearum Schwabe*.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

PH: Writing – review & editing, Writing – original draft, Methodology, Investigation, Funding acquisition. MS: Writing – original draft, Formal analysis. EF: Writing – review & editing, Formal analysis. AT: Writing – original draft, Software, Formal analysis.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the project Embrapa – SEG, “Wheat breeding for Brazil 2023 – 2027”, (20.22.01.001.00), The Genomics for Climate Change Research Center, and The São Paulo Research Foundation (FAPESP) # 2016/23218-0 - Engineering Research Centers (CPE).

Acknowledgments

The authors would like to thank The Genomics for Climate Change Research Center, The São Paulo Research Foundation (FAPESP) # 2016/23218-0 - Engineering Research Centers (CPE), the Nanotechnology Laboratory for Agribusiness (LNNA) and Embrapa’s National Reference Laboratory on Precision Agriculture (Lanapre) from Embrapa Instrumentation, located in São Carlos, SP (BRAZIL).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

References

Angelocci, L. R., Marin, F. R., Oliveira, R. F., and Righi, E. Z. (2004). Transpiration, leaf diffusive conductance, and atmospheric water demand relationship in an irrigated acid lime orchard. *Braz. J. Plant Physiol. Londrina* 16, 53–64. doi: 10.1590/S1677-04202004000100008

Arve, L. E., and Torre, S. (2015). Ethylene is involved in high air humidity promoted stomatal opening of tomato (Lycopersicon esculentum) leaves. *Funct. Plant Biol.* 42, 376–386. doi: 10.1071/FP14247

Basal, O., and Szabó, A. (2020). Physiology, yield and quality of soybean as affected by drought stress. *Asian J. Agric. Biol.* 8, 247–252. doi: 10.35495/ajab.2019.11.505

Bray, E. A., Bailey-Serres, J., and Weretilnyk, E. (2000). "Responses to abiotic stresses," in *Biochemistry and Molecular Biology of Plants*. Eds. B. Buchanan, W. Gruissem and R. Jones (Rockville, USA: American Society of Plant Physiologists), 1160.

Chatterjee, S., Castro, M., and Feller, J. F. (2013). An e-nose made of carbon nanotube-based quantum resistive sensors for the detection of eighteen polar/nonpolar VOC biomarkers of lung cancer. *J. Mater. Chem. B* 1, 4563–4575. doi: 10.1039/c3tb20819b

de Vries, R., Muller, M., van der Noort, V., Theelen, W. S. M. E., Schouten, R. D., Hummelink, K., et al. (2019). Prediction of response to anti-PD-1 therapy in patients with non-small cell lung cancer by electronic nose analysis of exhaled breath. *Ann. Oncol.* 30, 1660–1666. doi: 10.1093/annonc/mdz279

Embrapa Soja *Soja em números (safra 2022/23)*. Available online at: https://www.embrapa.br/en/soja/cultivos/soja1/dados-economicos (Accessed June 2023).

Farquhar, G. D., von Caemmerer, S., and Berry, J. A. (2001). Models of photosynthesis. *Plant Physiol.* 125, 42–45. doi: 10.1104/pp.125.1.42

Feng, L., Raza, M. A., Li, Z., Chen, Y., Khalid, M. H. B., Du, J., et al. (2019). The influence of light intensity and leaf movement on photosynthesis characteristics and carbon balance of soybean. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01952

Ferreira, L.C., Neiverth, W., Maronezzi, L.F.F., Sibaldelli, R.N.R., Nepomuceno, A.L., Farias, J.R.B., et al. (2015). Efficiency of cover materials in preventing evaporation in drought-stressed soybeans grown in pots. *Rev. Cienc. Agrar.*. 58 (4), 359–356. doi: 10.4322/rca.1861

Feyyad, U. M. (1996). Data mining and knowledge discovery: making sense out of data. *IEEE Expert* 11, 20–25. doi: 10.1109/64.539013

Fiorani, F., and Schurr, U. (2013). Future scenarios for plant phenotyping. *Annu. Rev. Plant Biol.* 64, 267–291. doi: 10.1146/annurev-arplant-050312-120137

FOX Analyzer. (2000) *Hardware User's Guide – Manuel Number 001*.

Garcia-Berrios, E., Theriot, J. C., Woodka, M. D., and Lewis, N. S. (2013). Detection of ammonia, 2,4,6-trinitrotoluene, and common organic vapors using thin-film carbon black-metalloporphyrin composite chemiresistors. *Sensors and Actuators B-Chemical* 188, 761–767. doi: 10.1016/j.snb.2013.07.006

Gardner, J. W., and Bartlett, P. N. (1994). A brief-history of electronic noses. *Sensors Actuators B-Chemical* 18, 211–220. doi: 10.1016/0925-4005(94)87085-3

Gomes, A. R. S., Kozlowski, T. T., and Reich, P. B. (1987). Some physiological responses of Theobroma cacao var. Catongo seedlings to air humidity. *New Phytol.* 107, 591–602. doi: 10.1111/j.1469-8137.1987.tb02929.x

Hale, M. G., and Orcutt, D. M. (1987). *The Physiology of Plants Under Stress* Vol. i-xii (New Jersey, USA: ED: Wiley-Interscience, John Wiley & Sons, Inc), 1–206.

Han, J., Kammber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. *3. ed* (Massachusetts: Morgan Kaufmann Publishers), 740.

Hazarika, S., Choudhury, R., Montazer, B., Medhi, S., Goswami, M. P., and Sarma, U. (2020). "Detection of Citrus tristeza virus in mandarin orange using a custom-developed electronic nose system," in *IEEE Transactions on Instrumentation and Measurement*, Vol. 6. 9010–9018. doi: 10.1109/TIM.2020.2997064

Jumrani, K., and Bhatia, V. S. (2019). Interactive effect of temperature and water stress on physiological and biochemical processes in soybean. *Physiol. Mol. Biol. Plants* 25, 667–681. doi: 10.1007/s12298-019-00657-5

Keller, P. E., and Kouzes, R. (2017). *Water Vapor Permeation in Plastics, Revision 1 Prepared for the U.S. Department of Energy under U.S. Department of Energy Contract DE-AC05-76RL01830* (USA: Pacific Northwesr National Laboratory).

Kiendler-Scharr, A., Wildt, J., Dal Maso, M., Hohaus, T., Kleist, E. T., Mentel, F., et al. (2009). New particle formation in forests inhibited by isoprene emissions. *Nature* 461, 381–384. doi: 10.1038/nature08292

Lambers, H., Chapin, F. S., and Pons, T. L. (2008). "Plant water relations," in *Plant Physiological Ecology* (Springer, New York, NY). doi: 10.1007/978-0-387-78341-3_5

Liu, W.–Y., Wang, B.-W., Yu, J.-X., Li, F., Wang, S.-X., and Hong, W.-X. (2008). "Visualization classification method of multi-dimensional data based on radar chart mapping," in *2008 International Conference on Machine Learning and Cybernetics*, Kunming, China: Proceedings of the Seventh International Conference on Machine Learning and Cybernetics. 857–862. doi: 10.1109/ICMLC.2008.4620524

Loh, W.-Y. (2011). Classification and regression trees. *WIREs Data Min. Knowl. Discovery* 1, 14–23. doi: 10.1002/widm.8

Lozano, J., Santos, J. P., and Horrillo, M. C. (2005). Classification of white wine aromas with an electronic nose. *Talanta* 67, 610–616. doi: 10.1016/j.talanta.2005.03.015

Mano, N. A., Madore, B., and Mickelbart, M. V. (2023). Different leaf anatomical responses to water deficit in maize and soybean. *Life* 13, 290. doi: 10.3390/life13020290

Manzoli, A., Steffens, C., Paschoalin, R. T., Correa, A. A., Alves, W. F., Leite, F. L., et al. (2011). Low-cost gas sensors produced by the graphite line-patterning technique applied to monitoring banana ripeness. *Sensors (Basel)* 11, 6425–6434. doi: 10.3390/s110606425

Manzoli, A., Steffens, C., Paschoalin, R. T., Graboski, A. M., De Mello Brandão, H., de Carvalho, B. C., et al. (2019). Volatile compounds monitoring as indicative of female cattle fertile period using electronic nose. *Sensors Actuators B: Chem.* 282, 609–616. doi: 10.1016/j.snb.2018.11.109

Niederbacher, B., Winkler, J. B., and Schnitzler, J. P. (2015). Volatile organic compounds as non-invasive markers for plant phenotyping. *J. Exp. Bot.* 66, 5403–5416. doi: 10.1093/jxb/erv219

Pallas, J. J.E. (1965). Transpiration and stomatal opening with changes in carbon dioxide content of the air. *Science* 147, 171–173. doi: 10.1126/science.147.3654.171

Pasqualotto, G., Carraro, V., Menardi, R., and Anfodillo, T. (2019). Calibration of granier-type (TDP) sap flow probes by a high precision electronic potometer. *Sensors* 19, 2419. doi: 10.3390/s19102419

Patakas, A., Noitsakis, B., and Chouzouri, A. (2005). Optimization of irrigation water use in grapevines using the relationship between transpiration and plant water status. *Agriculture Ecosyst. Environ.* 106, 253–259. doi: 10.1016/j.agee.2004.10.013

Peng, W. (2022). Improved radar chart for lighting system scheme selection. *Appl. Opt.* 61, 5619–5625. doi: 10.1364/AO.455779

Persaud, K., and Dodd, G. (1982). Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose. *Nature* 299, 352–355. doi: 10.1038/299352a0

Ratzmann, G., Zakharova, L., and Tietjen, B. (2019). Optimal leaf water status regulation of plants in drylands. *Sci. Rep.* 9, 3768. doi: 10.1038/s41598-019-40448-2

Rodrigues, F. A., Fuganti-Pagliarini, R., Marcolino-Gomes, J., Nakayama, T. J., Molinari Correa, H. B., Lobo, F. P., et al. (2015). Daytime soybean transcriptome fluctuations during water deficit stress. *BMC Genomics* 16, 505. doi: 10.1186/s12864-015-1731-x

Schaller, E., Bosset, J. O., and Escher, F. (1998). Electronic noses and their application to food. *Food Sci. Technology-Lebensmittel-Wissenschaft Technologie* 31, 305–316. doi: 10.1006/fstl.1998.0376

Sharma, C., Barkataki, N., and Sarma, U. (2023). A deep neural network with electronic nose for water stress prediction in Khasi Mandarin Orange plants. *Measurement Sci. Technol.* 34. doi: 10.1088/1361-6501/acf8e3

Silva, J. A., Santos, P. A. B., Carvalho, L. G., Moura, E. G., and Andrade, F. R. (2020). Gas exchanges and growth of soybean cultivars submitted to water deficiency. *Pesquisa Agropecuária Trop.* 50, e58854. doi: 10.1590/1983-40632020v5058854

Sinclair, T. R., Messina, C. D., Beatty, A., and Samples, M. (2010). Assessment across the United States of the benefits of altered soybean drought traits. *Agron. J.* 102, 475–482. doi: 10.2134/agronj2009.0195

Smith, S. E., Facelli, E., Pope, S., and Smith, F. A. (2010). Plant performance in stressful environments: interpreting new and established knowledge of the roles of arbuscular mycorrhizas. *Plant Soil* 326, 3–20. doi: 10.1007/s11104-009-9981-5

Steffens, C., Leite, F. L., Manzoli, A., Sandoval, R. D., Fatibello, O., and Herrmann, P. S. P. (2014). Microcantilever sensors coated with a sensitive polyaniline layer for detecting volatile organic compounds. *J. Nanoscience Nanotechnology* 14, 6718–6722. doi: 10.1166/jnn.2014.9348

Sun, Y., Guo, H., Yuan, L., Wei, J., Zhang, W., and Ge, F. (2015). Plant stomatal closure improves aphid feeding under elevated CO_2 Global Change Biology. *Global Change Biology* 21 (7), 2739–2748. doi: 10.1111/gcb.12858

Takenaka, T., Nakamura, K., Ukai, T., and Ohsawa, Y. (2018). Stability of the area of radar chart to evaluate the accessibility of facility location. *J. City Plann. Institute Japan* 53 (3), 640-645. doi: 10.11361/journalcpij.53.640

Tan, P. N., Kumar, V., and Srivastava, J. (2006). "Selecting the right interestingness measure for association patterns," in *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. Edmonton: KDD - 2002 Proceedings of the Eight ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery. 32–41.

THE UNIVERSITY OF WAIKATO WEBPAGE. Available online at: https://www.cs.waikato.ac.nz/ml/weka/ Accessed March, 2021).

Vernat-Rossi, V., Garcia, C., Talon, R., Denoyer, C., and Berdague, J. L. (1996). Rapid discrimination of meat products and bacterial strains using semiconductor gas sensors. *Sens. And Actuat. B* 37, 43–48. doi: 10.1016/S0925-4005(97)80070-6

Wang, X., Wu, Z., Zhou, Q., Wang, X., Song, S., and Dong, S. (2022). Physiological response of soybean plants to water deficit. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.809692

Wang, X., Zhou, Y., Zhao, Z., Feng, X., Wang, Z., and Jiao, M. (2023). Advanced algorithms for low dimensional metal oxides-based electronic nose application: A review. *Crystals* 13, 615. doi: 10.3390/cryst13040615

Wei, W., Li, J., and Huang, L. (2017). Discrimination of producing areas of Astragalus membranaceus using electronic nose and UHPLC-PDA combined with chemometrics. *Czech J. Food Sci.* 35, 40–47. doi: 10.17221/126/2016-CJFS

WEKA WIKI. Available online at: https://waikato.github.io/weka-wiki/ (Accessed March, 2021).

Wijewardana, C., Alsajri, F. A., Irby, J. T., Krutz, L. J., Golden, B., Henry, W. B., et al. (2019). Physiological assessment of water deficit in soybean using midday leaf water potential and spectral features. *J. Plant Interact.* 14, 533–543. doi: 10.1080/17429145.2019.1662499

Yang, L., Song, W., Xu, C., Sapey, E., Jiang, D., and Wu, C. (2023). Effects of high night temperature on soybean yield and compositions. *Front. Plant Sci.* 17. doi: 10.3389/fpls.2023.1065604

Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., et al. (2017). Temperature increase reduces global yields of major crops in four independent estimates. *Proc. Natl. Acad. Sci. U.S.A.* 114, 9326–9331. doi: 10.1073/pnas.1701762114

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 26 June 2024
DOI 10.3389/fpls.2024.1402835

OPEN ACCESS

EDITED BY
Muhammad Fazal Ijaz,
Melbourne Institute of Technology, Australia

REVIEWED BY
Rashid Ali,
University West, Sweden
Mustansar Fiaz,
IBM Research, United States
Muhammad Arslan Usman,
Kingston University, United Kingdom

*CORRESPONDENCE
Muhammad Adnan Khan
adnan@gachon.ac.kr
Daesik Jeong
jungsoft97@smu.ac.kr

†These authors have contributed equally to this work and share first authorship

RECEIVED 18 March 2024
ACCEPTED 31 May 2024
PUBLISHED 26 June 2024

CITATION
Tariq M, Ali U, Abbas S, Hassan S, Naqvi RA, Khan MA and Jeong D (2024) Corn leaf disease: insightful diagnosis using VGG16 empowered by explainable AI. *Front. Plant Sci.* 15:1402835. doi: 10.3389/fpls.2024.1402835

Corn leaf disease: insightful diagnosis using VGG16 empowered by explainable AI

Maria Tariq[1,2†], Usman Ali[3†], Sagheer Abbas[4], Shahzad Hassan[5], Rizwan Ali Naqvi[6], Muhammad Adnan Khan[7*] and Daesik Jeong[8*]

[1]Department of Computer Science, National College of Business Administration and Economics, Lahore, Pakistan, [2]Department of Computer Science, Lahore Garrison University, Lahore, Pakistan, [3]Department of Computer Science and Engineering, Sejong University, Seoul, Republic of Korea, [4]College of Computer Engineering and Science, Prince Mohammad Bin Fahd University, Al Khobar, Saudi Arabia, [5]Marine Engineering Department, Military Technological College, Muscat, Oman, [6]Department of Artificial Intelligence and Robotics, Sejong University, Seoul, Republic of Korea, [7]Department of Software, Faculty of Artificial Intelligence and Software, Gachon University, Seongnam, Republic of Korea, [8]College of Convergence Engineering, Sangmyung University, Seoul, Republic of Korea

The agricultural sector is pivotal to food security and economic stability worldwide. Corn holds particular significance in the global food industry, especially in developing countries where agriculture is a cornerstone of the economy. However, corn crops are vulnerable to various diseases that can significantly reduce yields. Early detection and precise classification of these diseases are crucial to prevent damage and ensure high crop productivity. This study leverages the VGG16 deep learning (DL) model to classify corn leaves into four categories: healthy, blight, gray spot, and common rust. Despite the efficacy of DL models, they often face challenges related to the explainability of their decision-making processes. To address this, Layer-wise Relevance Propagation (LRP) is employed to enhance the model's transparency by generating intuitive and human-readable heat maps of input images. The proposed VGG16 model, augmented with LRP, outperformed previous state-of-the-art models in classifying corn leaf diseases. Simulation results demonstrated that the model not only achieved high accuracy but also provided interpretable results, highlighting critical regions in the images used for classification. By generating human-readable explanations, this approach ensures greater transparency and reliability in model performance, aiding farmers in improving their crop yields.

KEYWORDS

intelligent agriculture system, machine learning (ML), corn leaf disease, explainable artificial intelligence (XAI), Visual Geometry Group 16 (VGG16), layer-wise relevance propagation (LRP)

1 Introduction

Economic development remains highly dependent on the agricultural sector (Mgomezulu et al., 2024), particularly in low-income nations where the industry significantly depends on the total labor force for income (Dethier and Effenberger, 2012). After rice and wheat, corn is one of the most important food crops in the world. People in central and south America generally get their carbohydrates from it. In the United States, corn is an essential alternate food source. Corn is a staple grain consumed by people in several Indonesian areas. In addition to providing humans with energy, corn is grown for animal feed, cooking oil is made from grains, and flour (sometimes known as cornstarch) is also made from grains). Corn and cob flour are also industrial raw materials (Kusumo et al., 2018). Corn is susceptible to many diseases, some of which can make it difficult for the crops to grow to their full potential (Widmer et al., 2024). The intensity of attacks on corn plants dictates how much of an impact it has. The disease usually causes irregular cell and tissue activity and stunted growth in affected plants. Some plants experience stunting and withering, while others show chromatic changes such as leaf drying or yellowing (Khoirunnisak, 2020). Early diagnosis mainly prevents and controls plant diseases, so agricultural production management and decision-making depend heavily on them. Plant disease identification has become a critical issue in recent years. Usually, leaves, stems, flowers, or fruits of disease-infected plants have visible scars or markings. Every disease or pest issue typically has a distinct visual pattern that can be utilized to identify anomalies. Most disease symptoms can initially manifest on the leaves of plants, making the leaves the primary source of information when diagnosing plant disease (Li et al., 2021). Figure 1 shows the different steps involved in intelligent agriculture systems in smart cities to detect plant diseases.

Smart agriculture utilizes various sensors to collect data (Rajamohana et al., 2024) that can be used to make better decisions and increase crop production. This generates large datasets that can be processed and analyzed by artificial intelligence (AI) and machine learning (ML) algorithms with high accuracy. While developing these algorithms improves decision-making, it will take more time to fully understand and leverage their capabilities. This decision-making process must be transparent so that people can trust AI as a part of their daily routine (Hagras, 2018). Machine learning and interoperability mean presenting machine learning models in a way understandable to humans (Linardatos et al., 2020); Bridgelall, 2024). While interpretability ensures the model is transparent before deployment, explainability explains the black box model *post hoc*. While the definition of interpretability differs from explainability in machine learning, both terms denote more or less the same meaning (Carvalho et al., 2019). Figure 2 demonstrates the difference between explainable artificial intelligence (XAI) and non-explainable artificial intelligence (non-XAI) by enhancing people's abilities with broader contextual knowledge, logical inference, and problem-solving, ultimately improving human–machine collaboration (Alsaleh et al., 2023). Systems with non-XAI may find it challenging to convey higher contextual concreteness and transparency, thus limiting their ability to interact and collaborate with humans as shown in Table 1.

XAI is crucial for anticipating corn leaf disease since it gives farmers explicit knowledge of the logic of the predictions. This transparency helps farmers manage their crops.

To effectively safeguard their crops, it is essential to enhance farmers' confidence in AI systems. So, XAI facilitates knowledge sharing and collaboration among academics, agronomists, and farmers, which support the development of more effective disease prediction models and farming methods tailored to corn leaf production.

This study focuses on leveraging explainable AI for diagnosing corn leaf diseases, emphasizing early detection, precise diagnosis, and informed decision-making. This technology aims to help farmers improve crop health, minimize yield losses, and optimize resource management in agriculture.

The paper is structured as follows: Section 2 discusses the previous research on corn leaf diseases, disease prediction methods, and the application of AI in agriculture. This section also highlights the limitations of traditional disease prediction approaches. Section 3 provides a detailed description of the methodology employed in this study, including the data collection process, preprocessing techniques, feature selection methods, and model development. It emphasizes using XAI techniques, such as interpretable machine learning algorithms, to predict corn leaf diseases while providing transparent insights into decision-making. Section 4 discusses the study's findings, including the performance metrics of the XAI model in predicting corn leaf diseases. The results are presented clearly and concisely to understand the extent to which the developed model can predict corn leaf diseases accurately. The conclusion summarizes the study's key findings and highlights the significance of the research in the context of agriculture and AI.

Main contributions

The following are the main contributions of this article:

- This study employed Explainable AI (XAI) to elucidate the decision-making process, setting it apart from previously published works that lacked such transparency.

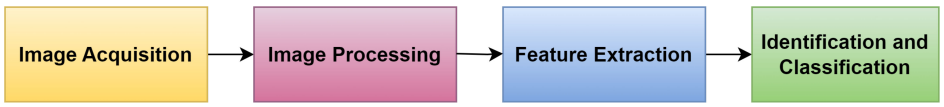


FIGURE 1
Steps involved in an intelligent agriculture system.

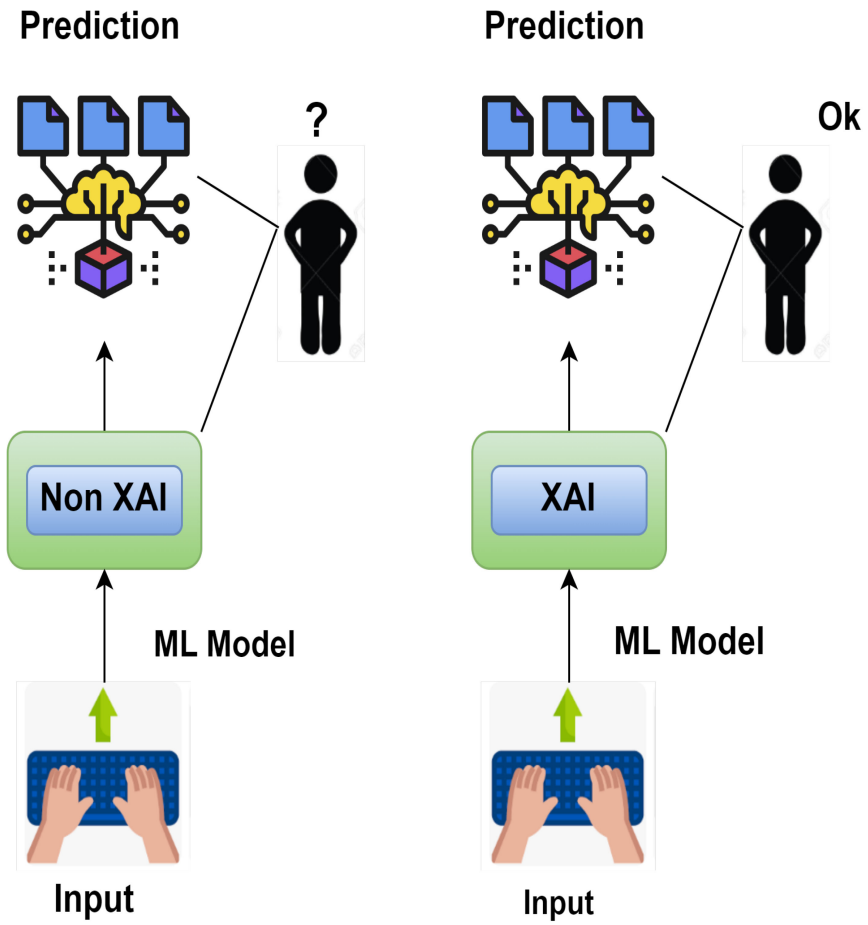


FIGURE 2
Comparison between XAI and non-XAI.

- This paper effectively presents a visual geometry group 16 (VGG16) model for utilizing a dataset containing images that address four classes of corn leaf diseases: healthy, common rust, blight, and gray leaf spot.
- The incorporation of layer-wise relevance propagation (LRP) enhances the accuracy of the analysis by providing valuable insights into the model's decision-making process.
- The combination of VGG16 and LRP offers a viable method for perceiving and investigating corn leaf diseases, enabling precise disease classification and facilitating a deeper understanding of the underlying mechanisms influencing the model's predictions.

TABLE 1 Comparison between XAI and non-XAI.

Aspect	Non-XAI	XAI
Transparency	Operates as black boxes	Designed for transparency
Interpretability	Lack of interpretability	Prioritizes interpretability
Accountability	Challenges in accountability	Enhances accountability

2 Literature review

In recent times, smart agriculture has been a field of active research. However, it is essential to note that authors have been unable to find a solution with the necessary features of customizability, interpretability, and anomaly detection in the smart agriculture field. In this section, we will discuss the existing literature related to different modules.

Using a convolutional neural network (CNN) model, Yang et al. (2019) assessed maize seedlings by analyzing spectral characteristics in the visible near-infrared region. Each maize variety's 3,600-pixel samples were utilized for CNN modeling, and an extra 400 samples were used for testing to achieve a correlation coefficient of 0.8219 with chemical methods for cold damage detection.

Agarwal et al. (2019) used a CNN with three convolution layers, three max-pooling layers, and two fully connected layers. The dataset contains corn leaves with three diseases: corn gray leaf spot, corn common rust, and corn northern leaf blight, and obtained an accuracy of 94%.

Zhen et al. (2020) used regression-guided detection network (RDNet) with the VGG16 model as a foundation and replaced the global pooling layer with a fully connected layer. Based on the encoder–decoder structure, a regional segmentation network

(RSNet) was created. The use of multi-scale kernels of varying sizes enabled the model to detect different features on different scales. The shallow field of the original convolution kernel is near the given image and accurately isolates the suspect area. Segmentation experiments were conducted on a dataset comprising field photographs of various crop diseases such as corn leaf spot, corn round spot, wheat stripe rust, wheat anthracnose, cucumber target spot disease, and cucumber anthracnose. This model achieved an accuracy of 87.04%.

Saeed et al. (2021) proposed an automated crop disease detection system by using partial least squares (PLS) for the feature selection. The authors used the pre-trained VGG19 network to extract-deep features from the plant village dataset, which included images of tomato, corn, and potato. Then, the PLS parallel fusion approach was employed to merge the features acquired through the 6th and 7th layers of the VGG19 network. Moreover, the PLS method was used to select the best features and achieved an accuracy of 90.1%.

A study by Sandotra et al. (2023) included the implementation of pre-trained DL models for corn leaf disease detection and compared various CNN architectures. Residual network (ResNet50), VGG16, VGG19, InceptionV3, and EfficientNetB0 were trained and used on a leaf dataset of corn leaf, and achieved accuracy rates of 70.02%, 91.37%, 89.69%, 87.77%, and 92.33%, respectively.

Table 2 shows the different AI models for the diagnosis of corn leaf disease, involving their types, accuracies, limitations, and their applied datasets. Two authors used the CNN model that was applied to two different studies based on supervised learning, and reached accuracy rates of 82.19% and 94% on the hyperspectral data and the multi-corn leaf diseases, respectively. Other studies employed the RDNet method using unsupervised learning, achieving an accuracy of 84.04% in detecting both corn leaf spot and corn round spot. Some authors focused on the fusion of VGG19, CNN, and PLS using supervised learning. It gave an accuracy of 90.01% on the images of tomato, corn, and potato. The last one used five different CNN models to detect corn leaf disease. So, all the given models did not use XAI to predict corn leaf disease.

3 Materials and methods

The dataset contained 4,188 total images that included four corn leaf disease classes. There were 1,146 images of blight leaf, 1,306 images of common rust, 574 images of gray spot, and 1,162 images of healthy leaf, and this dataset was divided randomly into 70% (2,930 images) for training and 30% (1,258 images) for testing. The data were acquired from the Kaggle repository (Smaranjit Ghose, 2024). Figure 3 shows the samples of corn leaf disease.

Image classification involves several steps: First, a labeled dataset of images is created. Second, the images are preprocessed by way of resizing, normalizing, and augmenting them; then, the features are extracted using pre-trained CNN or other methods. On the basis of the extracted features and their corresponding labels, a model is trained, after which it is validated for generalization, fine-tuned as required, and evaluated on a different test dataset. Lastly, the trained model is put to use in the real world, which classifies the new images by their visual content. This iterative process enables the creation of robust models for tasks such as object detection, classification, segmentation, and generation across diverse domains (Fadhilla et al., 2023). Figure 4 is an illustrated representation of this process.

The proposed model design has two phases: training testing. This design contains five essential steps, as shown in Figure 5. In the first step, the data are incorporated, and then used for preprocessing in the second step. After applying DL models to the data, the XAI model is used to explain the results. Ultimately, the last step ensures the model's performance.

Figure 6 depicts the architecture of the proposed approach, providing a general overview. The dataset (Smaranjit Ghose, 2024) is processed in two phases: training and testing. So, a DL model is

TABLE 2 Comparison of different AI models used to predict corn leaf disease.

Ref.	Model	Model type	Accuracy (%)	Applied on	Limitations
Yang et al. (2019)	CNN	Supervised learning	82.19	hyperspectral images	Non – Explainable Artificial Intelligence (XAI) Used
Agarwal et al. (2019)	CNN	Supervised learning	94	corn gray leaf spot, corn common rust, corn northern leaf blight, and healthy	
Zhen et al. (2020)	RDNet	Un-supervised learning	84.04	corn leaf spot and corn round spot	
Saeed et al. (2021)	VGG19, CNN, and PLS	Supervised learning	90.01	images of tomato, corn, and potato	
Sandotra et al. (2023)	ResNet50, VGG16, VGG19, InceptionV3, and EfficientNetB0	Supervised learning	70.02, 91.37, 89.69, 87.77, and 92.33	corn blight, corn common rust, corn gray leaf spot, and corn healthy	

FIGURE 3
Samples of corn leaf disease (Smaranjit Ghose, 2024).

used for training in which the VGG16 model is employed; after that, XAI techniques are implemented to visualize the essential features using the trained model employing LRP. This study used the LRP method for XAI. LRP, one of the primary algorithms for network explainability, uses the backpropagation algorithm (Bach et al., 2015). LRP explains a classifier's prediction for a specific data point by attributing 'Relevance values' (Ri) to important input components using the topology of the trained model. It is effective for images, videos, and text (Ullah et al., 2021). The DL model is used for the basic prediction of preprocessed data (Makridakis et al., 2023). The XAI model contrasts these expectations and the preprocessed data and utilizations for the correlations to make sense of the prediction made by the DL model. So, the clarifications given by the XAI model are good and show fair thinking behind the prediction, and the testing data are applied to the trained model to check the performance of the model.

The proposed model for detecting corn leaf disease, which incorporates XAI, is presented in detail in Figure 7. In this proposed model, during the training phase, the data acquisition step is responsible for obtaining the raw dataset (Smaranjit Ghose, 2024) of corn leaf disease images from the Kaggle repository. This dataset includes four categories: blight, gray spot, common rust, and healthy. The initial step is to preprocess the raw dataset. This involves resizing the images and normalizing the data. After that, the dataset is then divided into training and testing steps by the requirements of the DL model implementation. In this approach, the CNN–VGG16 model used for this research includes a convolution layer, a pooling layer, a dropout layer, a flattened layer, and a dense layer (Asriny and Jayadi, 2023); Mardiana et al., 2023). The VGG16 model was modified to include the four classes required by the sample dataset (Smaranjit Ghose, 2024). During the testing phase, the testing dataset is used for assessment. The trained model stored in the cloud is then used to classify the corn leaf diseases into four classes: blight, gray spot, common rust, and healthy, which explain the corn leaf disease.

In this study, a previously trained VGG16 model was used for transfer learning in AI. This technique is precious for AI developers as it gives them a shortcut to building good models, which is a bonus for both time and computer resources. The process usually involves the utilization of the VGG16 as a feature extractor, which in turn captures the particular characteristics of images related to the new task from the dataset. This model has several advantages, such as better performance, a smaller amount of labeled data, and efficient use of computational resources. So, transfer learning with

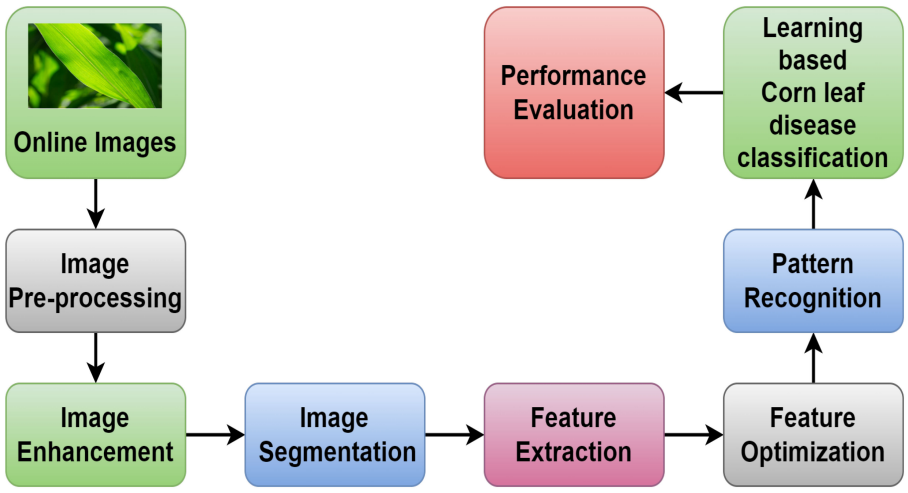


FIGURE 4
A general view of the prediction flow.

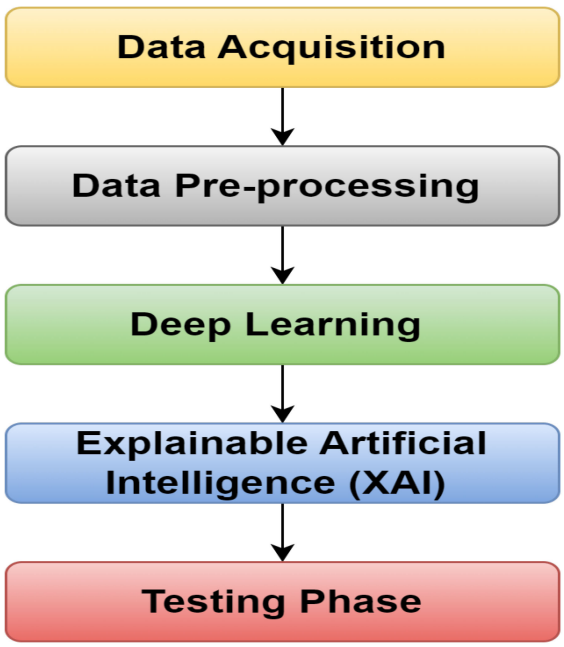


FIGURE 5
Five steps of the proposed model.

pre-trained models such as VGG16 basically makes the development of AI applications easier by using previous knowledge for adaptation to the new tasks (Özden, 2023).

XAI in predictive modeling is the key to the enhancement of the transparency and the trust of the users by giving them information about the decision-making process of an AI model. XAI methods are designed to make complicated models comprehensible to people, thus, it would be easy for users to understand why a particular prediction is made. These techniques produce explanations by focusing on the vital attributes, visualizing the model behavior, or providing context-oriented insights specific to the application domain. Through the introduction of XAI into predictive models, stakeholders will be able to not only acquire useful information about model predictions but also reduce the risks related to bias, mistakes, and the lack of transparency, hence fostering the acceptance and trust of AI systems in the real world (Ullah et al., 2021).

As shown in Table 3, the pseudo-code frames a proposed model for foreseeing corn leaf diseases utilizing XAI with the VGG16 calculation. The interaction includes two phases: training and testing. In the training phase, the dataset (Smaranjit Ghose, 2024) is gathered from Kaggle and then split into training and testing sets. A DL model is then applied to the training dataset, and the model's forecasts are done by utilizing XAI procedures. If the clarifications meet the standards for agreeable execution, the trained model is stored in the cloud. If not, it sets off a retraining cycle. In the testing phase, testing data are used with the trained model and then predict the corn leaf disease. After that, the reasons for corn leaf disease are explained, and such are quite helpful for farmers in smart agriculture systems.

4 Simulation and results

Some key metrics were evaluated to critically examine various aspects of the model's performance. These include accuracy, precision, true positive rate, false positive rate, and misclassification rate (Shahinfar et al., 2020). Accuracy is a performance metric that measures how well a model classifies images, regardless of the classification error type (Valero-Carreras et al., 2023).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100 \quad (1)$$

TP, FP, TN, and FN are symbols that indicate true positives, false positives, true negatives, and false negatives, respectively (Heydarian et al., 2022).

Precision measures how many images were correctly classified by the model as a fraction of the total number of images classified (Heydarian et al., 2022).

$$Precision = \frac{TP}{TP + FP} * 100 \quad (2)$$

The false negative rate measures the fraction of incorrectly classified images from all negative pictures (Renshaw, 1997).

$$False\_negative\_rate = \frac{FN}{FN + TP} * 100 \quad (3)$$

In a confusion matrix, specificity is calculated by taking the TN for a given class and dividing it by the sum of TN and FP for that class (Van Stralen et al., 2009).

$$Specificity = \frac{TN}{TN + FP} * 100 \quad (4)$$

The misclassification rate in the confusion matrix represents the proportion of cases classified incorrectly by the model (Llullaku et al., 2009).

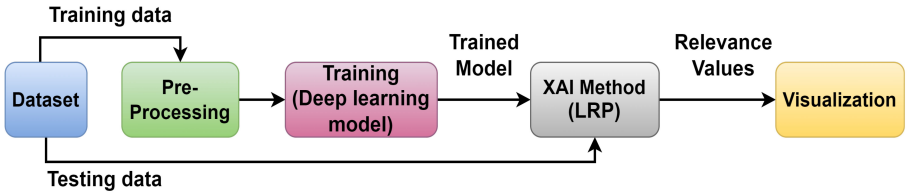


FIGURE 6
A general view of the proposed model.

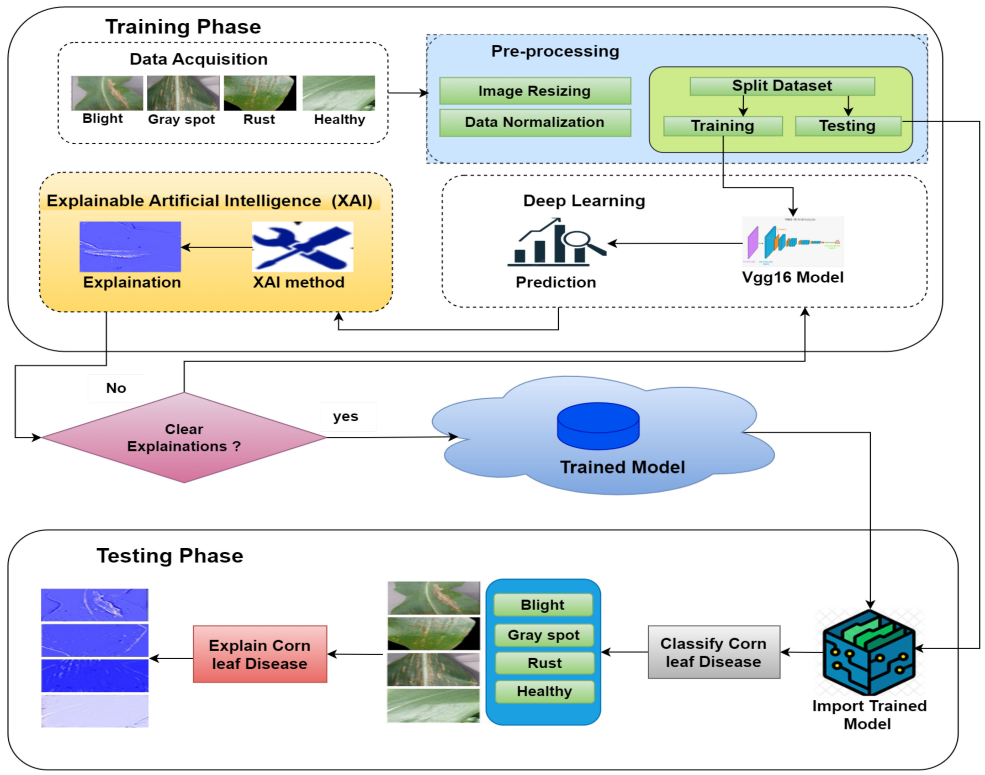


FIGURE 7
Detailed proposed model for predicting corn leaf disease with XAI using VGG16.

$$Misclassification\_rate = \frac{FN + FP}{TP + TN + FP + FN} * 100 \quad (5)$$

The evaluation metrics are used to measure the efficiency and effectiveness of the proposed approach, as shown in Equations 1–5.

The suggested model classifies images of corn leaf diseases into four categories: blight, common rust, gray leaf spot, and healthy. The model's training parameters are the number of epochs, optimization algorithm, input image size, batch size, and learning rate, as shown in Table 4.

Table 5 shows a confusion matrix generated during the training process for classifying corn leaf diseases. The matrix lists the actual classes as rows and the predicted classes as columns. Each matrix cell represents the number of instances classified accordingly during training. For example, the model accurately classified 800 instances of blight, 914 instances of common rust, 400 cases of gray leaf spot, and 813 instances of healthy. The off-diagonal elements of the matrix indicate the misclassifications, such as two instances of blight being incorrectly classified as gray leaf spot and one example of gray leaf spot being misclassified as blight. This matrix is a valuable tool for evaluating the performance of the classification model, helping to identify areas of accurate classification and pointing where errors occur, so the training results of the model are as illustrated in Figure 8.

TABLE 3 The pseudo-code of the proposed model with XAI using VGG16.

Training phase
1-Data Acquisition *raw_dataset = acquire_raw_dataset_from_Kaggle()* 2-Preprocessing preprocessed_dataset = *preprocess_dataset(raw_dataset)* Splitting into training and testing sets *training_set, testing_set = split_dataset(preprocessed_dataset)* 3-Deep Learning *trained_model = train_deep_learning_model(training_set)* 4- XAI explanations = explain_predictions(trained_model, testing_set) **If** *explanations_are_satisfactory(explanations)* Store_model_in_cloud(trained_model) **Else** retrain_model() **EndIf**

(Continued)

TABLE 3 Continued

Testing phase
5-Testing raw_testing_data = collect_raw_testing_data() preprocessed_testing_data = preprocess_testing_data(raw_testing_data) Classification using the trained Model classified_data = classify_data_using_trained_model(trained_model, preprocessed_testing_data) Import and utilize identified and predicted data import_and_utilize_data(classified_data)

TABLE 4 Training parameters.

Training parameters	Values
No. of epochs	10
Batch size	32
Learning rate	0.0001
Optimization algorithm	Adam
Input image size	224 × 224 × 3

Table 6 displays the confusion matrix for the corn leaf disease classification test. The rows represent the actual classes, whereas the columns indicate the predicted classes. Each cell in the table displays the count of instances classified accordingly during the testing phase. For example, the model correctly classified 306 instances of blight, 380 cases of common rust, 156 cases of gray leaf spot, and 349 instances of healthy. However, some misclassifications were observed, such as 33 instances of gray leaf spot being misclassified as blight and seven common rust misclassified as gray leaf spot. The testing accuracy, shown at the bottom of the table, is 94.67%. This represents the proportion of correctly classified instances from the total testing dataset. This matrix provides valuable insights into the model's performance, identifying areas where misclassifications occur. It helps further refine and evaluate the classification model, as shown in Figure 9.

Table 7 presents the per-class performance metrics of a classification model used to identify four different classes of corn leaf disease. The performance metrics, namely, accuracy, precision, false negative rate, specificity, and misclassification rate are all expressed as percentages. The model shows high performance over all the classes in the training phase. In the blight class, accuracy is 99.98%, precision is 99.87%, and specificity is 99.95%, where the false negative rate is 0.24% and misclassification rate is

TABLE 5 Training confusion matrix of the proposed model.

Actual/predicted	Blight	Common rust	Gray leaf spot	Healthy
Blight	800	0	2	0
Common rust	0	914	0	0
Gray leaf spot	1	0	400	0
Healthy	0	0	0	813
Training accuracy	99.89%			

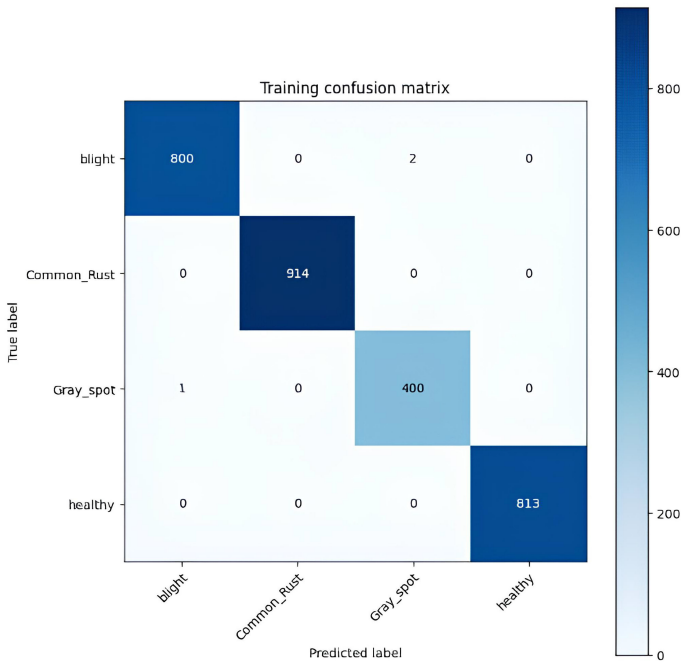


FIGURE 8
Training confusion matrix of the proposed model.

TABLE 6 Testing confusion matrix of the proposed model.

Actual/predicted	Blight	Common rust	Gray leaf spot	Healthy
Blight	306	5	33	0
Common rust	5	380	7	0
Gray leaf spot	14	2	156	1
Healthy	0	0	0	349
Testing accuracy	94.67%			

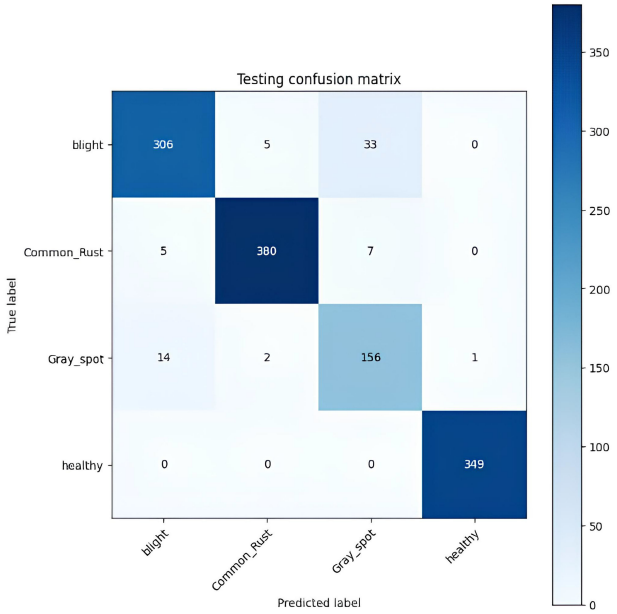


FIGURE 9
Testing confusion matrix of the proposed model.

0.1023%; common rust and healthy class show 100% accuracy, precision, and specificity, and have 0% false negative rate and misclassification rate; and in the gray leaf spot, the accuracy is 99.89%, precision is 99.50%, and specificity is 99.92%, where the false negative rate is 0.249% and misclassification rate is 0.1023%. Blight and gray leaf spots show equal accuracy and misclassification rates. Overall, the model shows high accuracy in identifying corn leaf diseases in the training phase.

Table 8 provides a detailed evaluation of a classification model's performance on the four classes of corn leaf disease in the testing phase. The performance of each class is assessed using five key metrics, namely, accuracy, precision, false negative rate, specificity, and misclassification rate. In the blight class, the accuracy is 95.46%, precision and specificity are 94.15%, the false negative rate is 11.046%, and the misclassification rate is 4.53%. The common rust class achieved 98.48% accuracy, precision is 98.19%, and

TABLE 7 Per-class performance metrics in training.

Classes	Accuracy (%)	Precision (%)	False negative rate (%)	Specificity (%)	Misclassification rate (%)
Blight	99.89	99.87	0.24	99.95	0.1023
Common rust	100	100	0	100	0
Gray leaf spot	99.89	99.50	0.249	99.92	0.1023
Healthy	100	100	0	100	0

TABLE 8 Per-class performance metrics in testing.

Classes	Accuracy (%)	Precision (%)	False negative rate (%)	Specificity (%)	Misclassification rate (%)
Blight	95.46	94.15	11.046	94.15	4.53
Common rust	98.48	98.19	3.06	99.19	1.51
Gray leaf spot	95.46	79.59	9.82	96.31	4.53
Healthy	99.92	99.71	0	99.88	0.07

specificity is 99.19%. The gray leaf spot has the same accuracy as the blight class, but the precision is 79.59%, the false negative rate is 9.82%, the specificity is 96.31%, and the misclassification rate is 4.53%. In the healthy class, the accuracy is 99.92%, precision is 99.71%, specificity is 99.88%, false negative rate is 0%, and misclassification rate is 0.07%. The model shows excellent results with respect to classifying different classes of corn leaf disease in the testing phase.

Table 9 displays the proposed model's performance metrics for training and testing. Figure 10 shows the results of the VGG16 model enhanced with LRP. This technique helps us understand the decisions made by the model. In this study, the LRP and VGG16 models are used to predict different types of corn leaf diseases, such as healthy, blight, common rust, and gray leaf spot. LRP generates results that allow us to understand the features and regions within the corn leaf images that contribute most significantly to the model decision-making process. This approach helps us interpret the model predictions more transparently and explainable, making it possible for researchers and practitioners to test the model performance. It also helps identify areas that require improvement or refinement in the classification task.

Table 10 compares the different models used to predict corn leaf disease, showing their accuracy, loss rate, and whether they used XAI. The suggested model used VGG16 with LRP and reached 94.67% accuracy. This is the only model that uses XAI to give transparency in the decision-making process. In other models compared, different AI models had lower accuracies and did not use XAI techniques. So, the proposed model achieves a good balance between high performance and interpretability using XAI.

TABLE 9 Overall performance metrics of training and testing.

Performance metrics	Training (%)	Testing (%)
Accuracy	99.89	94.67
Precision	99.84	92.91
False negative rate	0.12	5.98
Specificity	99.96	98.32
Misclassification rate	0.11	5.33

5 Conclusion

In this paper, the model VGG16 is employed to deal with the images used to detect the disease of corn leaves. This model achieves a better performance in terms of accuracy, specificity, misclassification rate, and false positive rate with respect to previously published works. The LRP with VGG16 model is used to accurately diagnose corn leaf diseases in agriculture. This technique makes it possible for farmers to get information

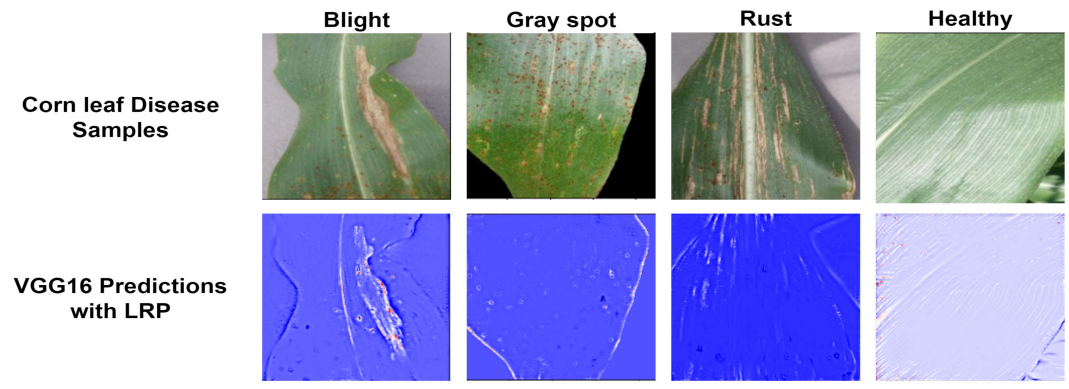


FIGURE 10
Results of the proposed model.

TABLE 10 Comparison of the proposed model with related works to predict corn leaf disease.

Ref.	Model	Accuracy (%)	Misclassification rate (%)	XAI
Yang et al. (2019)	CNN	82.19	17.81	Non – Explainable Artificial Intelligence (XAI) Used
Agarwal et al. (2019)	CNN	94	6	
Zhen et al. (2020)	RDNet	84.04	15.96	
Saeed et al. (2021)	VGG19, CNN, And PLS	90.01	9.99	
Sandotra et al. (2023)	ResNet50, VGG16, VGG19, InceptionV3, and EfficientNetB0	70.02, 91.37, 89.69, 87.77, and 92.33	29.98, 8.63, 10.31, 12.23, and 7.67	
Proposed model	**VGG16 empowered with LRP**	**94.67**	**5.33**	**Yes**

Bold values show the results of the proposed model.

regarding what is happening with corn leaf diseases at the current moment. This model enables farmers to take action at the right time and utilize various resources efficiently. This method to detect disease in the early steps can be used to prevent crop damage. Therefore, this model enhances farmers' understanding of disease transmission or crop management and other facets of sustainable agriculture with the help of proper explanations and visualizations. This study demonstrates the relevance of XAI in smart agriculture and acts as a foundation for future studies on how explainable methods can be employed to achieve further improvements in the performance and reliability of deep neural networks in agriculture.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

MT: Conceptualization, Software, Writing – original draft. UA: Formal analysis, Software, Visualization, Writing – review & editing. SA: Formal analysis, Methodology, Supervision, Writing – original draft. SH: Data curation, Formal analysis, Software, Writing – review & editing. RN: Funding acquisition, Software, Visualization, Writing – review & editing. MK: Investigation, Methodology, Project administration, Supervision, Writing – original draft. DJ: Data curation, Funding acquisition, Resources, Software, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the faculty research fund of Sejong University in 2024, and the Creative Challenge Research Program (2021R1I1A1A01052521) through the National Research Foundation (NRF) of Korea.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

Agarwal, M., Bohat, V. K., Ansari, M. D., Sinha, A., Gupta, S. K., and Garg, D. (2019). A convolution neural network based approach to detect the disease in corn crop. 176–181. doi: 10.1109/IACC48062.2019.8971602

Alsaleh, M. M., Allery, F., Choi, J. W., Hama, T., McQuillin, A., Wu, H., et al. (2023). Prediction of disease comorbidity using explainable artificial intelligence and machine learning techniques: A systematic review. *Int. J. Med. Inf.* 175, 105088. doi: 10.1016/j.ijmedinf.2023.105088

Asriny, D. M., and Jayadi, R. (2023). Transfer learning vgg16 for classification orange fruit images. *J. System Manage. Sci.* 13, 206–217. doi: 10.33168/JSMS.2023.0112

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* 10, e0130140. doi: 10.1371/journal.pone.0130140

Bridgelall, R. (2024). Alignment of unsupervised machine learning with human understanding: A case study of connected vehicle patents. *Appl. Sci.* 14, 474. doi: 10.3390/app14020474

Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 832. doi: 10.3390/electronics8080832

Dethier, J.-J., and Effenberger, A. (2012). Agriculture and development: A brief review of the literature. *Economic Syst.* 36, 175–205. doi: 10.1016/j.ecosys.2011.09.003

Fadhilla, M., Suryani, D., Labellapansa, A., and Gunawan, H. (2023). Corn leaf diseases recognition based on convolutional neural network. *IT J. Res. Dev. (ITJRD)* 8, 14–21. doi: 10.25299/itjrd.2023.13904

Hagras, H. (2018). Toward human-understandable, explainable ai. *Computer* 51, 28–36. doi: 10.1109/MC.2018.3620965

Heydarian, M., Doyle, T. E., and Samavi, R. (2022). Mlcm: Multi-label confusion matrix. *IEEE Access* 10, 19083–19095. doi: 10.1109/ACCESS.2022.3151048

Khoirunnisak, K. M. (2020). Sistem pakar diagnosa penyakit pada tanaman jagung dengan metode dempster shafer. *Publikasi Tugas Akhir S-1 PSTI FT-UNRAM*. 5, 15–27.

Kusumo, B. S., Heryana, A., Mahendra, O., and Pardede, H. F. (2018). "Machine learning-based for 293 automatic detection of corn-plant diseases using image processing," in *2018 International conference on computer, control, informatics and its applications (IC3INA) (IEEE)*. 93–97.

Li, L., Zhang, S., and Wang, B. (2021). Plant disease detection and classification by deep learning—a review. *IEEE Access* 9, 56683–56698. doi: 10.1109/ACCESS.2021.3069646

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy* 23, 18. doi: 10.3390/e23010018

Llullaku, S. S., Hyseni, N. S., Bytyc¸, i, C.I., and Rexhepi, S. K. (2009). Evaluation of trauma care using triss method: the role of adjusted misclassification rate and adjusted w-statistic. *World J. Emergency Surg.* 4, 1–6. doi: 10.1186/1749-7922-4-2

Makridakis, S., Spiliotis, E., Assimakopoulos, V., Semenoglou, A.-A., Mulder, G., and Nikolopoulos, K. (2023). Statistical, machine learning and deep learning forecasting methods: Comparisons and ways forward. *J. Operational Res. Soc.* 74, 840–859. doi: 10.1080/01605682.2022.2118629

Mardiana, B. D., Utomo, W. B., Oktaviana, U. N., Wicaksono, G. W., and Minarno, A. E. (2023). Herbal leaves classification based on leaf image using cnn architecture model vgg16. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)* 7, 20–26. doi: 10.29207/resti.v7i1

Mgomezulu, W. R., Chitete, M. M., Maonga, B. B., Dzanja, J., Mulekano, P., and Qutieshat, A. (2024). Agricultural subsidies in a political economy: Can collective action make smallholder agriculture contribute to development? *Res. Globalization* 8, 100212. doi: 10.1016/j.resglo.2024.100212

Özden, C. (2023). Corn disease detection using transfer learning. *Black Sea J. Eng. Sci.* 6, 387–393. doi: 10.34248/bsengineering.1322907

Rajamohana, S., Shri, S. P., Nithyapriyaa, V., Parthiban, B., Karthikeyan, A., Tharmentheran, V., et al. (2024). "Analysis of smart agriculture systems using iot," in *Heterogenous Computational Intelligence in Internet of Things*. Florida, United States: CRC Press (Taylor & Francis Group), 74–88.

Renshaw, A. A. (1997). Analysis of error in calculating the false-negative rate in the interpretation of cervicovaginal smears: the need to review abnormal cases. *Cancer Cytopathology: Interdiscip. Int. J. Am. Cancer Soc.* 81, 264–271. doi: 10.1002/(sici)1097-0142(19971025)81:5<264::aid-cncr2>3.0.co;2-n

Saeed, F., Khan, M. A., Sharif, M., Mittal, M., Goyal, L. M., and Roy, S. (2021). Deep neural network features fusion and selection based on pls regression with an application for crops diseases classification. *Appl. Soft Computing* 103, 107164. doi: 10.1016/j.asoc.2021.107164

Sandotra, N., Mahajan, P., Abrol, P., and Lehana, P. K. (2023). Analyzing performance of deep learning models under the presence of distortions in identifying plant leaf disease. *Int. J. Inf Commun. Technol.* 12, 115–126. doi: 10.11591/ijict.v12i2

Shahinfar, S., Meek, P., and Falzon, G. (2020). how many images do i need?" understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring. *Ecol. Inf.* 57, 101085. doi: 10.1016/j.ecoinf.2020.101085

Smaranjit Ghose. (2024). *Corn or maize leaf disease dataset*. Available online at: https://www.kaggle.com/datasets/smaranjitghose/corn-or-maize-leaf-disease-dataset.

Ullah, I., Rios, A., Gala, V., and Mckeever, S. (2021). Explaining deep learning models for tabular data using layer-wise relevance propagation. *Appl. Sci.* 12, 136. doi: 10.3390/app12010136

Valero-Carreras, D., Alcaraz, J., and Landete, M. (2023). Comparing two svm models through different metrics based on the confusion matrix. *Comput. Operations Res.* 152, 106131. doi: 10.1016/j.cor.2022.106131

Van Stralen, K. J., Stel, V. S., Reitsma, J. B., Dekker, F. W., Zoccali, C., and Jager, K. J. (2009). Diagnostic methods i: sensitivity, specificity, and other measures of accuracy. *Kidney Int.* 75, 1257–1263. doi: 10.1038/ki.2009.92

Widmer, J., Christ, B., Grenz, J., and Norgrove, L. (2024). Agrivoltaics, a promising new tool for electricity and food production: A systematic review. *Renewable Sustain. Energy Rev.* 192, 114277. doi: 10.1016/j.rser.2023.114277

Yang, W., Yang, C., Hao, Z., Xie, C., and Li, M. (2019). Diagnosis of plant cold damage based on hyperspectral imaging and convolutional neural network. *IEEE Access* 7, 118239–118248. doi: 10.1109/Access.6287639

Zhen, W., Shanwen, Z., and Baoping, Z. (2020). Crop diseases leaf segmentation method based on cascade convolutional neural network. *Comput. Eng. Appl.* 56, 242–250. doi: 10.3778/j.issn.1002-8331.1905-0193

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 12 July 2024
DOI 10.3389/fpls.2024.1418383

OPEN ACCESS

EDITED BY
Roger Deal,
Emory University, United States

REVIEWED BY
Zhen Fan,
University of Florida, United States
Xu Wang,
University of Florida, United States

*CORRESPONDENCE
Hyoung Seok Kim
hkim58@kist.re.kr

RECEIVED 16 April 2024
ACCEPTED 24 June 2024
PUBLISHED 12 July 2024

CITATION
Ndikumana JN, Lee U, Yoo JH, Yeboah S, Park SH, Lee TS, Yeoung YR and Kim HS (2024) Development of a deep-learning phenotyping tool for analyzing image-based strawberry phenotypes.
Front. Plant Sci. 15:1418383.
doi: 10.3389/fpls.2024.1418383

Development of a deep-learning phenotyping tool for analyzing image-based strawberry phenotypes

Jean Nepo Ndikumana[1,2], Unseok Lee[1], Ji Hye Yoo[1], Samuel Yeboah[1,2], Soo Hyun Park[1], Taek Sung Lee[1], Young Rog Yeoung[2] and Hyoung Seok Kim[1*]

[1]Smart Farm Research Center, Korea Institute of Science and Technology (KIST), Gangneung, Republic of Korea, [2]Department of Plant Science, Gangneung-Wonju National University, Gangneung, Republic of Korea

Introduction: In strawberry farming, phenotypic traits (such as crown diameter, petiole length, plant height, flower, leaf, and fruit size) measurement is essential as it serves as a decision-making tool for plant monitoring and management. To date, strawberry plant phenotyping has relied on traditional approaches. In this study, an image-based Strawberry Phenotyping Tool (SPT) was developed using two deep-learning (DL) architectures, namely "YOLOv4" and "U-net" integrated into a single system. We aimed to create the most suitable DL-based tool with enhanced robustness to facilitate digital strawberry plant phenotyping directly at the natural scene or indirectly using captured and stored images.

Methods: Our SPT was developed primarily through two steps (subsequently called versions) using image data with different backgrounds captured with simple smartphone cameras. The two versions (V1 and V2) were developed using the same DL networks but differed by the amount of image data and annotation method used during their development. For V1, 7,116 images were annotated using the single-target non-labeling method, whereas for V2, 7,850 images were annotated using the multitarget labeling method.

Results: The results of the held-out dataset revealed that the developed SPT facilitates strawberry phenotype measurements. By increasing the dataset size combined with multitarget labeling annotation, the detection accuracy of our system changed from 60.24% in V1 to 82.28% in V2. During the validation process, the system was evaluated using 70 images per phenotype and their corresponding actual values. The correlation coefficients and detection frequencies were higher for V2 than for V1, confirming the superiority of V2. Furthermore, an image-based regression model was developed to predict the fresh weight of strawberries based on the fruit size (R^2 = 0.92).

Discussion: The results demonstrate the efficiency of our system in recognizing the aforementioned six strawberry phenotypic traits regardless of the complex

scenario of the environment of the strawberry plant. This tool could help farmers and researchers make accurate and efficient decisions related to strawberry plant management, possibly causing increased productivity and yield potential.

KEYWORDS

deep learning, strawberry, phenotyping, YOLOv4, U-net

1 Introduction

The cultivated strawberry *Fragaria* × *ananassa* Duchesne is the most economically essential soft fruit worldwide, and its production and consumption are increasing in many parts of the world, including Korea (Simpson, 2018; Menzel, 2020). Given their significance in the global market and mounting year-round demand, strawberries are intensively grown under protected structures to ensure seasonal earliness, high-quality yield, and a continuous annual supply. Strawberry cultivation in Korea is primarily concentrated in greenhouses and is significant in the country's agricultural industry and economy (Hwang et al., 2020). However, cultivating strawberries under such conditions requires extensive input and labor (Ilyas et al., 2021; Khammayom et al., 2022; Mbarushimana et al., 2022). To generate higher outputs and make tangible profits, farmers should optimize resource use efficiency through consistent and timely plant monitoring and make accurate farm management decisions.

Early identification and timely quantification of key plant phenotypes may provide valuable insights that can predict subsequent stages of plant development and critical outcomes such as yield. In the case of strawberry farming, the measurement of phenotypic traits, such as crown diameter (CD), petiole length (PL), plant height (PH), flower, leaf, and fruit size, is common among growers and researchers, serving as phenotypic markers employed to monitor plant growth balance and manage cultivation conditions. The crown size of strawberry seedlings during the transplanting stage has been established as a reliable indicator of post-transplantation vigor, and transplants with initially larger CDs are associated with high-yield strawberry components (Fridiaa et al., 2016; Fagherazzi et al., 2021). PL and PH are used to assess the overall growth potential of strawberries, and the petiole size can be used as an indicator of plant dormancy, where the plant produces shorter petioles in dormant conditions (Robert et al., 1999; Sønsteby and Heide, 2006). Similarly, PH is considered an index of plant management among strawberry producers (Takahashi et al., 2020). The leaf size is vital as, and in addition to photosynthesis and transpiration, it guides cultural practices, such as plant training, pruning, irrigation, and nutrition supply (Takahashi et al., 2020; Zheng et al., 2021). In addition, the analysis of leaf area and climatic variables can be used to predict plant evolution and the quality of strawberry fruits (de Castro et al., 2020). Similarly, it has been argued that increased leaf size and number may cause an increased fruit yield (Ahn et al., 2021). Flower and fruit size are significant factors in strawberry plant productivity and yield predictions (Chen et al., 2019; Menzel, 2020). To acquire growth information on the abovementioned phenotypic traits, most growers rely on traditional visual and manual phenotyping approaches, which are highly criticized for being subjective, destructive, and error-prone (He et al., 2017; Mahmud Sultan et al., 2020). Thus, to overcome these limitations, farmers of intensive cash crops, such as strawberries, and researchers need a robust, fast, and cost-effective phenotyping tool to facilitate their daily farm management based on quantitative phenotypic data during the plant's life cycle.

Plant phenotyping combined with computer vision approaches provide better non-destructive options for plant monitoring through quantitative and qualitative analyses of complex plant traits, such as plant morphology, plant stress, crop yield, and plant physiological and anatomical traits (Costa et al., 2019). In the current state-of-the-arts in plant phenomics, many plant phenotyping methods are available, among which visible spectral imaging combined with deep-learning (DL) techniques present reliable advantages regarding affordability and quick measurement owing to the availability of various plant phenotyping hardware and software systems that facilitate image registration, processing, and data extraction (Fiorani and Schurr, 2013). Neural network-based DL techniques can be used to extract and analyze meaningful information on various plant traits from many collected image data; therefore, it is proposed that DL will dominate the future trends in image-based plant phenotyping (Tsaftaris et al., 2016). Among the recently published studies where DL techniques were applied for plant phenotyping, it is notable that convolutional neural networks, such as You Only Look Once (YOLO, also called single-shot detectors) and U-net, are among the most extensively used methods for object detection and segmentation (Zhang et al., 2018; Zheng et al., 2019; Ni et al., 2020).

Abbreviations: CD, Crown Diameter; DL, Deep-learning; Fl. A, Flower Area; Fr. A, Fruit Area; LA, Leaf Area; LL, Leaf Length; LW, Leaf Width; mPA, mean Precision Average; PH, Plant height; PL, Petiole Length; SPT, Strawberry Phenotyping Tool; V1, Version 1; V2, Version 2; VIA, VGG Image Annotator.

Studies involving a combination of image processing techniques and computational intelligence to acquire strawberry growth phenotypic information in the field or laboratory settings using various sensors and platforms of different scales have been conducted focusing on detection (Gan et al., 2020; Zhou et al., 2020; Shin et al., 2021), segmentation (Pérez-Borrero et al., 2020; Perez-Borrero et al., 2021), classification (Feldmann et al., 2020; Aish, 2021; Fatehi and Akhijahani, 2021), and quantification (Lee et al., 2017; de Castro et al., 2020). However, most available phenotyping methods for strawberries are research-scale, costly, and unaffordable for ordinary profit-oriented farmers or researchers with limited financial means. Additionally, although DL techniques have been explored in strawberries, they are limited mostly to qualitative fruit attributes, and extensive studies embracing major strawberry growth and development indicators, such as CD, plant height, leaf size, and fruit size, remain unavailable. Therefore, developing a cheap, precise, and high-throughput phenotyping tool that covers more parameters is essential and should sustainably advance farming efforts in the strawberry sector.

In this study, we developed a Strawberry Phenotyping Tool (SPT) based on deep learning (DL), which integrates two prominent DL architectures notably "YOLOv4" and "U-net" into a single system. This image-based tool was designed for phenotyping and data analysis of strawberry plant phenotypic traits focusing on six important growth and yield traits: plant height, petiole length, crown diameter, leaf characteristics (area, length, and width), as well as flowers and fruits. The accuracy and reliability of this tool was enhanced by increasing the number of training images and diversified annotation techniques. We expect that if SPT is integrated into the current strawberry farming systems, it is likely to alleviate various farm management existing challenges and boost strawberry farmers' productivity.

2 Materials and methods

2.1 Training image datasets acquisition

The strawberry-image-based SPT was developed using strawberry images collected from the Korean domestic cultivar 'Seolhyang', which was grown in the greenhouse facility of the Korea Institute of Science and Technology, Gangneung-si, Gangwon-do, Republic of Korea. Images of six phenotypes (crown, plant height, leaf, leaf petiole, flower, and fruit) of agronomic significance for strawberry growth and yield were acquired at variable distances using modern smartphones (iPhone 6S Plus, Apple Inc., United States and Galaxy S8 Samsung Electronics, South Korea) with iOS and Android operating systems for several days during the daytime (6:00–18:00) throughout the winter growing season between September 2019 and May 2020. The sampling devices used (smartphones) had both cameras of 12-megapixel, the exposure parameter was set automatically, and the objective focus system was set to autofocus mode.

In each instance, the image was acquired by steadily holding the Quick Response (QR) marker beside (parallel to) the target object, with the smartphone camera held perpendicularly against it, obtaining an image that included the target object and the QR marker. Spatial calibration for our measurement algorithm was conducted using a QR code (4.7 cm x 4.7 cm) as a reference. Both the growth parameters and the QR code were captured in a single image. The algorithm then recognized the QR code and its four corner points in the image, and applied distortion correction and length conversion based on those points for spatial calibration. The correct positioning and pose of the QR marker are critical for accurate image analysis (Teoh et al., 2022). Incorrectly positioned QR markers can cause overestimation or underestimation of the size of objectives captured in the same image (Data not shown).

Each phenotypic parameter has key areas detected or segmented by the deep learning model. Once these areas are detected, the distances between them are measured, and the lengths are converted based on the pixel size calibrated by the QR marker. Therefore, the parts that need to be parallel to the QR marker vary slightly for each phenotypic parameter. When the leaf area is calculated from the image, the entire leaf surface must be aligned and flat, parallel to the QR marker in the same image. These features also require hand support from the person taking the measurements during the process of acquiring phenotypic images. The SPT was designed for a single user to measure strawberry plants independently. Therefore, one hand holds the camera while the other hand holds the plant part parallel to the QR marker. This operation works smoothly when the strawberry plants are managed properly through the conventional pruning and defoliation. Our experiment was conducted under conventional strawberry cultivation practices where the SPT could operate smoothly. We provide explanations and example photos in Appendix A on how to position the QR marker for each phenotypic parameter during strawberry phenotyping, as well as how to perform the necessary hand support actions.

We initially collected 7,116 images to develop the first version of SPT (V1), and 7,850 images were used to construct the second version of SPT (V2). To obtain an RGB image dataset with a thorough variability of strawberry phenotypes under their natural habitat, the images were collected under different light intensity conditions (cloudy or sunny) and interferences, different days, and different growth stages. For each parameter, the shooting angle and shooting distance were continuously changed to collect images with various colors, postures, sizes, and backgrounds. The collected images were in the JPEG format, manually transferred, and stored in a computer for further processing. The detailed implications of each target phenotypic trait for strawberry farming and management are summarized in Table 1.

2.2 Image dataset construction, annotation, and model training

Our SPT was developed primarily through two phases subsequently named versions (V1 and V2). The primary distinction between V1 and V2 lies in the volume of image data and the specific annotation method employed during their development. The number of images used to develop the two

TABLE 1 Targeted strawberry phenotypes for imaging and features extracted.

Target phenotype	Features extracted in the image	Possible application of phenotyping results
1. Crown	Crown diameter (CD)	Strawberry yield is linked with the initial crown size (Torres-Quezada et al., 2015). Continuous monitoring of CD can provide early signs (prediction) of a plant's vigor and yield.
2. Plant height	Height of the plant (PH)	Tracking strawberry growth strength and speed through PH monitoring (Takahashi et al., 2020).
3. Petiole	Length of the petiole (PL)	PL size is linked to plant activity status (Robert et al., 1997). Dynamic size changes in PL of the strawberry leaf can be used to determine the fate of strawberry plants under field conditions.
4. Leaf	-Leaf area (LA) -Leaf length (LL) -Leaf width (LW)	Leaf size attributes (LA, LL, and LW) can assist in the modeling of photosynthesis, evaporation, and evaluation of crop growth and productivity (Takahashi et al., 2020; Zheng et al., 2021; Jo et al., 2022).
5. Flower	Flower area (Fl. A)	Non-destructive prediction of strawberry qualitative and quantitative yield based on flower number and size (Chen et al., 2019).
6. Fruit	Fruit area (Fr. A)	Non-destructive prediction of strawberry qualitative and quantitative yield based on fruit number and fruit size (Hortyński et al., 1991).

versions of the current phenotyping tool is presented in Table 2 and the annotation principles adopted for each version are illustrated by Figure 1. A batch of 7,116 images was initially collected and manually classified according to their phenotypes before annotation. Figure 2A shows the workflow of the two versions of the model training process. The original dataset was divided into training, validation, and test datasets at 8:1:1 ratio. Subsequently, the test set was excluded from the training set. The VGG Image Annotator (VIA) tool (version 1.0.5), an image annotation tool developed by Dutta and Zisserman (2019), was used to manually annotate the objects of interest to obtain ground truth information for the subsequent training of V1. Because our system was built

TABLE 2 Number of images collected and annotated for each version of the Strawberry Phenotyping Tool (SPT).

Phenotypic trait	Number of images used for model development		Number of images used for validation	
	Version 1	Version 2	Version 1	Version 2
1. Crown	698	950	70	70
2. Plant height	804	940	70	70
3. Petiole	779	915	70	70
4. Leaf	1,490	1,560	70	70
5. Flower	853	923	70	70
6. Fruit	2,492	2,562	70	70
Total	7,116	7,850	420	420

based on YOLOv4 (Bochkovskiy et al., 2020) and U-net (Ronneberger et al., 2015), an appropriate annotation technique that suffices for training these two architectures was considered. The YOLO algorithm series requires bounding box annotation for object localization to identify and detect specific objects in images. In contrast, U-net requires a class label and a pixel-level mask with an outline annotation of an object for semantic segmentation. Therefore, the annotation principle adopted to train these DL architectures needed to satisfy the above conditions. For phenotypic traits, such as crown, plant height, and petiole length, the regions of interest were annotated using a bounding box, whereas other phenotypes, such as leaves, flowers, and fruits, were annotated using polygon-shaped regions (Figure 1). To annotate the CD, a bounding box enclosing the thickest part of the strawberry crown was drawn with the lower side passing through the point of attachment of the leaves to the crown. To annotate the PH, two bounding boxes were used, with each placed at the bottom (crown) and the top (leaves) of the plant's most extreme boundaries. The PL was annotated by drawing two bounding boxes, with each at the plant's point of attachment of the leaves to the crown and at the point of attachment of the three leaflets to the petiole. The leaves, flowers, and fruits were annotated by carefully drawing polygons around the middle leaflets, flowers, and fruits, respectively. After annotation, the annotation information was downloaded and saved in CSV format and used to train and create the first version of our system.

Owing to the poor performance of V1, two approaches have been adopted to improve and change it to V2. First, we increased the number and variability of the training datasets. A total of 734 additional images were previously collected differently and added to the previous batch to make up 7,850 images. The PH and PL were targeted and acquired from a single image. Second, the annotation method was changed using an updated VIA version (VIA 2.0.10). Unlike V1, to annotate the second batch of images, all images were mixed to make a single folder, and all the objects to be annotated were pre-defined in the VIA 2.0.10 annotator and assigned class

10.3389/fpls.2024.1418383

FIGURE 1
Annotation approaches for the six target phenotypic traits. **(A)** displays the raw images of six phenotypic traits: crown diameter, plant height, petiole length, leaf size, flower size, and fruit size. **(B)** shows the annotations for Version 1 (V1), where traits are marked with bounding boxes for crown diameter, plant height, and petiole length, and with polygons for leaf, flower, and fruit sizes. **(C)** illustrates the improved annotations for Version 2 (V2), which include descriptive labels for each trait and employ multi-target annotations within a single image when multiple traits are present.

labels (Figure 1). Therefore, because more than one target objects co-occur in a single image, more than one target could be annotated in the same photograph (multitarget annotation). Similarly, to V1, the training, validation, and testing sets were fixed at 8:1:1 ratio. The resulting annotation information was downloaded, saved in JSON format, and used for V2 training and construction.

The images, annotation results and DL models subjected to V1 and V2 of STP are available at https://github.com/kist-smartfarm/SPT.

2.3 SPT SW architecture: YOLOv4 and U-net-based detection and segmentation pipeline and features extraction

The strawberry phenotype analysis pipeline workflow of the SPT is displayed in Figure 2B. We measured the phenotypic traits of strawberries in two ways using our system. First, the CD, PH, and PL were measured based on object detection (i.e., rectangular boxes). Second, the areas of flowers, fruits, and leaves were measured via object detection and segmentation (i.e., rectangular boxes and pixel-wise classification). As the U-net and YOLOv4 frameworks function differently, we designed a combination of YOLO v4 and U-net networks into a single system to build a robust and standalone SPT. YOLOv4 (Bochkovskiy et al., 2020) is a high-precision, real-time, one-stage object detection algorithm that involves using previous YOLO algorithms (Redmon et al., 2016; Redmon and Farhadi, 2017, 2018) with CSPDarknet53, PANet, and mosaic data augmentation. YOLOv4 performance was improved in previous YOLO algorithms through experiments and showed good performance in detecting small objects. Thus, we adopted this algorithm because it is suitable for detecting small objects, particularly strawberry fruits. For strawberry flower, fruit, and leaf measurements using our system, U-net-based semantic segmentation was adopted. U-net was designed to segment biomedical images in the original study (Ronneberger et al., 2015). The network is robust to small and thin objects, such as flowers, fruits, and petioles. Segmentation (i.e., pixel-wise classification) was performed using the object detection results, that is, a cropped detection image. Subsequently, we used our system to calculate the number of pixels (i.e., the area) based on the segmented results. Finally, all measurement results (e.g., length and area) were converted from pixels to actual distance or area using the detected QR code information. Our SPT equipped with the V2 DL was implemented on the web (https://www.cultigrowth.com) for the test of SPT with the

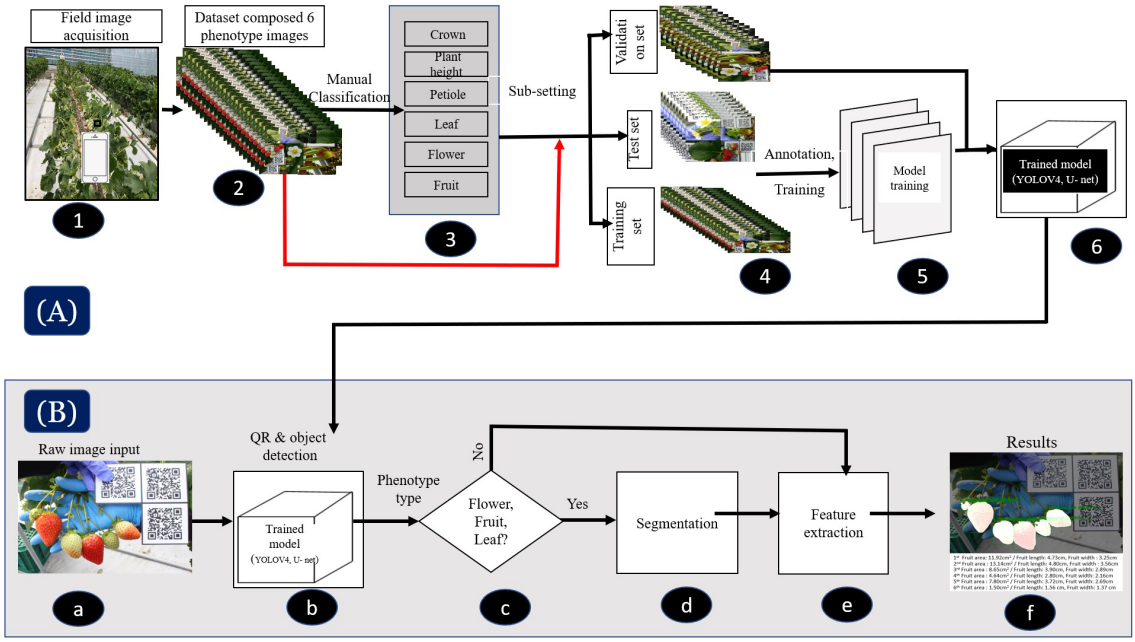


FIGURE 2
Flowchart of the deep learning-based phenotyping tool (SPT), which comprises two primary models: YOLOv4 and U-Net. **(A)** Illustration of the image acquisition, processing, and training procedures of the models for both versions. In Version 1, the models were trained through six processes (1–6), while in Version 2, they were trained through five processes (1, 2, 4–6). **(B)** (a–f) depicts the strawberry phenotyping analysis pipeline workflow using SPT, which involves image acquisition, detection, classification, segmentation and/or feature extraction, and results visualization.

independent datasets. The general information about how the SPT analyzes the studied strawberry phenotypic traits is found in section 3.5. For more detailed instructions on using the cloud-based SPT on the Cultilabs homepage, refer to the appendix A.

2.4 Ground truth data acquisition

To validate the functionality of our DL-based phenotyping tool, 70 plants (bearing flowers and fruits) were sampled from the Korean Institute of Science and Technology strawberry smart farm during the experimental period, and 70 images for each of the six phenotypes (crown, petiole, plant height, leaf, flower, and fruit) were captured using a smartphone camera and a QR marker. The resultant image files were renamed sample-wise to facilitate tracking and maintained for subsequent use as a validation dataset for the developed phenotyping tool. On the same day, after collecting digital images, the corresponding real data were acquired from the same plant parts directly through manual measurements or indirectly using phenobox-based methods (Czedik-Eysenberg et al., 2018). Manual data were collected using a measuring tape (PL and PH) and an electronic clipper (CD). The values of the remaining parameters, such as leaf area, leaf length, leaf width, flower area, and fruit area, were extracted from the phenobox-captured images using the Plant Analysis tool (Korea Scientific Technique Industry, Suwon, Republic of Korea). Additionally, the fresh weight of strawberries from the fruit cluster samples was recorded. The measured weight was used to construct a prediction model for strawberry fresh weight based on fruit size using our digital phenotyping tool. To evaluate the effectiveness of the Strawberry Phenotyping Tool (SPT) in monitoring and managing strawberry plants under greenhouse conditions, 'Seolhyang' strawberry cultivar transplants were sourced from a professional farmer in Pyeongchang, Gangwon-do, Republic of Korea. These transplants were planted at the Korean Institute of Science and Technology hydroponic greenhouse located in Gangneung-si, Gangwon-do, during the winter season from September 24, 2021, to April 30, 2022. The plants were cultivated in rectangular multi-potted containers measuring 20 cm x 15 cm x 60 cm, each equipped with six holes. The pots, filled with a soilless commercial medium and planted with 'Seolhyang' seedlings, were placed on 1-meter-high raised beds to enhance management accessibility. Cultivation was conducted using a standard hydroponic system, in compliance with established protocols for Korean strawberry farming within controlled environments. Weekly data on crown diameter, petiole length, plant height, and the sizes of leaves, flowers, and fruits were collected using both SPT and conventional tools on 27 randomly selected samples throughout the cultivation period.

2.5 Statistical analysis

After the completion of training, both models underwent consistent validation using a reserved subset of the dataset that had not been used during training. To evaluate the accuracy and reliability of the Strawberry Phenotyping Tool (SPT) throughout the development process, Pearson correlation coefficients and

visualization techniques were utilized to compare measurements obtained from the SPT with those obtained conventionally. Paired *t*-tests were then employed to statistically validate the improvements achieved by different versions of the SPT, determining which version produced values most comparable to those obtained conventionally. These tests were crucial for assessing whether the enhancements in phenotypic detection by each version of the SPT were significantly different from those obtained through conventional measurements.

During the field validation stage, the impact of phenotypic variations on fruit yield was explored by categorizing data collected from various phenotypic traits into distinct size clusters before fruit harvesting. For example, at the transplantation stage, crown diameter (CD) was categorized into three clusters: Cluster 1 for samples with large crowns, Cluster 2 for samples with medium crowns, and Cluster 3 for samples with small crowns. The K-means clustering algorithm was utilized to ensure effective categorization based on phenotypic sizes. Similarly, critical phenotypic parameters such as total flower area per plant, total weight of unripe fruits per plant, and leaf area per plant were also clustered. This clustering facilitated structured comparisons of weekly yields across different phenotypic categories. The comparisons were conducted using analysis of variance (ANOVA) and Tukey's Honestly Significant Difference (HSD) test to determine the statistical significance among the groups at a significance level of $p<0.05$. These comparisons allowed us to explore whether larger phenotypic sizes correlated with higher yields.

3 Results

3.1 Dataset size and annotation method effect on DL performance

The DL-based strawberry phenotyping digital tool was developed in two primary steps (subsequently called versions). The difference between these two versions is primarily owing to the different numbers of images and annotation techniques used during the development process. Table 2 presents the different numbers of images collected per phenotypic trait and used to develop the two SPT versions, and Figure 1 shows the annotation principle adopted correspondingly. V1 was initially developed using 7,116 images representing various phenotypic traits of interest. Annotation was performed by drawing a bounding box encompassing the target part for the CD, PL, and PH, and a polygon was drawn to include leaf, flower, and fruit areas. Because of the low detection frequency and precision of SPT-V1, we increased the number of images to 7,850 and changed the annotation techniques which led to V2. In V1, the bounding boxes for annotating the upper boundaries for PH and PL were drawn such that all leaves were enclosed inside the boxes; in V2, these boxes were reduced to include only the highest leaf part for PH and the junction point of the petiole and its leaflets for PL. Additionally, while in V1, each phenotype was annotated individually for each image without labels, in V2, more than one phenotype was annotated in one image. The resulting V2 of SPT showed higher detection precision (Figure 3) and frequency (Figure 4).

3.2 Ground truth versus image-based plant phenotype measurements

The measurements obtained through conventional methods (measured values) were systematically compared with those predicted using the two versions of the Strawberry Phenotyping Tool (SPT) (predicted values). Two analytical approaches were utilized to ensure robust evaluation and avoid potentially misleading conclusions based on the data presented in Table 3.

Initially, the analysis was conducted using Pearson correlation analysis (Figure 5), which confirmed the positive correlation between the measured and predicted values for all examined phenotypic traits. The analysis revealed that the fruit area exhibited the highest linear correlation, demonstrating exceptional precision in predictions with the highest R^2 and relatively low RMSE values, emphasizing the tool's accuracy in capturing this trait's variability. On the other hand, crown diameter displayed the lowest linear correlations in both SPT versions. Highly statistical significance was confirmed for the correlations across all variables ($p<0.001$), reinforcing the reliability of the predictions made by the Strawberry Phenotyping Tool (SPT). Comparative analysis showed that Version 2 (V2) consistently demonstrated higher correlation coefficients (R^2) and generally lower RMSE values than Version 1 (V1) across most phenotypic traits. This improvement highlights V2's enhanced algorithmic performance and overall efficacy in predicting phenotypic traits more accurately. The results collectively underline the significant advancements in V2, offering more reliable and precise measurements critical for strategic strawberry plant monitoring and management under controlled farming systems

Furthermore, measured values were compared against the predictions made by SPT, focusing on the average values calculated for each phenotypic trait. Similarly, across both versions of the SPT, the predicted averages generally approximated the conventionally measured values closely. Specifically, except for leaf length, the average values predicted by SPT Version 2 did not significantly differ from the real values, as shown by *p*-values greater than 0.05. However, in SPT V1, notable differences were observed in traits such as crown diameter, plant height, leaf length, and leaf width where the average values differed significantly from the ground truth values, as indicated by *p*-values less than 0.05. This differential accuracy highlights the variations in the effectiveness of the two SPT versions when estimating specific phenotypic traits.

3.3 DL-based regression model for predicting strawberry fresh weight

As previously detailed in the methodology section, our study employed a specialized dataset comprising 420 images, distributed evenly among the six target phenotypic traits, with 70 images dedicated to each trait. This dataset was intentionally prepared to evaluate the performance of the Strawberry Phenotyping Tool (SPT) across different developmental phases. Specifically targeting the 'fruit size' trait, strawberries from the samples corresponding to the fruit size image batch were harvested immediately after imaging

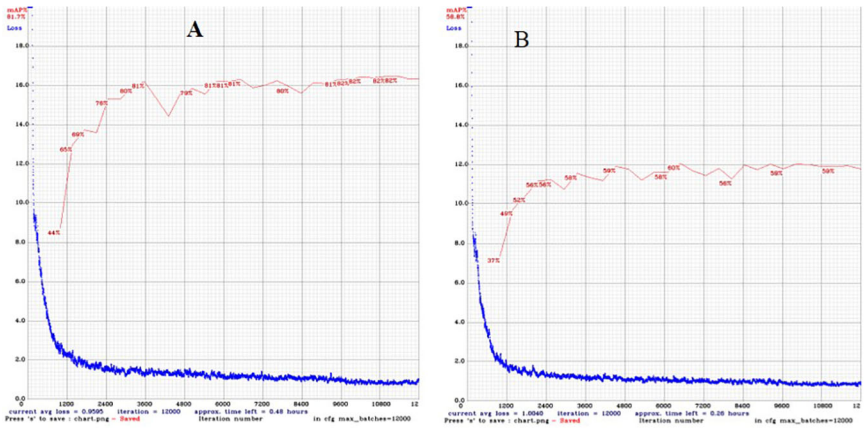


FIGURE 3
Change in training loss and validation mean precision average with the number of epochs of **(A)** Version 1 and **(B)** Version 2 using the training dataset.

to ensure data accuracy and freshness and, subsequently, measurement of each strawberry's fresh weight was conducted using an HR-200 electronic balance (A&D Company, Limited, Tokyo, Japan) by ensuring that each fruit's weight and its precise position within the images were carefully maintained. These fruits included fruits of various sizes and developmental stages and they were used to develop a regression model to predict strawberry fruit fresh weight based on the measured fruit area. The results are illustrated in Figure 6 and show a strong positive correlation between the predicted fruit sizes from SPT and the actual measured weights. Both Version 1 and Version 2 demonstrated relatively similar results, with Version 1 achieving an R^2 value of 0.90 while Version 2 showed a slight improvement with an R^2 value of 0.91

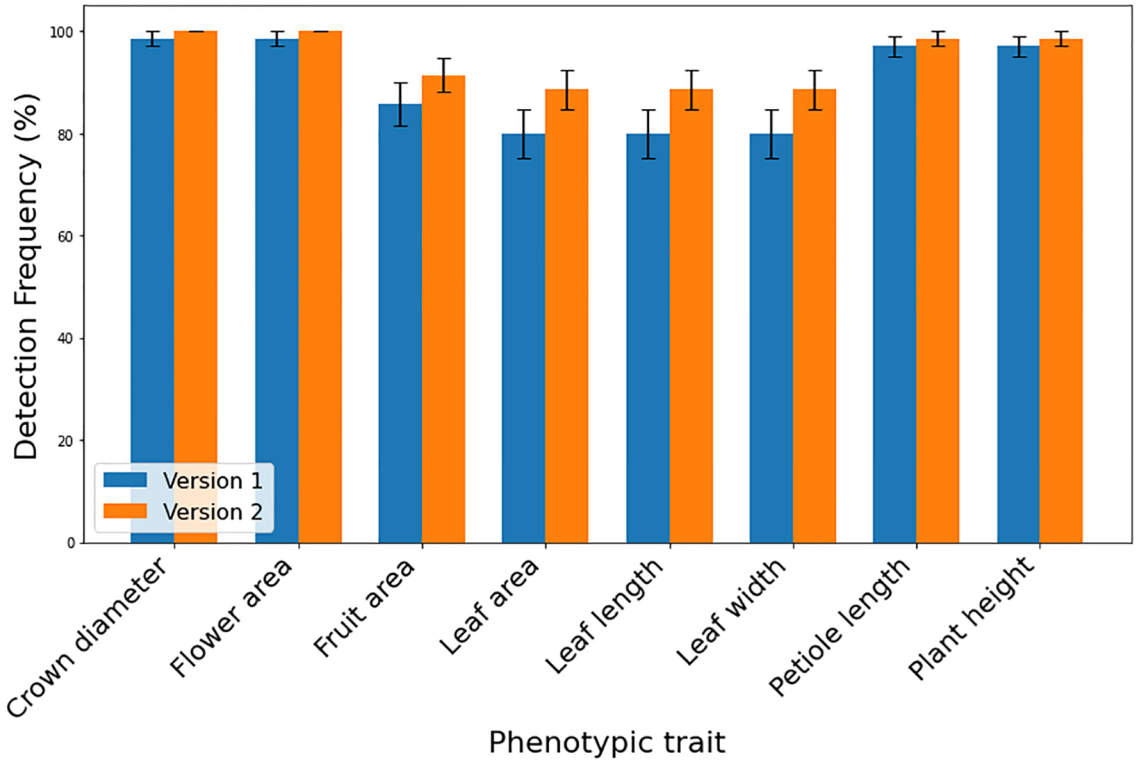


FIGURE 4
Detection frequencies of various strawberry target phenotypic traits using different SPT versions (V1 and V2) (n=70) displayed as percentages. Error bars shows the standard error of the detection frequencies.

TABLE 3 Comparison of strawberry phenotypic traits measured conventionally and with SPT-V1 and SPT-V2.

Phenotypic trait	Measurements of phenotypic trait			P-values	
	Conventional	SPT-V1	SPT-V2	Conventional vs. SPT-V1	Conventional vs. SPT2
1. Crown diameter (mm)	14.36 ± 2.62	14.84 ± 2.56	15.23 ± 2.62	0.027*	0.058
2. Plant height (cm)	24.62 ± 5.36	22.95 ± 5.01	23.17 ± 5.09	0.038*	0.051
3. Petiole length (cm)	15.09 ± 4.49	14.61 ± 4.06	14.87 ± 4.10	0.192	0.322
4.1. Leaf area (cm^2)	25.67 ± 8.65	25.9 ± 7.26	27.10 ± 9.49	0.437	0.187
4.2. Leaf length (cm)	7.04 ± 1.29	6.39 ± 1.08	6.56 ± 1.18	0.001*	0.015*
4.3. Leaf width (cm)	5.51 ± 1.02	5.11 ± 0.77	5.31 ± 0.94	0.007*	0.17
5.1. Flower area (cm^2)	3.26 ± 1.04	3.05 ± 1.08	3.06 ± 1.04	0.287	0.311
6. Fruit area (cm^2)	6.52 ± 3.73	6.32 ± 3.68	6.41 ± 3.72	0.311	0.372

Means ± standard deviation (SD) are shown for each variable, along with corresponding P-values for paired *t*-tests for comparing conventional measurements with the two SPT versions (SPT-V1 and SPT-V2) measurements. The p-values are presented in the last two columns. A p-value less than 0.05 (*) indicates a statistically significant difference between the two means.

3.4 SPT can be used to monitor strawberry growth and predict yield under greenhouse settings

The Strawberry Phenotyping Tool (SPT) was successfully utilized at the Korean Institute of Science and Technology hydroponic greenhouse to monitor the 'Seolhyang' strawberry cultivar. The tool proved to be as effective as traditional methods in tracking the growth and yield phenotypic parameters across 27 samples from September 24, 2021, to April 30, 2022. Figures 7A–D depicts how SPT accurately captured the temporal dynamics and patterns of crown diameter, plant height, leaf length, and leaf width, while also providing additional insights beyond the capabilities of conventional methods. Notably, SPT excelled in measuring leaf area and monitoring the occurrence and sizes of flowers and fruits. Figure 7E illustrates SPT's ability to track the decrease in leaf area as plants matured, offering valuable insights into the leaf-changing pattern during the crop cycle. Furthermore, Figure 7F demonstrates SPT's unique capability to monitor the sizes of flowers, unripe (non-harvestable), and ripe (harvestable) fruits over time, a feature unavailable with manual measurement methods. These advanced functionalities highlight the comprehensive understanding of

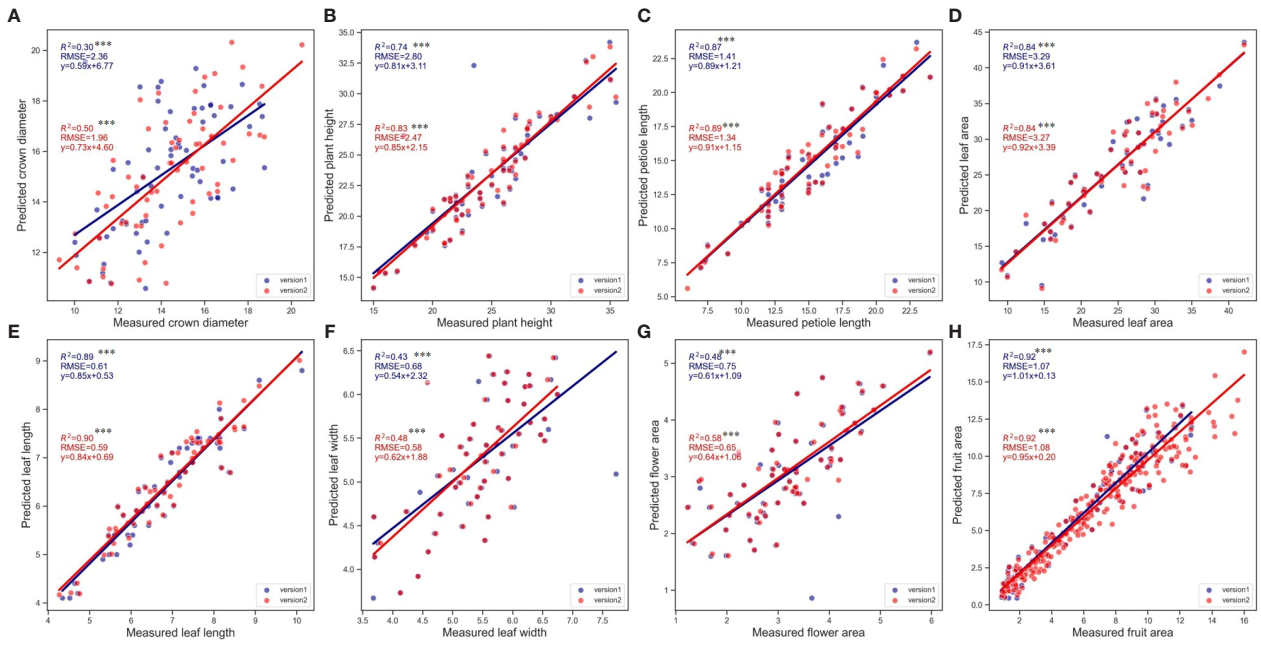


FIGURE 5
Correlation between real and predicted values (using V1 and V2 of SPT) for the six phenotypic traits. **(A)** Crown diameter, **(B)** Plant height, **(C)** Petiole length, **(D)** Leaf area, **(E)** Leaf length, **(F)** Leaf width, **(G)** Flower area, **(H)** Fruit area. Except for fruit where n=312. For the rest of the variables, n=70. Asterisks denote statistical significance at *p=0.05* (*** *p<0.001*).

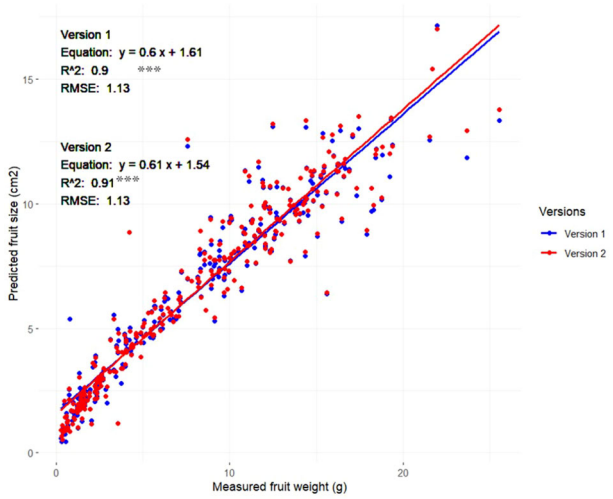


FIGURE 6
Relationship between strawberry fruit size (area) measured using SPT (Versions 1 and 2) and fruit weight (n=310) Asterisks denote statistical significance at *p=0.05* (*** *p<0.001*).

phenotypic changes provided by SPT, crucial for effective crop management.

Subsequently, we conducted an analysis using Analysis of Variance (ANOVA) and Tukey's Honestly Significant Difference (HSD) test to explore the impact of early phenotypic size variation on strawberry fruit yield. For example, data segmentation based on crown diameter (CD) at the transplantation stage, categorized as large, medium, and small, showed that early crown sizes were statistically significant predictors of yield outcomes (Figure 8A). Additionally, when samples were categorized based on other important phenotypic parameters such as flower area, unripe fruits, and leaf area into different size-based clusters at specified times before fruit harvesting, a notable trend emerged: an increase in the size of these phenotypes was associated with a substantial increase in yield (p<0.05) (Figures 8B–D).

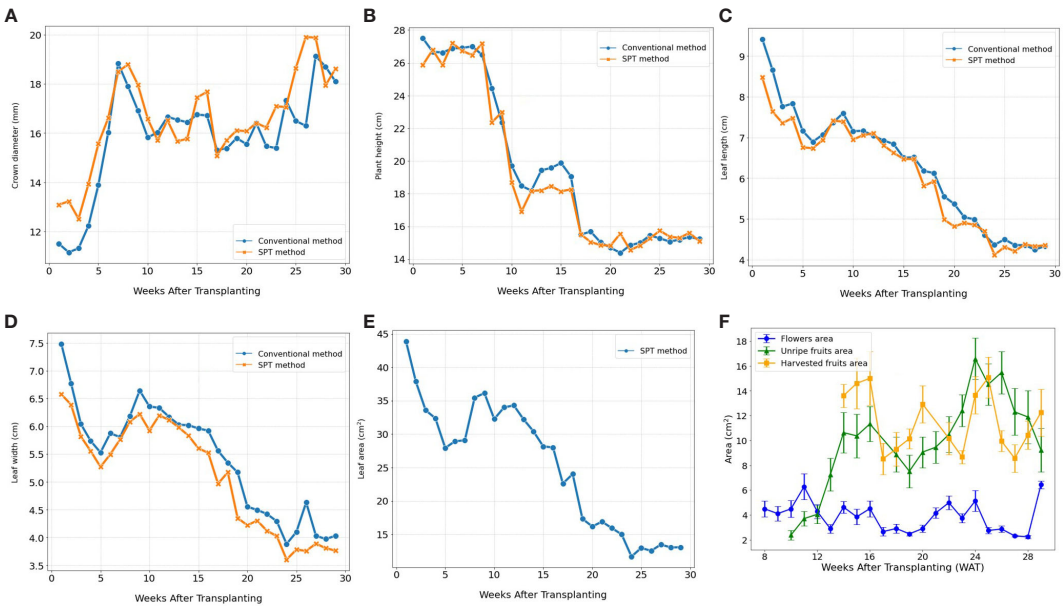


FIGURE 7
Illustrates the temporal dynamics of various phenotypic traits related to growth and yield in strawberry plants over the growing season. The phenotypic data plotted include, **(A)** crown diameter, **(B)** plant height, **(C)** leaf length, and **(D)** leaf width collected by both conventional and SPT methods. Additionally, **(E, F)** depict the evolution of leaf area, flower size, the area of unripe (non-harvestable) and ripe (harvestable) fruits using data exclusively collected by SPT. (n=27). Error bars represent the standard error of the mean.

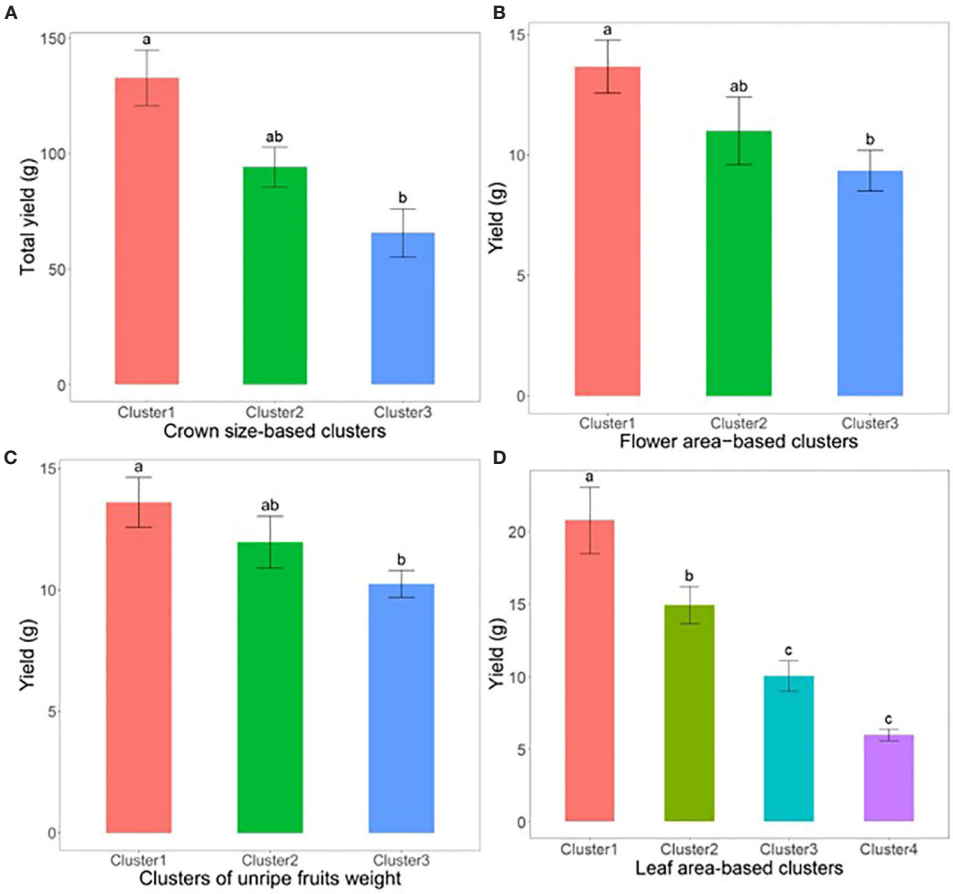


FIGURE 8
Illustrates the yield of strawberry fruits derived from plant samples sorted into phenotypic clusters based on size at critical growth stages. **(A)** showcases total yield per plant among crown size-based clusters, measured 2 weeks post-transplanting (n=22). **(B)** displays weekly yield per plant across flower size-based clusters, taken 5 weeks prior to harvest (n=125). **(C)** reveals weekly yield per plant for clusters categorized by the weight of unripe fruits, recorded 2 weeks before harvest (n=198). Lastly, **(D)** illustrates weekly yield per plant for leaf area-based clusters, assessed 4 weeks pre-harvest (n=80). The error bars represent the standard error, and groups labeled with the same letters within the same subfigure are not statistically different according to Tukey's Honestly Significant Difference test, $p<0.05$.

3.5 Data analysis and visualization using SPT

In this section, we describe the details related to the functionality of the SPT focusing on image data analysis and results visualization. Our phenotyping tool is used to analyze strawberry images and provides quantitative data in two alternative ways, and both require registration of the user as well the plant samples before usage. In addition, both methods require internet access. The one is a direct real-time method and operates on mobile phones. To acquire phenotypic data using this method, the user needs a smartphone and a QR code. Then chooses the plant ID and take the image of the target phenotype so that both the object and the QR code appear in the same image and confirm the task by clicking ok. The detected and analyzed phenotype is displayed along with the results within a short time (Figure 9). The second method is an indirect and works on both smartphone and computer. To analyze an image using this method, you choose the plant ID and the corresponding image to be analyzed is uploaded to the system from the computer or smartphone storage. The image should contain the target phenotype and the QR code. Once the image enters the system, the target objects are immediately detected and analyzed, and the results are displayed, in a manner similar to that of direct methods. The results can be downloaded as CSV files for further processing.

4 Discussion

In this study, we propose an image-based digital phenotyping tool, utilizing deep learning (DL), for analyzing strawberry phenotypes focusing on six essential parameters for strawberry growth and yield analysis. These images can be acquired using modern smartphones and analyzed directly at the field level or stored for subsequent analysis. Several machine-learning-based solutions for strawberry plant phenotyping to acquire growth and yield information have been proposed, most of which have concentrated on fruit detection, classification, segmentation (Chuah, 2018; Kirk et al., 2020; Basak et al., 2022; Wu et al., 2022), fruit size (Lee et al., 2017; Yue et al., 2020), and leaf area

FIGURE 9
Example of data analysis and visualization (screenshots) for the six strawberry phenotypic traits (**A**. Crown diameter, **B**. Plant height, **C**. Petiole length, **D**. Leaf size, **E**. Flower size, **F**. Fruit size) using SPT. The left images (under input column) are the raw input images, and the right images (under output column) show the output of the analyzed images, with quantitative results at the bottom.

(de Castro et al., 2020). However, a comprehensive digital phenotyping tool to measure most agronomic traits of strawberry growth, such as CD, PH, and PL, remains unreported in the existing literatures. Therefore, the efficiency of this tool in the non-destructive extraction of strawberry growth and yield phenotypic data, including those that were not attempted digitally, makes it a promising tool that can assist farmers and ordinary researchers in proper decision-making. The core robustness of our system was rendered by integrating two DL architectures, YOLOv4 and U-net, which are reportedly efficient regarding high-precision and real-time object detection (Bochkovskiy et al., 2020) and image segmentation (Ronneberger et al., 2015), respectively. The combination of YOLO and U-net series together or with other architectures to build a more robust phenotyping system or for comparison purposes has also been successfully performed in other crops such as grape (Barbole and Jadhav, 2021), wheat (Ullah et al., 2021), and mango (Koirala et al., 2019).

During the development, the real data values of the six phenotypes were acquired using conventional approaches, and the corresponding images were maintained. The latter served as the benchmark dataset for evaluating the performance of our digital system throughout the development process. Upon completion of V1, most of the performance metrics assessed, such as the mean precision average, detection frequency and Pearson correlation coefficients, were relatively poor in V1. However, by increasing the number of training datasets and changing the annotation method from single-target no-labeling annotation to multi-target labeling annotation, the abovementioned metrics improved for most of the various phenotypic parameters studied, making up V2. Such improvements in the performance and robustness of our DL framework in V2 can be attributed to the consistency and multi-target labeling annotations. During data collection, several objects, including nontarget objects, may be included in the image. In such cases, if the nontarget object is similar to the target, it confuses the annotator. Therefore, the annotation criteria should be pre-defined and consistently maintained throughout the annotation process.

For example, if the target strawberry fruit and the fruit next (not the target) to it are of the same size or distance (distance from the camera), both fruits need to be annotated; however, if it is blurred and not clearly visible or it is distant from the camera, only the target fruit should be annotated. By creating such consistent criteria to avoid annotating far or invisible targets, it is possible to prevent the model from being incorrectly trained on objects that look almost similar when learning. In addition, in a single image, the co-occurrence of more than one but different objects is normal. For example, while capturing a fruit cluster, flowers may be captured together, or petioles may be included while capturing plant height.

By labeling and annotating all the available objects in the image (those investigated) and training the AI accordingly, the overall performance and robustness of the system were increased. Low-quality annotation severely affects training models (Guo et al., 2020). According to Mullen et al. (2019), the annotation technique impacts the deep neural network, and the inconsistency of the annotation technique may cause incorrect conclusions regarding model performance. Recently, Yun et al. (2021) applied multiple labeling annotations on ImageNet (Russakovsky et al.,

2015), and without modifying the models, they improved the classification accuracy solely by revising the models from single-object labeling annotations to multi-labeling, which is consistent with our attempt to improve our phenotyping tool from V1 to V2. These findings are consistent with those reported in other studies (Barbedo, 2018; Shorten and Khoshgoftaar, 2019; Alhazmi et al., 2021) that dataset size significantly affects the performance of DL models. Furthermore, the strong correlation between strawberry fruit weight and fruit area (R^2>0.9) confirmed the efficiency of our system for predicting strawberry fruit yield based on the measured fruit area. Such options will help users forecast yields non-destructively and accurately, which, in the case of professional farmers, can assist in planning harvesting and marketing activities. This model can also be applied to strawberry studies.

In this study, we evaluated the efficiency of the SPT in collecting strawberry growth and yield data under real field conditions. The results revealed that SPT was comparable to conventional approaches in collecting growth information and could be used to monitor phenotypic traits, such as leaf area, flower area, and fruit area, which were previously challenging to obtain using traditional methods. A relatively large increase in the size of the aforementioned phenotypes before harvesting was associated with a significant increase in the yield. These results underline the potential of the SPT as a valuable tool for farmers engaged in professional and smart farming and its significance in yield prediction. Accurate yield prediction is crucial for farmers, as Kerfs et al. (2017) reported that weekly strawberry yields can vary significantly and emphasized that farmers should regularly monitor their fields for smooth planning of farm operations, particularly postharvest activities, for adequate resource distribution. These findings are consistent with those of previous studies. For example, Abd-Elrahman et al. (2021) investigated the integration of ground-based canopy images into modeling approaches to improve strawberry yield. Using canopy images captured with a handheld digital camera and machine-learning algorithms caused a significant improvement in yield prediction accuracy compared with traditional approaches. Yoon et al. (2023) combined immature fruit information with AI techniques to develop a strawberry yield prediction model. They also concluded that DL-based strawberry fruit detection results could contribute to yield prediction. In a previous study (Hassan et al., 2018), hyperspectral remote sensing imagery was used to acquire leaf area index parameters and six other vegetation indices to investigate the relationship between these parameters and yield under different growing conditions, which revealed that the leaf area index was highly related to yield. These findings provide valuable insights into the relationship between leaf indices and yield and contribute to a better understanding of the factors influencing fruit productivity. Finally, using the data collected through the SPT, we investigated the impact of initial crown size on strawberry yield. Our analysis involved comparing the total yields obtained from different crown classes, including small, medium, and large crowns. Our results indicated that strawberries with a larger initial CD produced significantly higher yields than those with smaller crowns, confirming the findings of previous studies (Torres-Quezada et al., 2015; Fridiaa et al., 2016; Fagherazzi et al., 2021). These studies also reported similar results, suggesting that the initial crown size is a factor that significantly affects strawberry yield. Therefore, if SPT is integrated into strawberry farming, it will possibly alleviate various farm management-related challenges and boost farm production.

Although the increase of the dataset and enhancement of annotation techniques significantly improved the core models (YOLOv4 and U-net) of our Strawberry Phenotyping Tool (SPT) in terms of precise detection, measurement, and analysis of strawberry phenotypic features, we encountered several challenges that require improvements in future studies to maximize the full potential of the SPT.

The two core models of our Strawberry Phenotyping Tool (SPT), YOLOv4 and U-net, require a relatively large volume of data, which necessitates extensive labor to annotate such a dataset. Additionally, managing this large volume of image data and the related logistics remains a challenge due to significant computational resource requirements.

The current version of the SPT was unable to accurately detect petiole length under field conditions, resulting in the omission of these results from our report. Additionally, detecting and measuring relatively small features poses a significant challenge, often requiring multiple captures of the same object from different angles to ensure target feature detection and accuracy. This issue is more aggravated when the plant canopy becomes bushy in later growth stages, especially if excessive leaves and side crowns are not pruned and managed properly. If the plants overgrow too much, their size may also exceed the capacity of a single person to acquire the images effectively, especially for plant height Enhancing the SPT's performance to provide more precise analysis of very small objects and operates well even in complex environments and growth conditions is a crucial area for future improvement.

Additionally, the SPT does not currently support the automatic categorization of fruits into different developmental stages, such as distinguishing between ripe and unripe fruits. Addressing this limitation could significantly enhance the tool's utility and applicability.

Lastly, flower initiation is a critically important event for the successful cultivation and production of june-bearing strawberry plants, directly impacting fruit yield (Van Delm et al., 2014; Li et al., 2021). This transient phenomenon is challenging to predict and is traditionally confirmed through destructive sampling and microscopic observation. By training our SPT models on potential phenotypic traits speculated to be indicators of flower initiation at the nursery stage, and supplementing this with robust statistical analysis, we anticipate that the SPT will provide valuable insights into the key phenotypic traits indicative of vegetative-to-reproductive changes.

We acknowledge that YOLOv8 has been released since the completion of our study. YOLOv8 has been reported to achieve high accuracy and fast inference speed (Liu et al., 2024). However, we chose to use YOLOv4 in our work due to its well-established performance and balance for server compatibility in our application. YOLOv4 had been extensively tested and was well-documented for deployment on servers. This was crucial for our study as we had already established a YOLOv4-based server infrastructure for our smartphone-based web application.

While YOLOv8 may offer potential performance gains, YOLOv4 was a suitable choice considering these factors. We provide the complete raw datasets (images subjected to YOLOv4),

along with the results of annotation and deep learning at https://github.com/kist-smartfarm/SPT, so that we and other research groups can use these datasets with the advanced deep learning architectures to develop more advanced practical phenotyping methods in future study.

Conclusively, in this study, a DL-based phenotyping tool was developed to collect, process, and analyze image-based strawberry phenotypes of six essential agronomic traits (CD, PL, PH, flower, leaf, and fruit size). The proposed approach involves integrating the YOLOv4 and U-net architectures into one system to make it more robust for better feature detection and extraction. An increased dataset size with various backgrounds, coupled with multi-labeling object annotation, improved the efficiency of our system in measuring target phenotypic traits with greater precision and accuracy. The evaluation of our phenotyping tool under real field settings showed the same efficiency in collecting strawberry growth data as conventional approaches, with additional capacity for predicting yield based on leaf, flower, and fruit indices. Real-time strawberry phenotyping with the current digitalized solution has potential applications in strawberry smart farming, assisting researchers and farmers in making appropriately informed decisions. In future studies, more phenotypic features, such as strawberry fruit maturity stage and canopy area quantification, should be added to the system to enable more in-depth strawberry phenotyping.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

Author contributions

JN: Data curation, Formal analysis, Validation, Visualization, Writing – original draft, Writing – review & editing. UL: Conceptualization, Formal analysis, Methodology, Software, Validation, Writing – review & editing. JY: Data curation, Investigation, Visualization, Writing – review & editing. SY: Data curation, Formal analysis, Writing – review & editing. SP: Formal analysis, Methodology, Software, Writing – original draft. TL: Formal analysis, Software, Supervision, Visualization, Writing – original draft. YY: Conceptualization, Investigation, Methodology, Supervision, Writing – review & editing. HK: Conceptualization, Funding acquisition, Investigation, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry (IPET) and Korea Smart Farm R&D (KosFarm) through the Smart Farm Innovation Technology Development Program. It was also funded by the Ministry of Agriculture, Food, and Rural Affairs (MAFRA), the Ministry of Science and ICT (MSIT), and the Rural Development Administration (RDA) (421002-04)

Acknowledgments

The authors thank Chang Geun Kim and Sang Mu Huh from Cultilabs company for their assistance in preparing and providing the test version of the SPT software with a user interface.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2024.1418383/full#supplementary-material

References

Abd-Elrahman, A., Wu, F., Agehara, S., and Britt, K. (2021). Improving strawberry yield prediction by integrating ground-based canopy images in modeling approaches. *ISPRS Int. J. Geo-Information* 10, 239. doi: 10.3390/ijgi10040239

Ahn, M. G., Kim, D. S., Ahn, S. R., Sim, H. S., Kim, S., and Kim, S. K. (2021). Characteristics and trends of strawberry cultivars throughout the cultivation season in a greenhouse. *Horticulturae* 7, 1–11. doi: 10.3390/horticulturae7020030

Aish, R. A. M. (2021). Strawberry classification using deep learning. *Int. J. Acad. Inf. Syst. Res.* 5, 6–13.

Alhazmi, K., Alsumari, W., Seppo, I., Podkuiko, L., and Simon, M. (2021). "Effects of annotation quality on model performance," in *3rd International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2021* . (Jeju Island, Korea (South): IEEE), 63–67. doi: 10.1109/ICAIIC51459.2021.9415271

Barbedo, J. G. A. (2018). Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. *Comput. Electron. Agric.* 153, 46–53. doi: 10.1016/j.compag.2018.08.013

Barbole, D., and Jadhav, P. (2021). Comparative analysis of deep learning architectures for grape cluster instance segmentation. *Inf. Technol. Ind.* 9, 344–352. doi: 10.17762/itii.v9i1.138

Basak, J. K., Paudel, B., Kim, N. E., Deb, N. C., Kaushalya Madhavi, B. G., and Kim, H. T. (2022). Non-destructive estimation of fruit weight of strawberry using machine learning models. *Agronomy* 12, 2487. doi: 10.3390/agronomy12102487

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). *YOLOv4: Optimal speed and accuracy of object detection*. Available online at: http://arxiv.org/abs/2004.10934.

Chen, Y., Lee, W. S., Gan, H., Peres, N., Fraisse, C., Zhang, Y., et al. (2019). Strawberry yield prediction based on a deep neural network using high-resolution aerial orthoimages. *Remote Sens.* 11, 1–21. doi: 10.3390/rs11131584

Chuah, N. L. M. C. (2018). "A strawberry detection system using convolutional neural networks," in *2018 IEEE International Conference on Big Data (Big Data)*. (Seattle, WA, USA: IEEE), 2515–2520. doi: 10.1016/B978-0-12-816176-0.00025-9

Costa, C., Schurr, U., Loreto, F., Menesatti, P., and Carpentier, S. (2019). Plant phenotyping research trends, a science mapping approach. *Front. Plant Sci.* 9, 1–11. doi: 10.3389/fpls.2018.01933

Czedik-Eysenberg, A., Seitner, S., Güldener, U., Koemeda, S., Jez, J., Colombini, M., et al. (2018). The 'PhenoBox', a flexible, automated, open-source plant phenotyping solution. *New Phytol.* 219, 808–823. doi: 10.1111/nph.15129

de Castro, A., Madalozzo, G. A., dos Santos Trentin, N., Castoldi da Costa, R., Calvete, E. O., Schardong Spalding, L. E., et al. (2020). BerryIP embedded: An embedded vision system for strawberry crop. *Comput. Electron. Agric.* 173, 105354. doi: 10.1016/j.compag.2020.105354

Dutta, A., and Zisserman, A. (2019). "The VIA annotation software for images, audio and video," in *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia*. (New York, NY, USA: Association for Computing Machinery), 2276–2279. doi: 10.1145/3343031.3350535

Fagherazzi, A. F., Suek Zanin, D., Soares Dos Santos, M. F., Martins de Lima, J., Welter, P. D., Francis Richter, A., et al. (2021). Initial crown diameter influences on the fruit yield and quality of strawberry pircinque. *Agronomy* 11, 1–16. doi: 10.3390/agronomy11010184

Fatehi, F., and Akhijahani, H. S. (2021). Classification of Parus strawberry fruit by combining image processing techniques and intelligent methods. *Iran. Food Sci. Technol. Res. J.* 16, 87–99. doi: 10.22067/ifstrj.v16i6.84187

Feldmann, M. J., Hardigan, M. A., Famula, R. A., López, C. M., Tabb, A., Cole, G. S., et al. (2020). Multi-dimensional machine learning approaches for fruit shape phenotyping in strawberry. *Gigascience* 9, giaa030. doi: 10.1093/gigascience/giaa030

Fiorani, F., and Schurr, U. (2013). Future scenarios for plant phenotyping. *Annu. Rev. Plant Biol.* 64, 267–291. doi: 10.1146/annurev-arplant-050312-120137

Fridiaa, A., Winardiantika, V., Lee, Y. H., Choi, I. Y., Yoon, C. S., and Yeoung, Y. R. (2016). Influence of crown size on plant growth, flowering and yield of day-neutral strawberry cultivars. *Acta Hortic.* 1117, 347–353. doi: 10.17660/ActaHortic.2016.1117.57

Gan, H., Lee, W. S., Peres, N., and Fraisse, C. (2019). Development of A multi-angle imaging system for automatic straw-berry flower counting. in *International Symposium on Artificial Intelligence and Mathematics* (Fort Lauderdale, Florida, USA: AAAI).

Guo, L., Xu, X., Xie, G., and Gao, J. (2020). Conceptual cognitive modeling for fine-grained annotation quality assessment of object detection datasets. *Discret. Dyn. Nat. Soc* 2020, 1–11. doi: 10.1155/2020/6195189

Hassan, A. H., Taha, S. S., and Aboelghar, M. (2018). Comparative the impact of organic and conventional strawberry cultivation on growth and productivity using remote sensing techniques under Egypt climate conditions. *Asian J. Agric. Biol.* 6, 228–244.

He, J. Q., Harrison, R. J., and Li, B. (2017). A novel 3D imaging system for strawberry phenotyping. *Plant Methods* 13, 93. doi: 10.1186/s13007-017-0243-x

Hortyński, J. A., Zebrowska, J., Gawroński, J., and Hulewicz, T. (1991). Factors influencing fruit size in the strawberry (Fragaria ananassa Duch.). *Euphytica* 56, 67–74. doi: 10.1007/BF00041745

Hwang, H. S., Jeong, H. W., Lee, H. R., Jo, H. G., Kim, H. M., and Hwang, S. J. (2020). Acceleration of flower bud differentiation of runner plants in "Maehyang" strawberries using nutrient solution resupply during the nursery period. *Agronomy* 10, 1127. doi: 10.3390/agronomy10081127

Ilyas, T., Khan, A., Umraiz, M., Jeong, Y., and Kim, H. (2021). Multi-scale context aggregation for strawberry fruit recognition and disease phenotyping. *IEEE Access* 9, 124491–124504. doi: 10.1109/ACCESS.2021.3110978

Jo, J. S., Sim, H. S., Jung, S., Moon, Y. H., Jo, W. J., Woo, U. J., et al. (2022). Estimation and validation of the leaf areas of five june-bearing strawberry (Fragaria × ananassa) cultivars using non-destructive methods. *J. Bio-Environment Control* 31, 98–103. doi: 10.12791/KSBEC.2022.31.2.098

Kerfs, J., Eagan, Z., and Liu, B. (2017). "Machine vision for strawberry detection," in *2017 ASABE Annual International Meeting*. (Washington, DC, USA: American Society of Agricultural and Biological Engineers), 1. doi: 10.13031/aim.201700925

Khammayom, N., Maruyama, N., Chaichana, C., and Hirota, M. (2022).). Impact of environmental factors on energy balance of greenhouse for strawberry cultivation. *Case Stud. Therm. Eng.* 33, 101945. doi: 10.1016/j.csite.2022.101945

Kirk, R., Cielniak, G., and Mangan, M. (2020). L*a*b*Fruits: A rapid and robust outdoor fruit detection system combining bio-inspired features with one-stage deep learning networks. *Sensors (Basel).* 20, 275. doi: 10.3390/s20010275

Koirala, A., Walsh, K. B., Wang, Z., and McCarthy, C. (2019). Deep learning for real-time fruit detection and orchard fruit load estimation: benchmarking of 'MangoYOLO.'. *Precis. Agric.* 20, 1107–1135. doi: 10.1007/s11119-019-09642-0

Lee, D. H., Cho, Y., and Choi, J. M. (2017). Strawberry volume estimation using smartphone image processing. *Korean J. Hortic. Sci. Technol.* 35, 707–716. doi: 10.12972/kjhst.20170075

Li, Y., Xiao, J., Guo, G., and Jeong, B. R. (2021). Transplant pre-chilling induces earlier flowering and fruiting for forcing-cultured June-bearing strawberries. *Scientia Horticulturae.* 288, 110371. doi: 10.1016/j.scienta.2021.110371

Liu, Z., Abeyrathna, R. D., Sampurno, R. M., Nakaguchi, V. M., and Ahamed, T. (2024). Faster-YOLO-AP: A lightweight apple detection algorithm based on improved YOLOv8 with a new efficient PDWConv in orchard. *Comput. Electron. Agriculture.* 223, 109118. doi: 10.1016/j.compag.2024.109118

Mahmud Sultan, M., Zaman, Q. U., Esau, T. J., Chang, Y. K., Price, G. W., and Prithiviraj, B. (2020). Real-time detection of strawberry powdery mildew disease using a mobile machine vision system. *Agronomy* 10, 1027. doi: 10.3390/agronomy10071027

Mbarushimana, J. C., Bosch, D. J., and Samtani, J. B. (2022). An economic comparison of high tunnel and open-field strawberry production in southeastern Virginia. *Horticulturae* 8, 1139. doi: 10.3390/horticulturae8121139

Menzel, C. M. (2020). A review of productivity in strawberries: marketable yield has a linear, but inconsistent relationship with total yield, and cannot be predicted from total yield. *J. Hortic. Sci. Biotechnol.* 96, 135–144. doi: 10.1080/14620316.2020.1808086

Mullen, J. F., Tanner, F. R., and Sallee, P. A. (2019). "Comparing the effects of annotation type on machine learning detection performance," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (Long Beach, CA, USA: IEEE), 855–861. doi: 10.1109/CVPRW47913.2019

Ni, X., Li, C., Jiang, H., and Takeda, F. (2020). Deep learning image segmentation and extraction of blueberry fruit traits associated with harvestability and yield. *Hortic. Res.* 7, 110. doi: 10.1038/s41438-020-0323-3

Pérez-Borrero, I., Marín-Santos, D., Gegúndez-Arias, M. E., and Cortés-Ancos, E. (2020). A fast and accurate deep learning method for strawberry instance segmentation. *Comput. Electron. Agric.* 178, 105736. doi: 10.1016/j.compag.2020.105736

Perez-Borrero, I., Marin-Santos, D., Vasallo-Vazquez, M. J., and Gegundez-Arias, M. E. (2021). A new deep-learning strawberry instance segmentation methodology based on a fully convolutional neural network. *Neural Comput. Appl.* 33, 15059–15071. doi: 10.1007/s00521-021-06131-2

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. (Las Vegas, NV, USA: IEEE), 779–788. doi: 10.1109/CVPR.2016.91

Redmon, J., and Farhadi, A. (2017). "YOLO9000: Better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. (Honolulu, HI, USA: IEEE), 7263–7271. doi: 10.1109/CVPR.2017.690

Redmon, J., and Farhadi, A. (2018). *YOLOv3: an incremental improvement*. Available online at: http://arxiv.org/abs/1804.02767.

Robert, F., Pétel, G., Risser, G., and Gendraud, M. (1997). Determination of the growth potential of strawberry plants (Fragaria x ananassa Duch.) by morphological and nucleotide measurements in relation to chilling. *Can. J. Plant Sci.* 77, 127–132. doi: 10.4141/P96-002

Robert, F., Risser, G., and Pétel, G. (1999).). Photoperiod and temperature effect on growth of strawberry plant (Fragaria x ananassa Duch.): Development of a morphological test to assess the dormancy induction. *Sci. Hortic. (Amsterdam).* 82, 217–226. doi: 10.1016/S0304-4238(99)00054-0

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. ed. A.F.F.N. Navab, J Hornegger and W.M. Wells (Berlin, Germany: Springer: Cham), 234–241.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Shin, J., Chang, Y. K., Heung, B., Nguyen-Quang, T., Price, G. W., and Al-Mallahi, A. (2021). A deep learning approach for RGB image-based powdery mildew disease detection on strawberry leaves. *Comput. Electron. Agric.* 183, 106042. doi: 10.1016/j.compag.2021.106042

Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6, 1–48. doi: 10.1186/s40537-019-0197-0

Simpson, D. (2018). "The Genomes of Rosaceous Berries and Their Wild Relatives," in *The Genomes of Rosaceous Berries and Their Wild Relatives*, eds. T. Hytönen, J. Graham and R. Harrison (Cham, Switzerland: Springer International Publishing), 1–7. doi: 10.1007/978-3-319-76020-9

Sønsteby, A., and Heide, O. M. (2006). Dormancy relations and flowering of the strawberry cultivars Korona and Elsanta as influenced by photoperiod and temperature. *Sci. Hortic. (Amsterdam).* 110, 57–67. doi: 10.1016/j.scienta.2006.06.012

Takahashi, M., Takayama, S., Umeda, H., Yoshida, C., Koike, O., Iwasaki, Y., et al. (2020). Quantification of strawberry plant growth and amount of light received using a depth sensor. *Environ. Control Biol.* 58, 31–36. doi: 10.2525/ecb.58.31

Teoh, M. K., Teo, K. T. K., and Yoong, H. P. (2022). Numerical computation-based position estimation for QR code object marker: mathematical model and simulation. *Computation* 10, 147. doi: 10.3390/computation10090147

Torres-Quezada, E. A., Zotarelli, L., Whitaker, V. M., Santos, B. M., and Hernandez-Ochoa, I. (2015). Initial crown diameter of strawberry bare-root transplants affects early and total fruit yield. *Horttechnology* 25, 203–208. doi: 10.21273/HORTTECH.25.2.203

Tsaftaris, S. A., Minervini, M., and Scharr, H. (2016). Machine learning for plant phenotyping needs image processing. *Trends Plant Sci.* 21, 989–991. doi: 10.1016/j.tplants.2016.10.002

Ullah, S., Henke, M., Narisetti, N., Hejatko, J., and Gladilin, E. (2021). Automated detection and segmentation of grain spikes in greenhouse images using shallow and deep learning neural networks: A comparison of six methods. *Res. Square*. doi: 10.21203/rs.3.rs-252740/v1

Van Delm, T., Melis, P., Stoffels, K., Van De Vyver, F., and Baets, W. (2014). Strawberry plant architecture and flower induction in plant production and strawberry cultivation. *Acta Hortic.* 1049, 489–494. doi: 10.17660/ActaHortic.2014.1049.72

Wu, H., Cheng, Y., Zeng, R., and Li, L. (2022). "Strawberry Image Segmentation Based on Û 2-Net and maturity calculation," in *2022 14th International Conference on Advanced Computational Intelligence, ICACI 2022*. (Wuhan, China: IEEE), 74–78. doi: 10.1109/ICACI55529.2022.9837483

Yoon, S., Jo, J. S., Kim, S. B., Sim, H. S., Kim, S. K., and Kim, D. S. (2023). Prediction of strawberry yield based on receptacle detection and Bayesian inference. *Heliyon.* 9, e14546. doi: 10.1016/j.heliyon.2023.e14546

Yue, X. Q., Shang, Z. Y., Yang, J. Y., Huang, L., and Wang, Y. Q. (2020). A smart data-driven rapid method to recognize the strawberry maturity. *Inf. Process. Agric.* 7, 575–584. doi: 10.1016/j.inpa.2019.10.005

Yun, S., Oh, S. J., Heo, B., Han, D., Choe, J., and Chun, S. (2021). "Re-labeling ImageNet: From single to multi-labels, from global to localized labels. Proc," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. (Nashville, TN, USA: IEEE), 2340–2350. doi: 10.1109/CVPR46437.2021.00237

Zhang, J., He, L., Karkee, M., Zhang, Q., Zhang, X., and Gao, Z. (2018). Branch detection for apple trees trained in fruiting wall architecture using depth features and Regions-Convolutional Neural Network (R-CNN). *Comput. Electron. Agric.* 155, 386–393. doi: 10.1016/j.compag.2018.10.029

Zheng, C., Abd-elrahman, A., and Whitaker, V. (2021). Remote sensing and machine learning in crop phenotyping and management, with an emphasis on applications in strawberry farming. *Remote Sens.* 13, 1–29. doi: 10.3390/rs13030531

Zheng, Y. Y., Kong, J. L., Jin, X. B., Wang, X. Y., and Zuo, M. (2019). CropDeep: The crop vision dataset for deep-learning-based classification and detection in precision agriculture. *Sensors (Basel).* 19, 1058. doi: 10.3390/s19051058

Zhou, C., Hu, J., Xu, Z., Yue, J., Ye, H., and Yang, G. (2020). A novel greenhouse-based system for the detection and plumpness assessment of strawberry using an improved deep learning technique. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00559

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 25 July 2024
DOI 10.3389/fpls.2024.1356078

OPEN ACCESS

EDITED BY
Roger Deal,
Emory University, United States

REVIEWED BY
Eetu Puttonen,
National Land Survey of Finland, Finland
Bingxiao Wu,
Guangdong Academy of Agricultural Sciences (GDAAS), China

*CORRESPONDENCE
Lin Cao
lincao@njfu.edu.cn

†These authors have contributed equally to this work and share first authorship

RECEIVED 15 December 2023
ACCEPTED 04 July 2024
PUBLISHED 25 July 2024

CITATION
Liang Y, Zhou K and Cao L (2024) An advanced three-dimensional phenotypic measurement approach for extracting *Ginkgo* root structural parameters based on terrestrial laser scanning.
Front. Plant Sci. 15:1356078.
doi: 10.3389/fpls.2024.1356078

An advanced three-dimensional phenotypic measurement approach for extracting *Ginkgo* root structural parameters based on terrestrial laser scanning

Yinyin Liang†, Kai Zhou† and Lin Cao*

Co-Innovation Center for Sustainable Forestry in Southern China, Nanjing Forestry University, Nanjing, China

The phenotyping of plant roots is essential for improving plant productivity and adaptation. However, traditional techniques for assembling root phenotyping information are limited and often labor-intensive, especially for woody plants. In this study, an advanced approach called accurate and detailed quantitative structure model-based (AdQSM-based) root phenotypic measurement (ARPM) was developed to automatically extract phenotypes from *Ginkgo* tree root systems. The approach involves three-dimensional (3D) reconstruction of the point cloud obtained from terrestrial laser scanning (TLS) to extract key phenotypic parameters, including root diameter (RD), length, surface area, and volume. To evaluate the proposed method, two approaches [minimum spanning tree (MST)-based and triangulated irregular network (TIN)-based] were used to reconstruct the *Ginkgo* root systems from point clouds, and the number of lateral roots along with RD were extracted and compared with traditional methods. The results indicated that the RD extracted directly from point clouds [coefficient of determination (R^2) = 0.99, root-mean-square error (RMSE) = 0.41 cm] outperformed the results of 3D models (MST-based: R^2 = 0.71, RMSE = 2.20 cm; TIN-based: R^2 = 0.54, RMSE = 2.80 cm). Additionally, the MST-based model (F1 = 0.81) outperformed the TIN-based model (F1 = 0.80) in detecting the number of first-order and second-order lateral roots. Each phenotyping trait fluctuated with a different cloud parameter (CP), and the CP value of 0.002 (r = 0.94, p < 0.01) was found to be advantageous for better extraction of structural phenotypes. This study has helped with the extraction and quantitative analysis of root phenotypes and enhanced our understanding of the relationship between architectural parameters and corresponding physiological functions of tree roots.

KEYWORDS

root phenotyping, LiDAR, 3D reconstruction, *Ginkgo*, root structural parameters

1 Introduction

Roots play a key role in supporting trees and the global carbon cycle, and they can also regulate ecosystem processes via plant–soil–microbe interactions by driving plants to obtain water and nutrients (Lynch, 1995; Rajendra et al., 2014; Villordon et al., 2014; Freschet et al., 2021b). Root system architecture (RSA) has become the "second green revolution" for global food security (Lynch, 2007). Structural and morphological characteristics, such as root diameter (RD), number, and lateral root isometry, are vital to understanding plant physiological functions (Gu et al., 2014; Wang, 2017; Seethepalli et al., 2021). Root phenotyping can track these structural and morphological characteristics and, thus, has great potential for bioenergy agroecosystems (York et al., 2022). Because of the challenge of directly gathering information on the roots underground, there are currently limited studies focusing on root phenotyping (Wilhelm et al., 2022).

Since the 1990s, innovative techniques and devices have been applied to root measurement, including non-destructive, manual, or automatic two-dimensional (2D) and three-dimensional (3D) digitizing techniques (Danjon and Reubens, 2008). Current research on root phenotyping primarily focuses on the cultivation of crops (e.g., rice, corn, and wheat), with higher yields, higher quality, and more excellent resistance to stress in the combination with genomic data (de Dorlodot et al., 2007; Kuijken et al., 2015; Topp et al., 2016; Tracy et al., 2020). The most extensively used method for root phenotyping is based on 2D images (Chen et al., 2018). However, 2D measurements are limited by the fact that pictures are typically taken from just one or two perspectives, where information can be lost as a result of roots overlapping (Shao et al., 2021). Traditional techniques for assembling root phenotyping information include minirhizotron techniques (Volkmar, 1993) and the agar gel culture method (Iyer-Pascuzzi et al., 2010). These methods are not only time-consuming and laborious, but also unable to describe the actual 3D structure of the root system (Delory et al., 2022). To further understand trait–function interactions, standardized and high-throughput approaches for acquiring root phenotypes are required (Wen et al., 2015; Delory et al., 2022). However, to the best of our knowledge, research on the 3D root structure of woody plants is still in its infancy (Li et al., 2015). In particular, there needs a systemic approach to accessing the architecture of tree roots (Zanetti et al., 2015). Although it is challenging to collect phenotypic information on the tree root system, evaluating the 3D root structure of woody perennials is crucial for understanding ecology (Liu, 1998), physiology and biochemistry (Steingrobe, 2001), morphology (Xi, 2019), biomechanics (Song et al., 2008), and bioenergy (York et al., 2022). More importantly, extracting tree root phenotyping traits is also critical in cultivating tree species with higher economic and ecological merits.

Structure from motion (SfM) has recently emerged as a digital tool for studying root structures (Koeser et al., 2016). This technique involves the acquisition of target point cloud data through photography, which is then used to perform 3D reconstruction (Lu et al., 2021). However, the performance of this method is influenced by the size of the object and the distance of measurement, and the process of dealing with background noise can be time-consuming (Okamoto et al., 2022). Computed tomography (CT) (Shao et al., 2021) and magnetic resonance imaging (MRI) (van Dusschoten et al., 2016) are currently popular techniques for the determination of 3D phenotypic information of roots, but less frequently to large woody root systems (Wu et al., 2021b). By digging up *Quercus petraea* and *Pinus pinaster*, Danjon et al. (1999) manually measured the diameter and topology of the root system and then reconstructed the 3D structure of the roots. However, manual measurements occupied an average of 2 to 3 h for each root. Danjon et al. (2005) adopted a 3D digitizer to measure the root structure of *P. pinaster*, reconstructing the 3D model of the roots and coloring the roots hierarchically, to link the structural properties of the roots with the stability of against wind. Yang (2021) used SketchUp software to simulate the 3D visualization of the root system of slope protection plants. The root configuration parameters, topological indexes, and fractal dimensions were extracted, which provided an important basis for the planting method and species selection of slope protection plants. However, owing to the time-consuming and laborious recording of coordinates, diameter, angle, and other parameters, the 3D structure of the root system cannot be directly obtained. Zhang et al. (2020) applied time-consuming 3D printing to simulate the 3D structure of roots with a physical model. This model utilized four fixed-sized RDs to represent the entire root system, which hardly capture the real RSA and morphology of roots. Spanos et al. (2008) obtained root structure information by uprooting *Abies cephalonica* Loudon with a 3D digitizer, which was limited to the lab analysis. Zhang et al. (2021) utilized ground-penetrating radar (GPR) to detect the roots of *Pinus sylvestris* var. *mongolica*, by connecting the root system's coordinate to determine its spatial distribution. Because of the influence of soil water content and resolution, as an emerging nondestructive detection technique, GPR cannot detect fine roots (RD less than 2 mm) and cannot directly obtain the 3D structure of roots. Quantitatively obtaining multidimensional information on plant roots, for constructing 3D models with a high efficiency, has become a challenging problem in root visualization research (Wu et al., 2021a).

Light detection and ranging (LiDAR) is a fast, non-destructive, and accurate remote sensing sensor for monitoring plant information (Cao et al., 2014; Lin, 2015; Zhou and Cao, 2021). Terrestrial laser scanning (TLS), a near-ground active remote sensing technique (Raumonen et al., 2013), can efficiently and accurately gather information about 3D point clouds of trees (Wang et al., 2021). It can also quantitatively extract the parameters and skeleton of trees for creating 3D models (Liu, 2016). Previous studies focused primarily on the aboveground components (e.g., branches, leaves, and trunks) (Liu et al., 2016; Disney, 2019), but only a few on roots underground with different typical root phenotyping acquisition methods and sensors (Table 1). Smith et al. (2014) and Todo et al. (2021) demonstrated that TLS point clouds are capable of accurately representing tree root architecture, thereby providing a robust technical foundation for 3D models to characterize root variables. *Ginkgo* (*Ginkgo biloba* L.) is a deep-rooted tree species (Fu and Zhang, 2019) and an essential economic tree species in China, with various valuable characteristics, namely, medicinal, edible, and ecological, and it can also be used in landscaping (Shen et al.,

TABLE 1 Comparison of root phenotyping acquisition methods with different sensors.

Data type	Sensor	Species	Advantages of techniques	Disadvantages of techniques	References
2-D	RGB camera	Crops, herbs	High throughput, low cost	Large amount of data, incomplete root system image information	(Yin et al., 2009); (Wilhelm et al., 2022)
	Electrical resistance tomography (ERT)	Trees	Non-destructive	Multiple sources of error, highly influenced by soil moisture	(Amato et al., 2008); (Zhao et al., 2019)
3-D	3-D digitizer	Trees, crops	High precision, semi-automatic	Time-consuming, complicated operation	(Danjon et al., 1999); (Spanos et al., 2008)
	X-ray computed tomography (CT)	Crops	Non-destructive, high precision	Costly, unable to detect coarse roots, indoor operation	(Perret et al., 2007); (Shao et al., 2021)
	Laser scanning (LiDAR)	Trees	Wide range of detection, high precision	Costly, uneven point cloud density	(Smith et al., 2014); (Todo et al., 2021)
	Magnetic resonance imaging (MRI)	Crops	Non-destructive, high precision	Costly, unable to detect coarse roots, indoor operation	(van Dusschoten et al., 2016)
	Ground-penetrating radar (GPR)	Trees	Non-destructive, *in situ* detection	Low resolution, unable to detect fine roots, highly influenced by soil and water	(Zhang et al., 2020); (Alani and Lantini, 2020)

2020). Research on *Ginkgo* now focuses on the aboveground part, and it is uncommon to find studies on its root systems (Men, 1986). Studies on the morphology and structure of the *Ginkgo* root system, as well as the establishment of 3D models, can contribute to understanding its physiological activities and mechanisms. Additionally, genetic data can also be combined with root phenotyping to create more tolerant and productive plants. A sound and organized system of plant research is made possible by quantitative descriptions of the root structural parameters.

However, to date, limited comparable investigations utilizing LiDAR have been carried out on the root systems of woody plants. It would be of high potential value for parameterizing 3D models of tree root systems to quantify the relationship between RD and other root phenotyping traits. Therefore, in this study, we utilized TLS to scan the *Ginkgo* root systems for obtaining its 3D structure, while establishing 3D models of the root system and automatically extracting several phenotyping traits. Specifically, the objectives of this study were (1) to develop an advanced approach of accurate and detailed quantitative structure model-based (AdQSM-based) root phenotypic measurement (ARPM) for extracting 3D phenotypic parameters of tree roots, (2) to evaluate the ability of the developed approach for extracting structural parameters of *Ginkgo* roots, and (3) to analyze the variations of the *Ginkgo* root structural parameters automatically extracted based on the developed approach by considering different parameters.

2 Materials and methods

Figure 1 shows the overall framework for measuring Ginkgo root phenotypes, which includes the operation of the developed ARPM approach (A), as well as the specific workflow of the study (B). The ARPM is divided into four modules: data acquisition, operator-assisted processing, 3D visualization and modeling, and phenotyping extraction. Specifically, LiDAR data were collected by setting up site scans, followed by preprocessing such as stitching, denoising, cropping, and coordinate conversion. Secondly, two approaches were further used to reconstruct the root: minimum spanning tree (MST)-based and triangulated irregular network (TIN)-based models. Accordingly, we employed various metrics to assess the models' performance. The coefficient of determination (R^2), root-mean-square error (RMSE), and mean absolute error (MAE), for example, are used to assess the RD; Recall, Precision, F1-score, and Accuracy are used to assess the number of roots. Thirdly, we used the AdQSM algorithm to automatically extract essential root traits and the Pearson correlation coefficient to assess the relationship between aspiration rate and RD.

2.1 Data acquisition

The information was gathered on 27 December 2021 at the Xiashu Experimental Forestry Site of Nanjing Forestry University, Jiangsu Province (119°22′E, 32°12′N). The climate of the study area is northern subtropical monsoon climate, with an average annual temperature of 15.5°C, an average annual precipitation of 1,099.1 mm, and a landscape of hilly areas. The plot size is 20 m × 20 m with a tree density of 1,475 trees/ha. The *Ginkgo* trees in the sample plot are planted artificially, and the average age of the trees is 22 years old. Before selecting the sample trees, we considered the tree height, diameter at breast height (DBH), and uprightness of the trees in the plot. We selected six trees with good growth and upright trunks, labeled a, b, c, d, e, and f, based on a combination of three sizes of DBH (>12 cm, 9–12 cm, and 6–9 cm). The roots of

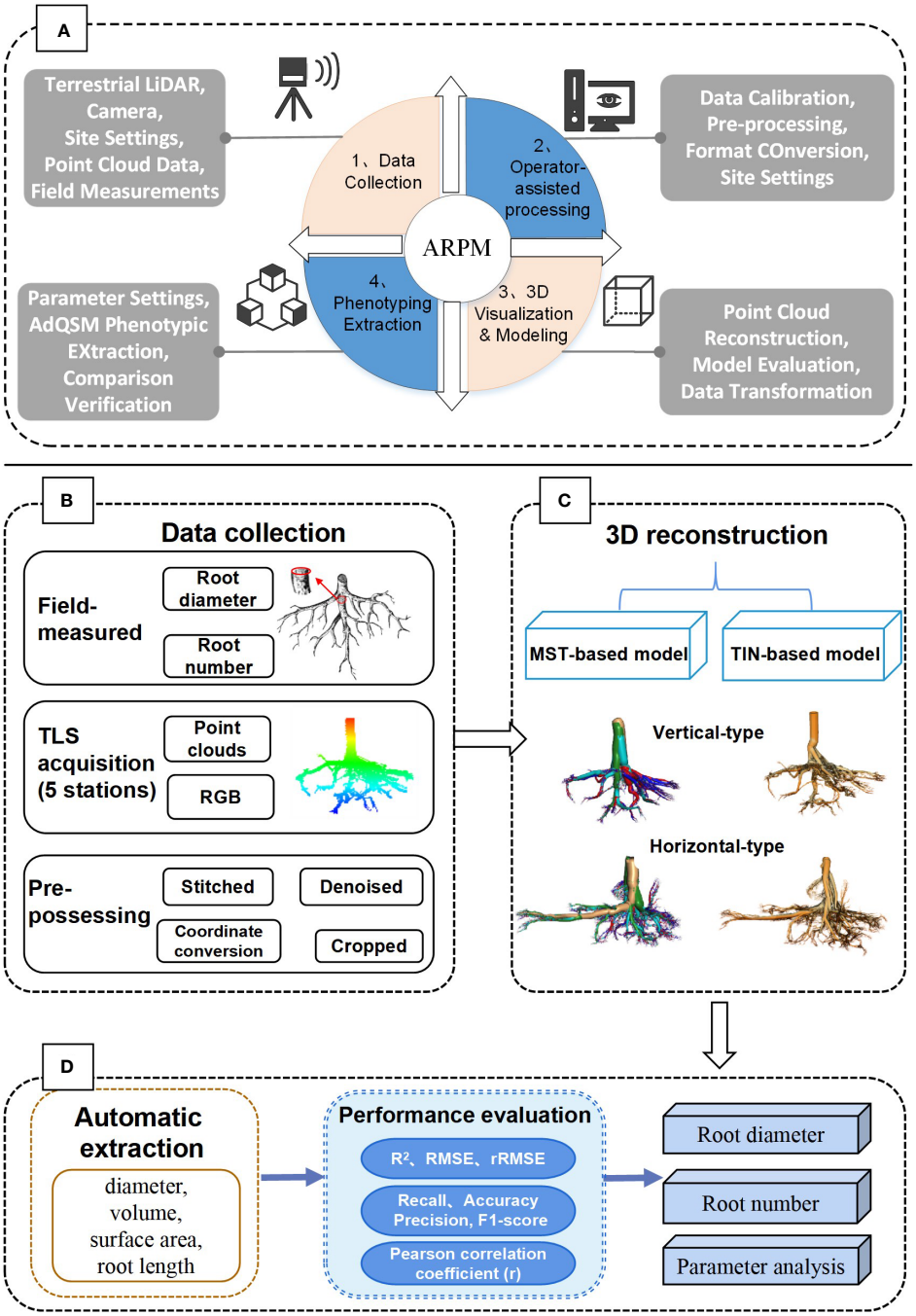


FIGURE 1
The overall process framework of the study. **(A)** The proposed ARPM approach for phenotypic measurement and extraction of tree roots, which includes four steps: data collection, operator-assisted processing, 3D visualization modeling, and phenotyping extraction. **(B)** Data collection and processing steps. **(C)** Methods of 3D reconstruction modeling. **(D)** Parameter extraction and performance evaluation.

the trees were cleaned and manicured before being raised entirely by an excavator and marked with red paint precisely for pointing the south direction of the trunk. The 22-year-old *Ginkgo* tree has a deep root system (root depth approximately 2 m), which causes its roots to be buried deep underground for a very long period. Given this fact, the surface of the roots is tangled up with soil and fine roots, making it challenging to determine the topological structure of the roots. The fine roots were cut back to highlight the RSA. Coarse roots with a base diameter larger than 0.5 cm were measured.

The TLS, which combines LiDAR and a digital camera, is one of the more crucial pieces of hardware in the ARPM. The fixed *Ginkgo* trees' roots were scanned using the RIEGL VZ-400i Terrestrial Laser Scanning (RIEGL Laser Measurement Systems, Inc., Horn, Austria) and integrated with a Nikon D810 camera (resolution: 7,380 × 4,928 pixels) to produce true color images and high-density

3D point clouds. The device has a measurement accuracy of ≤5 mm, a range of 800 m, a field of view of 100° × 360° (vertical × horizontal), and a maximum laser pulse repetition rate of up to 1.2 MHz. The angular resolution was set to 0.0007°C and 0.0005°C for the vertical and horizontal angles, respectively. Five different scanning positions were evenly distributed around the target in the center, with an interval angle of approximately 70°C.

2.2 Data processing

2.2.1 Preprocessing

In this study, the ARPM approach provided the first attempt to use TLS as a sensor to obtain 3D point clouds of the roots. To achieve automatic site data stitching, the raw data from TLS scanning are fed into the accompanying RiSCAN PRO software (http://www.riegl.com/products/software-packages/riscan-pro/), with a registration error of less than 2 mm (Henning and Radtke, 2008). ICP (iterative closest point) is the foundation of the point clouds stitching algorithm. If the automatic stitching effect is inadequate, fine-tune manually to get each site point cloud as tightly fitted as possible. The TLS is equipped with a digital camera with a fixed focal length to acquire texture information from the target object's surface. The stitched point clouds were imported into LiDAR360 software (Beijing Green Valley Technology. Co., Ltd., China, https://www.lidar360.com), and the root point clouds were cropped out separately and then denoised. Usually, modeling is bottom-up; in terms of morphology, it is from the apical side to the basal side. Aboveground branches and underground roots are similar in morphology that they are classified as trunk or taproot, first-order lateral, second-order lateral, etc. Furthermore, given underground roots do not have leaves, the noise source and shielding area are reduced as compared to aboveground modeling. Hence, we proposed and tested a hypothesis for conducting bottom-up branch separation to root reconstruction and extraction. Since the root system is oriented from the root base to the root tip, the typical upright root systems need to be inverted for further modeling.

2.2.2 3D reconstruction

Three-dimensional quantitative structural modeling (3D QSM) can contribute to the knowledge of spatial distribution characteristics, traits, and growth of the root system (Smith et al., 2014). 3D reconstruction is a critical step in the developed ARPM. The developed ARPM takes 10 min to capture point clouds of one entire root system, and the visualization is done on the computer side by reading the data. The pipeline of the 3D reconstruction approaches is shown in Figure 1. For the plant reconstruction, there are mainly two approaches based on segmentation and skeleton (Raumonen et al., 2013). In this study, we primarily employ two different approaches to extract the skeleton and then generate the root model automatically. The first approach is MST-based, and the second is TIN-based. AdQSM is a method developed by improving on AdTree (Du et al., 2019) and TreeQSM (Raumonen et al., 2013). The workflow of AdQSM is shown in Figure 2. Figure 3 shows the front view of the processed TLS 3D point clouds of the Ginkgo root system. The key algorithm is to construct the MST using Dijkstra's shortest path algorithm (Dijkstra, 1959) to obtain the skeleton of a tree (Figure 4). The branches are then reconstructed based on K-means clustering and nonlinear least squares optimized cylinder fitting with the aim of obtaining a more refined geometric structure model (Fan, 2021). Tree roots can be divided into tap roots, first-order lateral roots, second-order lateral roots, etc (Ingram and Malamy, 2010; Wang, 2017). To grade the roots and assign them a specific color, the process is recreated from the bottom up by computing each branch node and its order (Figure 5). The algorithm is robust to issues like missing or insufficient regional point clouds.

The Point Cloud Automata Viewer (PCAV) (Tianhong Jiye Technology Development Co., Ltd., China, http://www.thjymap.com/pca) filters point clouds based on the Multi-Primitive TIN Progressive Densification (MPTPD) algorithm of object primitives to automate modeling by generating a triangular mesh tree skeleton (Lin and Zhang, 2014; Lin et al., 2016). It is a commercial software, and its interface is developed based on the opensource project CloudCompare (https://www.cloudcompare.org) (Rajendra et al., 2014) the root model reconstructed by PCAV is shown in Figure 6. The TIN representation uses the discrete data obtained from all sampling points and connects these discrete points (vertices of triangles) into continuous triangular surfaces according to the principle of optimized combination. When constructing a TIN from the point cloud, the normal vector and centroid of each triangle in the TIN are calculated as follows (Equations 1, 2) (Wu et al., 2021c). Assume the vertices of triangle Q_i are

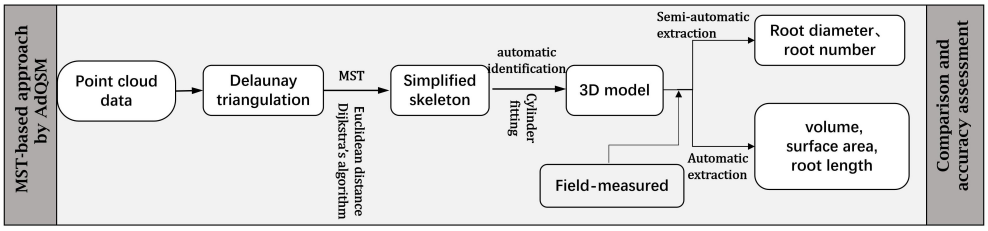


FIGURE 2
Flowchart of the MST-based approach by AdQSM for phenotypes measurement of *Ginkgo* roots. The point cloud data underwent initial Delaunay triangulation, followed by simplification using specific algorithms aimed at streamlining the skeleton. Subsequently, the simplified skeleton was molded to resemble cylinders to obtain a 3D model. Finally, the model was assessed through the extraction of root system parameters.

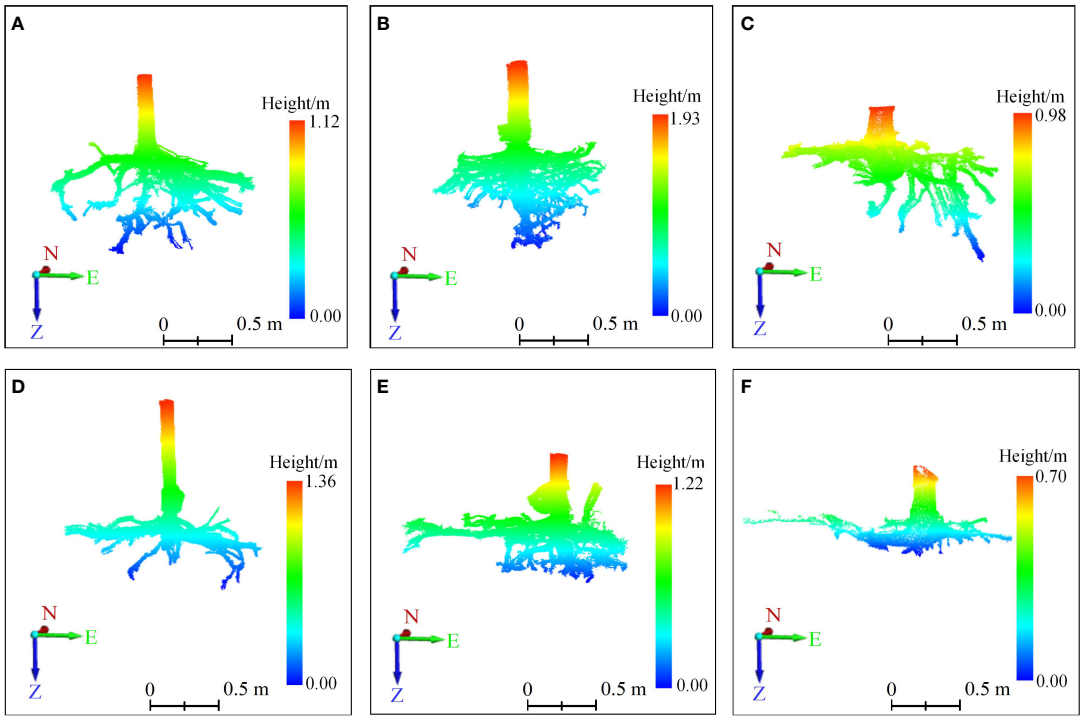


FIGURE 3
The 3D point clouds' front view of root systems. **(A–F)** correspond to six sample trees, with those labeled **(A–C)** indicating trees exhibiting vertical root growth, and those labeled **(D–F)** indicating trees with horizontal root growth. On the coordinate axis, N represents north, E stands for east, and Z indicates the direction of root growth. The root systems were scanned using TLS to generate point clouds, which were displayed after preprocessing steps like stitching, cropping, and denoising.

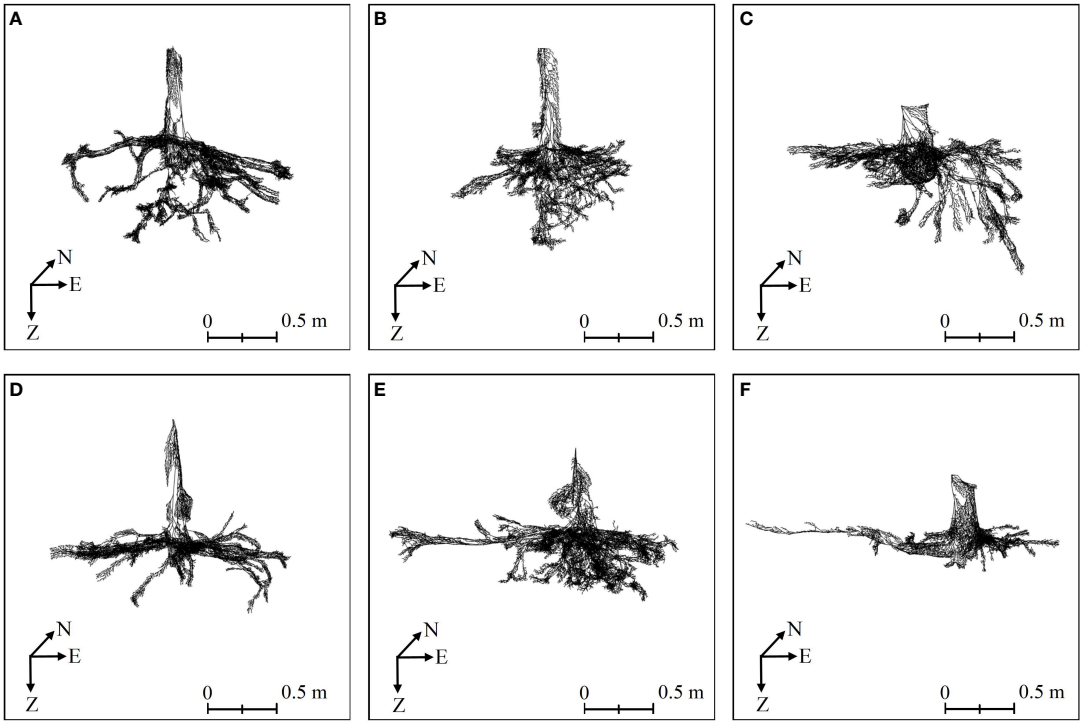


FIGURE 4
The root skeletons of six Ginkgo trees obtained by MST-based algorithm in AdQSM. **(A–F)** correspond to six sample trees. On the coordinate axis, N represents north, E stands for east, and Z indicates the direction of root growth.

G_1 (X_1, Y_1, Z_1), G_2 (X_2, Y_2, Z_2), and G_3 (X_3, Y_3, Z_3), in that order. The normal vector is N = (A, B, C), which can be expressed as the result of the cross-product between G_1G_2 and G_1G_3. The center of mass of triangle Q_i is C_k. Calculate the vectors between G_1 and G_2, and between G_1 and G_3, as $G_1G_2 = (X_2 - X_1, Y_2 - Y_1, Z_2 - Z_1)$ and $G_1G_3 = (X_3 - X_1, Y_3 - Y_1, Z_3 - Z_1)$, respectively.

$$\begin{aligned} A &= (Y_2 - Y_1)(Z_3 - Z_1) - (Z_2 - Z_1)(Y_3 - Y_1) \\ B &= (Z_2 - Z_1)(X_3 - X_1) - (X_2 - X_1)(Z_3 - Z_1) \\ C &= (X_2 - X_1)(Y_3 - Y_1) - (Y_2 - Y_1)(X_3 - X_1) \end{aligned} \quad (1)$$

The centroid C_k is calculated as follows:

$$C_k = \left(\frac{X_1 + X_2 + X_3}{3}, \frac{Y_1 + Y_2 + Y_3}{3}, \frac{Z_1 + Z_2 + Z_3}{3}\right) \quad (2)$$

2.2.3 Automatic extraction of phenotyping information

A quantitative understanding of the phenotyping traits of roots facilitates the understanding of the environmental–functional mechanisms of root action. In this study, multiple phenotyping traits (diameter, surface area, volume, and length) were extracted automatically using AdQSM v1.7 (open access: https://github.com/GuangpengFan/AdQSM) (Fan et al., 2020). The algorithms are modeled in a bottom-up manner to calculate the order and the number of bifurcation points of each grade of branches, which correspond to the grading number and the number of different grades of roots, respectively (Dong et al., 2021) (Ajmera et al., 2022). The branching order and basal diameter of the branches correspond to the grading order of the root system and the RD (Fan, 2021).

2.3 Model evaluation methods

2.3.1 Root diameter accuracy evaluation method based on point clouds and models

To evaluate the accuracy of the model, the basal diameters of approximately 2–4 RDs were randomly selected for each sample tree (19 RDs in total) and compared with the diameters extracted from the point cloud and the model, respectively. RDs were measured with vernier calipers serving as the true values, while those extracted from point clouds and models are considered as extracted values. The point clouds of roots are first segmented in LiDAR360. To make the upper end of the root morphology vertical, the projection and coordinate transformation were applied to the segmented individual roots. The least squares circle fitting algorithm was used to measure its basal diameter, with the average of several measurements adopted to determine the diameter. Likewise, CloudCompare was used to crop the point clouds of the root bases of the reconstructed models, and then importing it into LiDAR360 to obtain the model extracted values of RD. The R^2, RMSE, relative root-mean-square error (rRMSE), and MAE were calculated to estimate the level of consistency between the point clouds, 3D models' measurement data, and the raw data collected in the field (Equations 3–6). The metrics were calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2} \quad (4)$$

$$rRMSE = \frac{RMSE}{\bar{x}} \times 100\% \quad (5)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{x}_i - x_i| \quad (6)$$

where x_i is the measured root diameter; $\bar{x}$ is the mean of the measured root diameter; $\hat{x}_i$ is the estimated value of the root diameter model; n is the number of samples.

2.3.2 Accuracy evaluation methods for models to identify different levels of roots

The main traits to describe the root architecture are the number, diameter, and grade of the taproots and lateral roots (Xiao et al., 2014). From the 3D laser point clouds of *Ginkgo*, it can be found that its root type belongs to the horizontal or vertical root system (Zanetti et al., 2015; Xi, 2019), and the taproots are thick or not prominent. In this study, the number of first-order lateral roots (including the taproot) and second-order lateral roots was counted separately. To quantitatively distinguish taproots, first-order lateral roots, and second-order lateral roots, a computerized 3D visualization was utilized, combined with the visual interpretation of root point clouds or models. The accuracy of the models was evaluated by comparing it with the measured number of lateral roots measured by TLS. The performance of the model is evaluated using Recall, Precision, F1-score (F1), and Accuracy, all of which vary from 0 to 1 (Equations 7–10). The formulas are shown below, respectively.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (8)$$

$$\text{F1} = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (9)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

where TP, TN, FN, and FP refer to true positives, true negatives, false negatives, and false positives, respectively. F1-score is basically a harmonic mean of precision and recall.

2.3.3 Evaluation for the automatically extracted parameters

The correlation between the automatically extracted parameters and the true values is calculated by the Pearson correlation

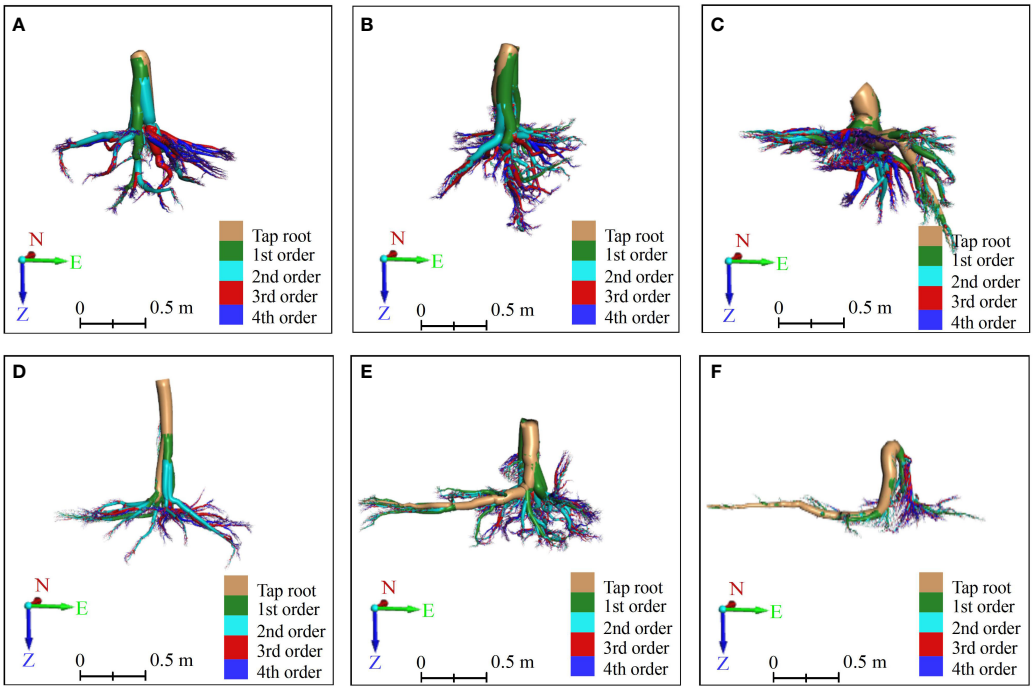


FIGURE 5
The front view of vertical-type **(A–C)** and horizontal-type **(D–F)** Ginkgo root systems, which were constructed by an AdQSM-based approach. The method reconstructed the roots in a bottom-up manner according to its growth rules, and calculated each branch node and order to grade the roots. Different orders of roots are colored by various colors. The tap roots, the first-order (1st) lateral roots, the second-order (2nd) lateral roots, the third-order (3rd) lateral roots, and the fourth-order (4th) lateral roots are colored brown, green, cyan, red, and blue, respectively. On the coordinate axis, N represents north, E stands for east, and Z indicates the direction of root growth.

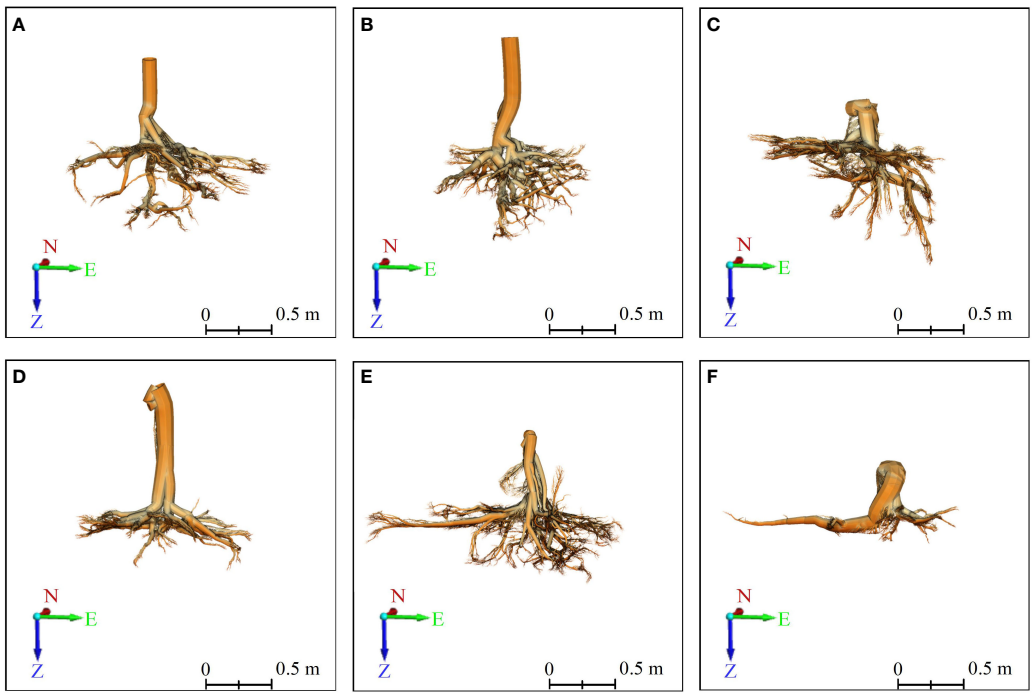


FIGURE 6
3D front view of six Ginkgo root systems, modeled by the TIN-based algorithm in PCAV, and CloudCompare presents the final model. **(A–F)** correspond to six sample trees, with those labeled **(A–C)** indicating trees exhibiting vertical root growth, and those labeled **(D–F)** indicating trees with horizontal root growth. On the coordinate axis, N represents north, E stands for east, and Z indicates the direction of root growth.

coefficient (r) (Equation 11). The calculation formula of r is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (11)$$

where x_i is the extracted parameter for the ith samples; $\bar{x}$ is the mean of x_i; y_i is the true value (typically measured manually) for the ith samples; $\bar{y}$ is the mean of y_i; n is the number of samples. The coefficient r ranges from −1 to 1. If the absolute value of r is close to 1, the linear correlation between the extracted values and the real value is stronger; if it is 0, there is no correlation between them.

3 Results

3.1 Accuracy assessment of the root diameter

In this study, two root models were selected to extract 19 RDs from six *Ginkgo* trees. A linear regression was fitted between the measured diameter (manually measured with a vernier caliper) and the values extracted from the point cloud and models, and a scatter plot was drawn (Figures 7A–C). For extracted diameters, the root point clouds acquired by TLS (R^2 = 0.99, MAE = 0.35 cm, RMSE = 0.47 cm, rRMSE = 8.21%) are highly consistent with the manually measured values, which are more accurate than the rebuilt root models. For the reconstructed MST-based (R^2 = 0.71, MAE = 1.79 cm, RMSE = 2.20 cm, rRMSE = 38.57%) and TIN-based (R^2 = 0.54, MAE = 5.62 cm, RMSE = 2.80 cm, rRMSE = 48.94%) 3D models, the former fits better and has higher accuracy when it is compared with manual measurement values. The RMSE of the two models ranged from 2 to 3 cm. There is a point with a large deviation (Figure 7C). Given the fact that the tree labeled c has more fine roots, this large deviation is mostly caused by the underestimation of this high-diameter root, as well as some root scans are not of good enough quality for the point cloud, resulting in a biased reconstructed model. This value reflects the variation in RD extracted by different models, with higher values in the TIN-based model tending to become saturated.

3.2 Accuracy evaluation of the model to identify different orders of roots

Table 2 demonstrates the accuracy of the two models for identifying different classes of root systems. In terms of the models' detection of the number of roots, the MST-based model (F1 = 0.81, Accuracy = 0.83) possesses slightly higher overall accuracy than the TIN-based model (F1 = 0.80, Accuracy = 0.82), and both Recall and Precision are roughly comparable, with mean values of approximately 0.8. In particular, the second-order lateral roots (F1 = 0.83) were somewhat better than the first-order lateral roots (F1 = 0.78), in terms of the root number for model detection. Although the accuracy of the first-order lateral roots is higher than that of second-order lateral roots, this is due to the fact that the number of second-order roots is typically more than that of the first-order roots, and an imbalanced distribution will impair the accuracy outcomes. Additionally, Figure 8 visualizes the comparison between Recall, Precision, F1, and Accuracy for the MST-based and TIN-based models, and the first- and second-order lateral roots.

3.3 Results of automatic phenotype extraction

Specific algorithms were implemented in AdQSM to quantitatively extract phenotypic parameters (e.g., volume, surface area, diameter, and length) while generating models (Table 3). The AdQSM algorithms were optimized by adjusting the height segmentation (HS) value and the cloud parameter (CP). The traits are extracted from the default values; i.e., the HS and the CP are set to 0.5 and 0.003, respectively. The CP represents the degree of point cloud thinning, i.e., down-sampling. Figure 9

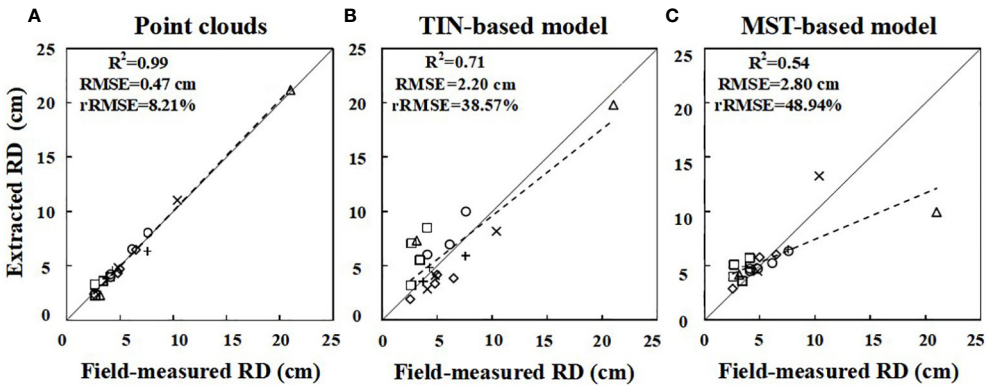


FIGURE 7
Linear fit of the root diameter (RD) extracted from point clouds **(A)** and the MST-based **(B)** and TIN-based **(C)** model to the measured values, respectively. The six symbols "□◇△+○×" represent roots from A, B, C, D, E, and F, respectively. The solid line represents the 1:1 line, and the dashed line indicates the regression equation.

TABLE 2 Overall accuracy assessment of the two models for identifying different orders of root systems.

Evaluation index	The first-order root		The second-order root		Overall evaluation	
	MST-based	TIN-based	MST-based	TIN-based	MST-based	TIN-based
Recall	0.76	0.75	0.87	0.87	0.82	0.81
Precision	0.83	0.83	0.80	0.81	0.81	0.82
F1-score	0.79	0.77	0.83	0.83	0.81	0.80
Accuracy	0.86	0.84	0.81	0.80	0.83	0.82

reveals the variation of branch volume, branch surface area, the average RD, and branch length with different CPs (range from 0.001 to 0.004). The figure indicates that the RD generally increases with the increase of CP values, while the root length decreases. The diameter and length of the roots showed opposite trends, resulting in an irregular variation of area and volume obtained by the balanced calculation of these two. Figure 10 shows the correlation analysis between the extracted RD at different CP values and the manually measured values. The extracted diameter values were significantly correlated with the measured values at CP values of 0.002 ($r = 0.94$, $p < 0.01$) and 0.003 ($r = 0.87$, $p < 0.05$). Therefore, it can be judged that 0.002 is the optimal CP value in this study. The CP values were directly related to the final extraction accuracy of the RD. Careful point cloud thinning is essential for reducing data redundancy and facilitating the accurate extraction of structural parameters. It is important to note that, given that only the diameter has been measured and other extracted characteristics were unavailable to be verified, the optimal value of CP should be considered purely as a reference point. The approach needs the coordinate file of the input point clouds to generate the model and finishes extracting phenotypes automatically. According to Table 3, it can be seen that the model's extracted diameters are smaller than those measured ones.

4 Discussion

The 3D point clouds from TLS show that the spatial distribution of the *Ginkgo* root system extends further horizontally than vertically, which is consistent with the results of Men (1986). Although the *Ginkgo*'s taproot was demonstrated to be distinguishable, certain sample trees lost their taproot dominance, possibly as a result of environmental factors such soil depth or water table, which, in turn, altered the root system's hierarchical structure (Freschet et al., 2021a). Liu et al. (2007) established a 3D static model of the root system of *Pinus tabuliformis* Carr. based on fractal theory. It concluded that there is a strong correlation between the RD and the root length, which can be used to predict the RD. However, this requires much more time and labor to measure the relevant parameters to establish the relationship model, which is complicated for tree species with more root branches. Guo et al. (2008) revealed a strong correlation between the RD and the root branch order. According to Li et al. (2016), the RD can be used to infer root biomass, while its variation is driven by soil, water, and nutrients. Conventional measurements of root system morphological and structural parameters are typically labor-intensive and time-consuming. Quantitative description of the link between phenotyping traits and the function of roots has

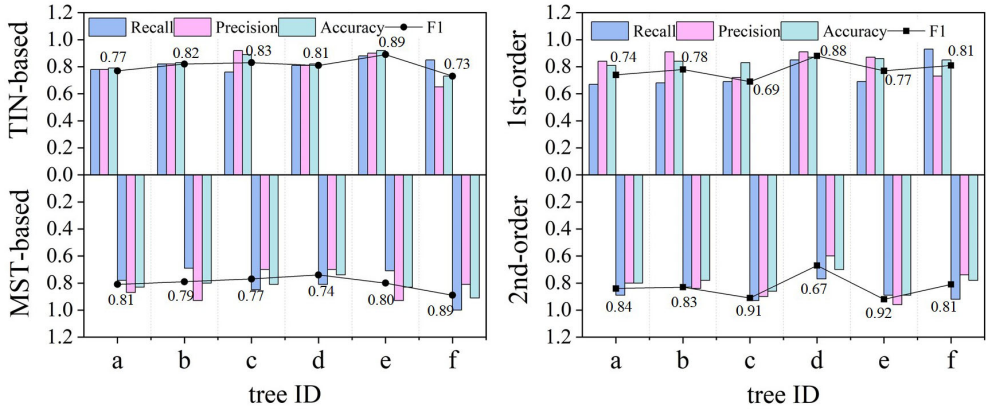


FIGURE 8
Comparison of the detection effect of the number of lateral roots from the two models regarding different evaluation metrics. The MST-based model (F1 = 0.81) surpassed the MST-based model (F1 = 0.80). The better performance of the second (2nd) order (F1 = 0.83) than the first (1st) order (F1 = 0.78) lateral roots is explained by the larger number of the former and their wider distribution, which makes them easier to be identified. Conversely, the 1st-order lateral roots are often misidentified (false negative) due to noise or occlusion.

TABLE 3 The phenotypic parameters of roots automatically extracted by the AdQSM method.

Tree ID	Number of roots ± SE	Measured diameter ± SE (cm)	Parameters for automatic extraction			
			Extracted diameter ± SE (cm)	Root volume (m^3)	Root surface area (m^2)	Root length (m)
a	30 ± 3	4.37 ± 0.57	3.73 ± 0.62	0.01 ± 0.00	2.06 ± 0.35	108.11 ± 10.21
b	52 ± 5	4.07 ± 0.37	2.77 ± 0.86	0.02 ± 0.00	3.36 ± 0.36	281.25 ± 16.58
c	28 ± 2	7.75 ± 0.88	5.44 ± 1.02	0.05 ± 0.01	6.59 ± 0.77	248.29 ± 13.57
d	32 ± 4	3.60 ± 0.82	2.50 ± 0.97	0.02 ± 0.00	2.44 ± 0.28	175.88 ± 14.23
e	55 ± 3	4.66 ± 0.35	4.51 ± 0.69	0.05 ± 0.01	7.39 ± 1.04	365.02 ± 18.69
f	19 ± 2	5.21 ± 1.15	3.70 ± 1.54	0.04 ± 0.01	4.57 ± 0.83	234.83 ± 15.47

SE refers to standard error. Each data value denotes the mean ± SE. The number of roots here represents only the number of the taproot and the first- and second-order lateral roots.

been the endeavor of many scholars. van Dusschoten et al. (2016) estimated root length in different RD classes for maize based on MRI images and established linear regression relationships with an R^2 and RMSE of 0.66 and 0.68 cm, respectively. In this study, the R^2 and RMSE of the RD of *Ginkgo* derived from TLS were 0.99 and 0.47 cm, respectively, which were highly consistent with the measured values and can substitute manual measurement. MRI is usually performed indoors, and the TLS is relatively flexible in the sites of use and less labor-intensive.

Zhu et al. (2014) employed GPR to obtain 3D images of pine root systems and indirectly estimate the underground biomass by establishing the model of RD. It took 2 to 3 h for GPR to scan a root system and could vaguely distinguish the root distribution position from an image with a low resolution. The results demonstrated that the RD error was between 13% and 16%, when the best RD estimation model was tested against the measured data. The electromagnetic waves emitted by GPR are affected by the dielectric constant. The resolution of the medium is mainly

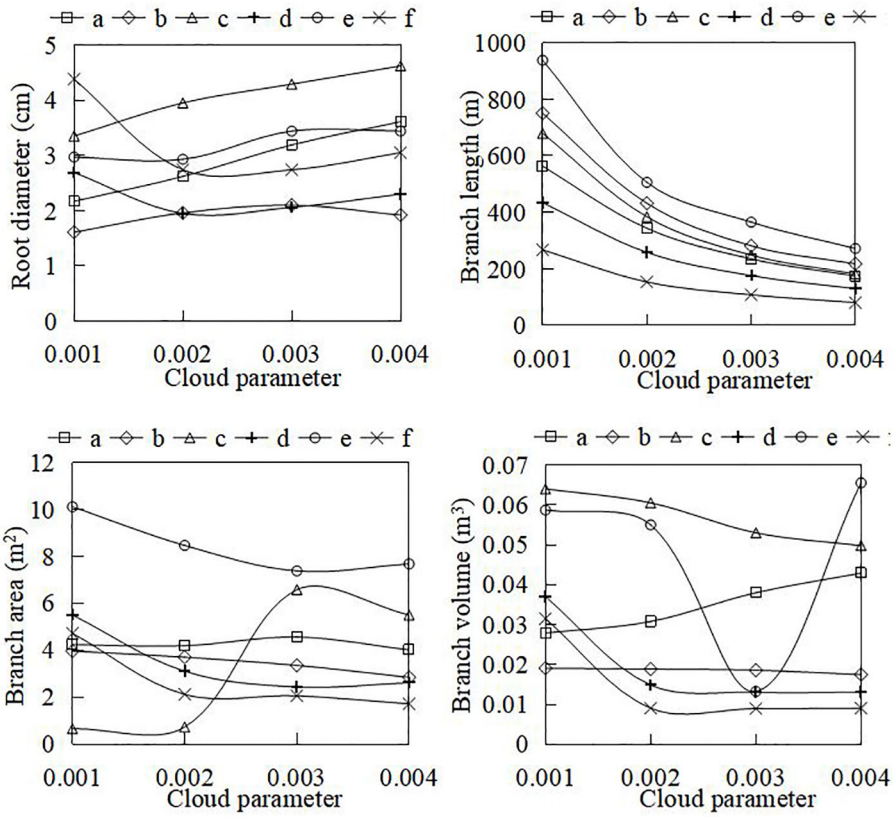


FIGURE 9
The variation of automatically extracted parameters (e.g., root diameter, branch volume, branch surface area, and branch length) at different cloud parameter (CP) coefficients. The CP represents the dilution rate during point cloud processing.

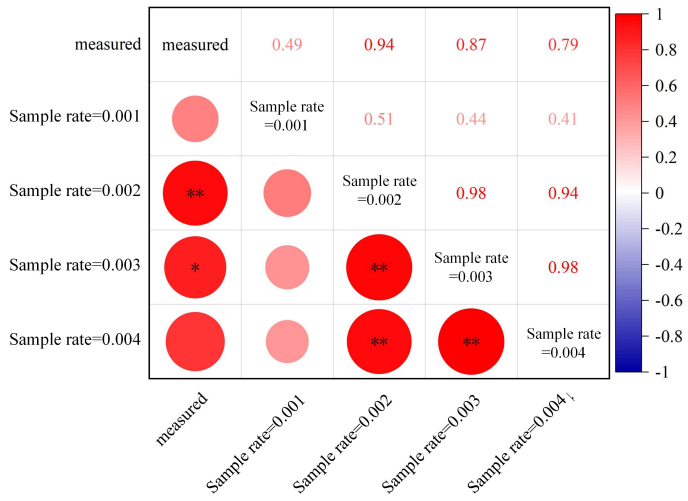


FIGURE 10
Correlations between automatically extracted root diameters at different sampling rates (cloud parameter, CP) of point clouds from AdQSM and manually measured values. $^{**}p < 0.01$; $^{*}p < 0.05$.

adopted to detect coarse roots and their distribution. Miao et al. (2022) utilized TLS to scan maize, convert point clouds into images, and assess stem thickness using elliptical fitting. The results achieved high accuracy, allowing for the quick determination of crop phenotypic traits. However, when 3D point clouds are turned into 2D images for measurement, it not only increases the source of error and decreases measurement accuracy, but also lengthens the data processing time. Furthermore, because there were more sites to scan, it took longer and increased the risk of error (Disney, 2019). In this study, the root system can be scanned by TLS to directly acquire the millimeter-level accurate 3D morphological structure and realistic texture information of the root system. The five-station scanning method takes approximately 20 min to obtain the object's high-density point clouds. In addition, the TLS point cloud-based measurements in this study improved at least 30% in terms of efficiency and achieved a higher level of accuracy. The 3D visualization of root systems and model reconstruction is essential for understanding the morphology, structure, and function of plant root systems (Wu et al., 2021a). Despite the fact that there are various ways for modeling roots, we still require a systematic approach as most have limitations and are not very universal. Studies on 3D model visualization of roots mostly focus on studying monocotyledons (Delory et al., 2022). Han and Kuo (2018) constructed a 3D image of the rice root system and quantified phenotypic traits, such as lateral root number and surface area.

To explore the dynamic growth process of the rice root system, Yang et al. (2020) proposed a 3D growth model based on a differential L-system. The model fitted the total root length and surface area to the measured values with an accuracy of more than 95%. In this study, woody plant roots were investigated. The results of this study indicated that the software for single-tree modeling could be implemented to root modeling, through the coordinate transformation to represent the complex architecture of roots completely. Regarding the RD extraction, the RMSE of the diameter was controlled within 2 to 3 cm for both models, and the models were well-performed. The model value of the RD is generally higher than the measured value. The overestimation may be caused by the misalignment of the point clouds, the incorrect recognition during modeling, and the mistaking noise as a component. By improving the registration accuracy and precise denoising, these issues can be alleviated and the model accuracy can be increased. In terms of root detection results, the MST-based and TIN-based models could correctly detect most first-order and second-order lateral roots. The results showed that the overall accuracy of the MST-based model (F1 = 0.81, Accuracy = 0.83) was slightly higher than that of the TIN-based model (F1 = 0.80, Accuracy = 0.82). The F1-score values of second-order lateral roots were higher than those of the first-order ones. Specifically, the sample tree labeled e (total number of roots is 55) had the highest accuracy, while the sample d (total number of roots is 32) had the lowest accuracy. These may be because the larger the number of roots, the wider the spatial distribution area, and the easier to be scanned by the laser, thereby weakening the influence of environmental factors and improving the detection accuracy.

Yang (2021) utilized SketchUp for 3D modeling of slope protection plants, and the three root architecture parameters extracted by the model were highly linearly correlated with the true values. The study required manual measurement of RDs and coordinates, which was relatively time-consuming and laborious. In this study, root scanning using TLS can obtain not only the 3D structure of the root system, but also the true color image as well as the coordinates. This can reduce the error of human measurement. Though the two models in this study performed worse than SketchUp, automatic modeling makes modeling reasonably straightforward (only takes 3–15 min) and enhances the efficiency of digitization processing. The possible reasons for the lower accuracy are that the model algorithm is developed mainly for the aboveground part of the plant and is prone to systematic errors. Alternatively, this was caused by the influence of environmental

factors in scanning the root system, which resulted in point cloud noise. Hui et al. (2022) extracted the number of first-order lateral roots by segmenting and scanning the image. However, this method is time-consuming and cannot directly capture the true 3D spatial distribution structure of the root system, as it is based on 2D images. Compared with other methods, the millimeter resolution and high penetration of TLS can accurately capture the structure and texture of the root system. Specific algorithms can also be implemented for the 3D reconstruction of the entire root system to quantify more traits quickly, efficiently, and with high throughput. The 3D reconstruction of root systems based on point clouds overcomes the limitation of traditional 2D image-based modeling with a single perspective.

Nevertheless, this study also has some limitations. The number of samples should be expanded, and differences in age, tree species, site conditions, and culture methods (cuttings and seedlings) of the samples should be taken into account. Currently, this paper is based on a semi-automated method for extracting the parameters of a 3D root model. The approaches primarily focus on separating and reconstructing branches and leaves aboveground. However, it is recommended that future research focuses on developing specialized algorithms for the separation and reconstruction of 3D root systems to enable automatic extraction of root parameters. The lack of measured data on root length in this study made it unavailable to validate the automated extraction of phenotypic traits, such as surface area and volume. In this study, fine roots less than 5 cm in diameter were clipped, due to their potential to introduce noise into the point cloud data. Consequently, this exclusion led to the loss of characterization of the fine roots and obscured their contribution to traits, such as surface area and the total root length. Owing to the presence of noise caused by fine roots in the point cloud, the accuracy analysis presented in this study has been restricted to the first-order lateral roots (including the tap root) and the second-order lateral roots. This study serves as an initial application of tree modeling methods for the extraction of phenotypic parameters of roots. We aim to validate further levels of roots using this approach in the future, thus optimizing the model and enhancing its robustness. This study provides a technical reference for the extraction of 3D root structure parameters of other trees. Refined extraction of root phenotypes can help improve our understanding of carbon and nitrogen allocation in tree organs and potentially improve future forest genetic gains. Based on TLS, it is hoped that future researchers will continue to develop methodological techniques to reconstruct tree root systems and be able to automatically extract phenotyping traits. That would be crucial for evaluating the analysis of spatial distribution structure, forest biomass, and growth structure for trees. This will facilitate improved understanding of precise plant cultivation, integration of phenotypes and genotypes, exploration of physiological and biochemical plant properties, and enhanced mechanical anchor of root systems.

5 Conclusion

There is still a lack of high-throughput data collection and modeling approach for root systems of trees. In this study, a new approach for quantifying root phenotyping based on ARPM was developed. This approach provides a potential avenue for improving 3D modeling algorithms and offers a new impetus for root phenotyping measurements. High-precision TLS point clouds can access sophisticated 3D structures of the root system. Compared to existing methods, the developed ARPM approach offers numerous advantages, including wider site applicability, reduced time and labor costs, and increased data collection and analysis efficiency and accuracy. Fitting of the diameter and the number of lateral roots showed that TLS is a reliable means to obtain root information effectively with high accuracies. The reconstructed models based on point clouds can not only present the spatial distribution and topology of the root system but also quantitatively extract the corresponding phenotyping traits.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

YL: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. KZ: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. LC: Conceptualization, Project administration, Resources, Supervision, Writing – review & editing.

.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was funded by the National Natural Science Foundation of China (32101521) and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (21KJB220003).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Ajmera, I., Henry, A., Radanielson, A. M., Klein, S. P., Ianevski, A., Bennett, M. J., et al. (2022). Integrated root phenotypes for improved rice performance under low nitrogen availability. *Plant Cell AND Environ.* 45, 805–822. doi: 10.1111/pce.14284

Alani, A. M., and Lantini, L. (2020). Recent advances in tree root mapping and assessment using non-destructive testing methods: A focus on ground penetrating radar. *Surv. Geophys.* 41, 605–646. doi: 10.1007/s10712-019-09548-6

Amato, M., Basso, B., Celano, G., Bitella, G., Morelli, G., and Rossi, R. (2008). *In situ* detection of tree root distribution and biomass by multi-electrode resistivity imaging. *Tree Physiol.* 28, 1441–1448. doi: 10.1093/treephys/28.10.1441

Cao, L., Coops, N. C., Innes, J., Dai, J. S., and She, G. H. (2014). Mapping above- and below-ground biomass components in subtropical forests using small-footprint liDAR. *Forests* 5, 1356–1373. doi: 10.3390/f5061356

Chen, F., Zhou, X., Li, Y., Yang, Z., Wang, G., and Yan, L. (2018). Design and implementation of plant root 3D architecture measurement system. *Comput. Eng. Appl.* 54, 249–255. doi: 10.3778/j.issn.1002-8331.1612-0381

Danjon, F., Fourcaud, T., and Bert, D. (2005). Root architecture and wind-firmness of mature Pinus pinaster. *New Phytol.* 168, 387–400. doi: 10.1111/j.1469-8137.2005.01497.x

Danjon, F., and Reubens, B. (2008). Assessing and analyzing 3D architecture of woody root systems, a review of methods and applications in tree and soil stability, resource acquisition and allocation. *Plant Soil* 303, 1–34. doi: 10.1007/s11104-007-9470-7

Danjon, F., Sinoquet, H., Godin, C., Colin, F., and Drexhage, M. (1999). Characterisation of structural tree root architecture using 3D digitising and AMAPmod software. *Plant Soil* 211, 241–258. doi: 10.1023/A:1004680824612

de Dorlodot, S., Forster, B., Pages, L., Price, A., Tuberosa, R., and Draye, X. (2007). Root system architecture: opportunities and constraints for genetic improvement of crops. *Trends Plant Sci.* 12, 474–481. doi: 10.1016/j.tplants.2007.08.012

Delory, B. M., Hernandez-Soriano, M. C., Wacker, T. S., Dimitrova, A., Ding, Y., Greeley, L. A., et al. (2022). A snapshot of the root phenotyping landscape in 2021. *bioRxiv*., 2022-01. doi: 10.1101/2022.01.28.478001

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numer. Math. (Heidelb)* 1, 269–271. doi: 10.1007/BF01386390

Disney, M. (2019). Terrestrial LiDAR: a three-dimensional revolution in how we look at trees. *New Phytol.* 222, 1736–1741. doi: 10.1111/nph.15517

Dong, Y. Q., Fan, G. P., Zhou, Z. W., Liu, J. C., Wang, Y. G., and Chen, F. X. (2021). Low cost automatic reconstruction of tree structure by adQSM with terrestrial close-range photogrammetry. *Forests* 12, 1021. doi: 10.3390/f12081020

Du, S. L., Lindenbergh, R., Ledoux, H., Stoter, J., and Nan, L. L. (2019). AdTree: accurate, detailed, and automatic modelling of laser-scanned trees. *Remote Sens.* 11, 2074. doi: 10.3390/rs11182074

Fan, G. P. (2021). *Development and application of Tree Quantitative Structure Models Based on LiDAR Point Clouds* (Dissertation). Beijing Forestry University, Beijing, China.

Fan, G., Nan, L., Dong, Y., Su, X., and Chen, F. (2020). AdQSM: A new method for estimating above-ground biomass from TLS point clouds. *Remote Sens.* 12, 3089. doi: 10.3390/rs12183089

Freschet, G. T., Pages, L., Iversen, C. M., Comas, L. H., Rewald, B., Roumet, C., et al. (2021a). A starting guide to root ecology: strengthening ecological concepts and standardising root classification, sampling, processing and trait measurements. *New Phytol.* 232, 973–1122. doi: 10.1111/nph.17572

Freschet, G. T., Roumet, C., Comas, L. H., Weemstra, M., Bengough, A. G., Rewald, B., et al. (2021b). Root traits as drivers of plant and ecosystem functioning: current understanding, pitfalls and future research needs. *New Phytol.* 232, 1123–1158. doi: 10.1111/nph.17072

Fu, Y. R., and Zhang, W. G. (2019). Effects of soil moisture content and root depth on anti-overturing performance of Ginkgo biloba seedings. *J. Civ. Environ. Eng.* 41, 42–48. doi: 10.11835/j.issn.2096-6717.2019.093

Gu, J. C., Xu, Y., Dong, X. Y., Wang, H. F., and Wang, Z. Q. (2014). Root diameter variations explained by anatomy and phylogeny of 50 tropical and temperate tree species. *Tree Physiol.* 34, 415–425. doi: 10.1093/treephys/tpu019

Guo, D. L., Xia, M. X., Wei, X., Chang, W. J., Liu, Y., and Wang, Z. Q. (2008). Anatomical traits associated with absorption and mycorrhizal colonization are linked to root branch order in twenty-three Chinese temperate tree species. *New Phytol.* 180, 673–683. doi: 10.1111/j.1469-8137.2008.02573.x

Han, T. H., and Kuo, Y. F. (2018). Developing a system for three-dimensional quantification of root traits of rice seedlings. *Comput. Electron. Agric.* 152, 90–100. doi: 10.1016/j.compag.2018.07.001

Henning, J., and Radtke, P. (2008). Multiview range-image registration for forested scenes using explicitly-matched tie points estimated from natural surfaces. *Isprs J. Photogrammetry Remote Sens. - ISPRS J. PHOTOGRAMM* 63, 68–83. doi: 10.1016/j.isprsjprs.2007.07.006

Hui, F., Xie, Z. W., Li, H. G., Guo, Y., Li, B. G., Liu, Y. L., et al. (2022). Image-based root phenotyping for field-grown crops: An example under maize/soybean intercropping. *J. Integr. Agric.* 21, 1606–1619. doi: 10.1016/S2095-3119(20)63571-7

Ingram, P. A., and Malamy, J. E. (2010). "Root System Architecture" in Advances in Botanical Research, eds.," J.-C. Kader and M. Delseny (New York, United States: Academic Press), 75–117. doi: 10.1016/B978-0-12-380868-4.00002-8

Iyer-Pascuzzi, A. S., Symonova, O., Mileyko, Y., Hao, Y. L., Belcher, H., Harer, J., et al. (2010). Imaging and analysis platform for automatic phenotyping and trait ranking of plant root systems. *Plant Physiol.* 152, 1148–1157. doi: 10.1104/pp.109.150748

Koeser, A. K., Roberts, J. W., Miesbauer, J. W., Lopes, A. B., Kling, G. J., Lo, M., et al. (2016). Testing the accuracy of imaging software for measuring tree root volumes. *Urban For. Urban Greening* 18, 95–99. doi: 10.1016/j.ufug.2016.05.009

Kuijken, R. C. P., van Eeuwijk, F. A., Marcelis, L. F. M., and Bouwmeester, H. J. (2015). Root phenotyping: from component trait in the lab to breeding. *J. Exp. Bot.* 66, 5389–5401. doi: 10.1093/jxb/erv239

Li, Z. J., Chen, X., Shu, J. H., Sun, H. Y., and Cong, R. C. (2015). Research methods for tree root system distribution and structure: A review. *World For. Res.* 28, 13–18. doi: 10.13348/j.cnki.sjlyyj.2015.03.003

Li, X. Q., Dai, S. G., Long, H. L., Zhang, W., and Gu, Y. J. (2016). Fine root morphology and biomass of phoebe zhennan provenance seedlings. *J. Trop. Subtrop. Bot.* 24, 208–214. doi: 10.11926/j.issn.1005-3395.2016.02.012

Lin, Y. (2015). LiDAR: An important tool for next-generation phenotyping technology of high potential for plant phenomics? *Comput. Electron. Agric.* 119, 61–73. doi: 10.1016/j.compag.2015.10.011

Lin, X. G., and Zhang, J. X. (2014). Segmentation-based filtering of airborne liDAR point clouds by progressive densification of terrain segments. *Remote Sens.* 6, 1294–1326. doi: 10.3390/rs6021294

Lin, X., Zhang, J., Ning, X., Duan, M., and Zang, Y. (2016). Filtering of point clouds using fusion of three types of primitives including points, objects and key points. *Acta Geod. Cartogr. Sin.* 45, 1308–1317. doi: 10.11947/j.AGCS.2016.20160372

Liu, J. J. (1998). A review on root ecology of forest trees. *J. Northwest For. Coll.*, 13, 76–80. doi: CNKI:SUN:XBLX.0.1998-03-014

Liu, J. P. (2016). *Trees Parameters Extraction Study from Terrestrial Laser Scanning Data*. Beijing, China: Chinese Academy of Forestry.

Liu, X. P., Chen, L. H., Song, W. F., and Wu, Y. L. (2007). Fractal analysis on morphology distribution of the pinus tabuliformis carr. Root system. *Bull. Soil Water Conserv.* 27, 47–50+54. doi: 10.13961/j.cnki.stbctb.2007.01.011

Liu, J. P., Zhang, H. Q., Liu, M., Li, Y. L., and Li, W. N. (2016). Extraction of individual tree branch parameters from skeleton model based on point cloud data. *J. Beijing For. Univ.* 38, 15–20. doi: 10.13332/j.1000-1522.20150490

Lu, Y., Wang, Y., Chen, Z., Khan, A., Salvaggio, C., and Lu, G. (2021). 3D plant root system reconstruction based on fusion of deep structure-from-motion and IMU. *Multimedia Tools Appl.* 80, 17315–17331. doi: 10.1007/s11042-020-10069-3

Lynch, J. (1995). Root architecture and plant productivity. *Plant Physiol.* 109, 7–13. doi: 10.1104/pp.109.1.7

Lynch, J. P. (2007). Roots of the second green revolution. *Aust. J. Bot.* 55, 493–512. doi: 10.1071/BT06118

Men, X. Y. (1986). Study on root distribution and growth dynamics of Ginkgo biloba L. *For. Sci. Technol.*, 4–6. doi: 10.13456/j.cnki.lykt.1986.10.003

Miao, Y. L., Peng, C., Wang, L. Y., Qiu, R. C., Li, H., and Zhang, M. (2022). Measurement method of maize morphological parameters based on point cloud image conversion. *Comput. Electron. Agric.* 199, 107174. doi: 10.1016/j.compag.2022.107174

Okamoto, Y., Ikeno, H., Hirano, Y., Tanikawa, T., Yamase, K., Todo, C., et al. (2022). 3D reconstruction using Structure-from-Motion: a new technique for morphological measurement of tree root systems. *Plant Soil* 477, 829–841. doi: 10.1007/s11104-022-05448-8

Perret, J. S., Al-Belushi, M. E., and Deadman, M. (2007). Non-destructive visualization and quantification of roots using computed tomography. *Soil Biol. Biochem.* 39, 391–399. doi: 10.1016/j.soilbio.2006.07.018

Rajendra, Y. D., Mehrotra, S. C., Kale, K. V., Manza, R. R., Dhumal, R. K., Nagne, A. D., et al. (2014). Evaluation of partially overlapping 3D point cloud's registration by using ICP variant and cloudcompare. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 40, 891–897.

Raumonen, P., Kaasalainen, M., Akerblom, M., Kaasalainen, S., Kaartinen, H., Vastaranta, M., et al. (2013). Fast automatic precision tree models from terrestrial laser scanner data. *Remote Sens.* 5, 491–520. doi: 10.3390/rs5020491

Seethepalli, A., Dhakal, K., Griffiths, M., Guo, H. C., Freschet, G. T., and York, L. M. (2021). RhizoVision Explorer: open-source software for root image analysis and measurement standardization. *AoB Plants* 13, plab056. doi: 10.1093/aobpla/plab056

Shao, M. R., Jiang, N., Li, M., Howard, A., Lehner, K., Mullen, J. L., et al. (2021). Complementary phenotyping of maize root system architecture by root pulling force and X-ray imaging. *Plant Phenomics* 2021, 12. doi: 10.34133/2021/9859254

Shen, X., Cao, L., Coops, N. C., Fan, H. C., Wu, X. Q., Liu, H., et al. (2020). Quantifying vertical profiles of biochemical traits for forest plantation species using advanced remote sensing approaches. *Remote Sens. Environ.* 250, 20. doi: 10.1016/j.rse.2020.112041

Smith, A., Astrup, R., Raumonen, P., Liski, J., Krooks, A., Kaasalainen, S., et al. (2014). Tree root system characterization and volume estimation by terrestrial laser scanning and quantitative structure modeling, *Forests.* 5, 3274–3294. doi: 10.3390/f5123274

Song, W. F., Chen, L. H., and Liu, L. P. (2008). Root reinforcement of soil: a review. *J. Zhejiang A F Univ.* 25, 376–381. doi: 10.3969/j.issn.2095-0756.2008.03.022

Spanos, I., Ganatsas, P., and Raftoyannis, Y. (2008). The root system architecture of young Greek fir (Abies cephalonica Loudon) trees. *Plant Biosyst.* 142, 414–419. doi: 10.1080/11263500802151082

Steingrobe, B. (2001). Root renewal of sugar beet as a mechanism of P uptake efficiency. *J. Plant Nutr. Soil Sci.* 164, 533–539. doi: 10.1002/1522-2624(200110)164:5<533::AID-JPLN533>3.0.CO;2-D

Todo, C., Ikeno, H., Yamase, K., Tanikawa, T., Ohashi, M., Dannoura, M., et al. (2021). Reconstruction of conifer root systems mapped with point cloud data obtained by 3D laser scanning compared with manual measurement. *Forests*. 12, 1117. doi: 10.3390/f12081117

Topp, C. N., Bray, A. L., Ellis, N. A., and Liu, Z. B. (2016). How can we harness quantitative genetic variation in crop root systems for agricultural improvement? *J. Integr. Plant Biol.* 58, 213–225. doi: 10.1111/jipb.12470

Tracy, S. R., Nagel, K. A., Postma, J. A., Fassbender, H., Wasson, A., and Watt, M. (2020). Crop improvement from phenotyping roots: highlights reveal expanding opportunities. *Trends Plant Sci.* 25, 105–118. doi: 10.1016/j.tplants.2019.10.015

van Dusschoten, D., Metzner, R., Kochs, J., Postma, J. A., Pflugfelder, D., Bühler, J., et al. (2016). Quantitative 3D analysis of plant roots growing in soil using magnetic resonance imaging. *Plant Physiol.* 170, 1176–1188. doi: 10.1104/pp.15.01388

Villordon, A. Q., Ginzberg, I., and Firon, N. (2014). Root architecture and root and tuber crop productivity. *Trends Plant Sci.* 19, 419–425. doi: 10.1016/j.tplants.2014.02.002

Volkmar, K. M. (1993). A comparison of minirhizotron techniques for estimating root length density in soils of different bulk densities. *Plant Soil* 157, 239–245. doi: 10.1007/BF00011052

Wang, F. G. (2017). Comparision of fine roots of Pinus koraienschangeis, Picea mongolica and Fraxinus mandahurica. *Shangdong For. Sci. Technol.* 47, 66–68. doi: 10.3969/j.issn.1002-2724.2017.03.014

Wang, D., Liang, X. L., Mofack, G., and Martin-Ducup, O. (2021). Individual tree extraction from terrestrial laser scanning data *via* graph pathing. *For. Ecosyst.* 8, 1-11. doi: 10.1186/s40663-021-00340-w

Wen, W. L., Guo, X. Y., Zhao, C. J., Wang, C. Y., and Xiao, B. X. (2015). Crop roots configuration and visualization: A review. *Sci. Agric. Sin.* 48, 436–448. doi: 10.3864/j.issn.0578-1752.2015.03.04

Wilhelm, J., Wojciechowski, T., Postma, J. A., Jollet, D., Heinz, K., Böckem, V., et al. (2022). Assessing the storage root development of cassava with a new analysis tool. *Plant Phenomics* 2022, 9767820. doi: 10.34133/2022/9767820

Wu, P. P., Tang, Z. Z., Yang, L., Peng, J., Zhang, H. H., and Shi, J. L. (2021a). Visualization of rice root system by 3D modeling: A review. *Fujian J. Agric. Sci.* 36, 972–980. doi: 10.19303/j.issn.1008-0384.2021.08.015

Wu, X., Wang, F. Y., Wang, M. C., Zhang, X. Q., Wang, Q., and Zhang, S. (2021c). A new method for automatic extraction and analysis of discontinuities based on TIN on rock mass surfaces. *Remote Sens.* 13, 2894. doi: 10.3390/rs13152894

Wu, Q., Zhang, W. X., Zhang, L. L., Sun, C. L., Liu, N. S., Yue, Y. B., et al. (2021b). Research progress on acquisition of plant root phenotype information. *Jiangsu Agric. Sci.* 49, 31–37. doi: 10.15889/j.issn.1002-1302.2021.05.006

Xi, B. (2019). Morphology, distribution, dynamic characteristics of poplar roots and its water uptake habits. *J. Beijing For. Univ. (Chin. Ed.)* 41, 37–49. doi: 10.12171/j.1000–1522.20190400

Xiao, Y. S., Peng, F. T., Zhang, Y. F., Qi, Y. J., Wang, G. F., Wang, X. L., et al. (2014). Effects of aeration cultivation on root architecture and nitrogen metabolism of young peach trees. *Sci. Agric. Sin.* 47, 1995–2002. doi: 10.3864/j.issn.0578-1752.2014.10.013

Yang, L. (2021). *3D Visual Simulation of Root System of Main Slope Protecting Plants in Loess Region and Extraction of Root Architecture Parameters and Application.* (Dissertation) Northwest A & F University, Xianyang City, Shaanxi Province, China.

Yang, L., Wu, P., Yang, S., and Shao, P. (2020). Research on the construction and visualization of A three-dimensional model of rice root growth. *Appl. Eng. Agric.* 36, 847–857. doi: 10.13031/aea.13543

Yin, J. D., Zhao, J. Y., Ge, Y. H., Feng, X., Tang, Y., and Ren, Z. Y. (2009). Application of digital techniques to dynamic monitoring of roots hydroponic cultured plants. *J. Northeast For. Univ.* 37, 71–74. doi: 10.3969/j.issn.1000-5382.2009.07.024

York, L. M., Cumming, J. R., Trusiak, A., Bonito, G., von Haden, A. C., Kalluri, U. C., et al. (2022). Bioenergy Underground: Challenges and opportunities for phenotyping roots and the microbiome for sustainable bioenergy crop production. *Plant phenom. J.* 5, e20028. doi: 10.1002/ppj2.20028

Zanetti, C., Vennetier, M., Mériaux, P., and Provansal, M. (2015). Plasticity of tree root system structure in contrasting soil materials and environmental conditions. *Plant Soil* 387, 35. doi: 10.1007/s11104-014-2253-z

Zhang, X., Knappett, J. A., Leung, A. K., Ciantia, M. O., Liang, T., and Danjon, F. (2020). Small-scale modelling of root-soil interaction of trees under lateral loads. *Plant Soil* 456, 289–305. doi: 10.1007/s11104-020-04636-8

Zhang, T., Song, L. N., Zhu, J. J., Wang, G. C., Li, M. C., Zheng, X., et al. (2021). Spatial distribution of root systems of Pinus sylvestris var. mongolica trees with different ages in a semi-arid sandy region of Northeast China. *For. Ecol. Manage.* 483, 118776. doi: 10.1016/j.foreco.2020.118776

Zhao, P.-F., Wang, Y.-Q., Yan, S.-X., Fan, L.-F., Wang, Z.-Y., Zhou, Q., et al. (2019). Electrical imaging of plant root zone: A review. *Comput. Electron. Agric.* 167, 105058. doi: 10.1016/j.compag.2019.105058

Zhou, K., and Cao, L. (2021). The status and prospects of remote sensing applications in precision silviculture. *Natl. Remote Sens. Bull.* 25, 423–438. doi: 10.11834/jrs.20210506

Zhu, S. P., Huang, C. L., Su, Y., and Sato, M. (2014). 3D ground penetrating radar to detect tree roots and estimate root biomass in the field. *Remote Sens.* 6, 5754–5773. doi: 10.3390/rs6065754

frontiers | Frontiers in Plant Science

TYPE Review
PUBLISHED 12 August 2024
DOI 10.3389/fpls.2024.1417912

OPEN ACCESS

EDITED BY
Dmitri Voronine,
University of South Florida, United States

REVIEWED BY
Dhananjay K. Pandey,
Amity University, Jharkhand, India
Boubacar Gano,
Donald Danforth Plant Science Center,
United States

*CORRESPONDENCE
Jauhar Ali
J.Ali@irri.org

RECEIVED 15 April 2024
ACCEPTED 15 July 2024
PUBLISHED 12 August 2024

CITATION
Khatibi SMH and Ali J (2024) Harnessing the power of machine learning for crop improvement and sustainable production. *Front. Plant Sci.* 15:1417912. doi: 10.3389/fpls.2024.1417912

Harnessing the power of machine learning for crop improvement and sustainable production

Seyed Mahdi Hosseiniyan Khatibi and Jauhar Ali*

Rice Breeding Platform, International Rice Research Institute, Los Baños, Laguna, Philippines

Crop improvement and production domains encounter large amounts of expanding data with multi-layer complexity that forces researchers to use machine-learning approaches to establish predictive and informative models to understand the sophisticated mechanisms underlying these processes. All machine-learning approaches aim to fit models to target data; nevertheless, it should be noted that a wide range of specialized methods might initially appear confusing. The principal objective of this study is to offer researchers an explicit introduction to some of the essential machine-learning approaches and their applications, comprising the most modern and utilized methods that have gained widespread adoption in crop improvement or similar domains. This article explicitly explains how different machine-learning methods could be applied for given agricultural data, highlights newly emerging techniques for machine-learning users, and lays out technical strategies for agri/crop research practitioners and researchers.

KEYWORDS

artificial intelligence, machine learning, deep learning, precision crop improvement, prediction model

1 Introduction

Naturally, humans' learning procedure is carefully or randomly monitoring surrounding events, grasping some experience and then predicting the next event, mainly occurring without human awareness. For instance, consider a human child who is learning how to talk. Basically, children do not know language learning techniques, procedures, or linguistics. Nevertheless, by listening to surrounding sounds, experimenting, and making mistakes, children gradually adjust their listening skills and simultaneously learn to talk and communicate in different situations. These procedures will continue until children feel confident enough to speak. Technically, they are learning how to talk by establishing a sound and an adequately accurate model of a whole set of procedures automatically and by testing the developed model again and again with surrounding voice

data and improving it to build a more precise model. The term "machine learning" typically refers to the procedures of finding relevant groups within data or fitting prediction models to a target dataset. In essence, machine learning aims to mimic or resemble human capacity and the ability to identify patterns using computation approaches. Machine learning is especially handy when the dataset being analyzed is huge or sophisticated beyond human ability to analyze it or when we want to build an automated platform for analyzing a target dataset by considering it to be time-efficient and repeatable. Agricultural data often have these characteristics. Over the past few decades, agricultural databases have experienced remarkable growth in quantity and multi-layer complexity. Having a solid grasp of the methods being employed and some valuable tools for interpreting this wealth of data is becoming increasingly crucial. Although machine learning has been engaged in the crop domain for many years, its usage in agriculture and crop improvement has now become so widespread that it is used in almost every discipline. Only recently, though, has the field started to examine the various strategies more closely to determine which ones work best in certain situations or whether they are suitable at all. This review aims to offer compact, sufficient, and explicit information and details on how to use machine-learning techniques for agricultural and crop improvement researchers. We do not seek to provide a comprehensive analysis and investigate the literature on machine-learning applications for crop improvement problems nor to get into the specific mathematical details of different machine-learning techniques (Liakos et al., 2018; Sharma et al., 2020). We focus on connecting specific machine-learning methods to various kinds of agricultural data. In addition, we will try to explain some best practices for approaching training and modeling improvement in real-world scenarios. The intrinsic intricacy of agricultural data poses opportunities and challenges for analytical methods in machine learning. We highlight common problems that undermine the validity of research and offer advice on how to overcome these challenges. The discussion of several machine-learning methods takes up most of this review, and we also provide explicit examples of how to use the strategy appropriately and understand the outcomes in each case. Traditional machine-learning techniques are included in the discussion as, in many situations; they continue to be the best options to apply. Our discussion covers techniques of deep learning, which shows satisfactory performance and is the best option for various machine-learning responsibilities. We also cover federated learning as a robust technique for having a machine-learning global crop improvement model to deal with future challenges such as climate change. We conclude by outlining the prospects for integrating machine learning into agricultural data analysis pipelines. When using machine learning in agriculture, there are two primary objectives. First, even though the collected data are sufficient or deficient, precise predictions should be made and used to direct further research endeavors. Since scientists are interested in understanding mechanisms, the other objective is to apply machine learning to enhance and increase the comprehension of crop improvement mechanisms, including several types of phenotypical, genotypical, biological, agronomic, and climatic mechanisms. We also summarize some of the limitations and applications of machine-learning approaches along with some data-related concerns for researchers in the crop improvement domain.

2 Shortlist of machine-learning applications for crop improvement and production

With emerging new technologies and approaches, large datasets are generated from different agricultural domains, particularly from the crop production domain. These vast datasets can easily feed into machine-learning approaches to help all beneficiaries optimize crop improvement systems. Even though machine-learning applications are extensive, their subcategories, mainly in crop quality (Elbasi et al., 2023; Attri et al., 2024), crop phenotyping (Gano et al., 2024), crop weed identification (Hu et al., 2021; Modi et al., 2023; Venkataraju et al., 2023), disease detection (Kulkarni and Shastri, 2024; Srinivas et al., 2024), crop recognition (Tian et al., 2021; Fu et al., 2023; Gafurov et al., 2023), crop-related microbiome improvements (Chang et al., 2017; Aguilar-Zambrano et al., 2023), and yield prediction (Van Klompenburg et al., 2020; Morales and Villalobos, 2023), were separated into crop development, production, and improvement, as shown in Figure 1.

3 Essential concepts

We discuss several fundamental ideas in machine learning and, whenever possible, present examples from agricultural literature to clarify these concepts.

3.1 Basic terms in machine learning

A dataset consists of several instances, or data points, that are conceptualized as individual experimental observations. Several fixed features describe each data point. Phenotype, genotype (SNPs), product price, and climatic parameters are a few examples of these features. Whatever we aim to do with a machine-learning model is specified objectively by a machine-learning task. For instance, we could predict the rate of price fluctuation at a particular point in time for a specific agricultural product with an experiment examining the cost of the crop product over time. In this instance, the features "cost of crop product" and "time" could be referred to as input features. The conversion rate, which would represent the anticipated output of the target model at a specific moment, is the quantity we are interested in forecasting. Input and output features of a model can be as many as desired. Features could be either categorical (accepting just discrete values) or continuous (continuous numerical values are used). Technically, categorical features are usually binary in nature, meaning they can be 1 (true) or 0 (false).

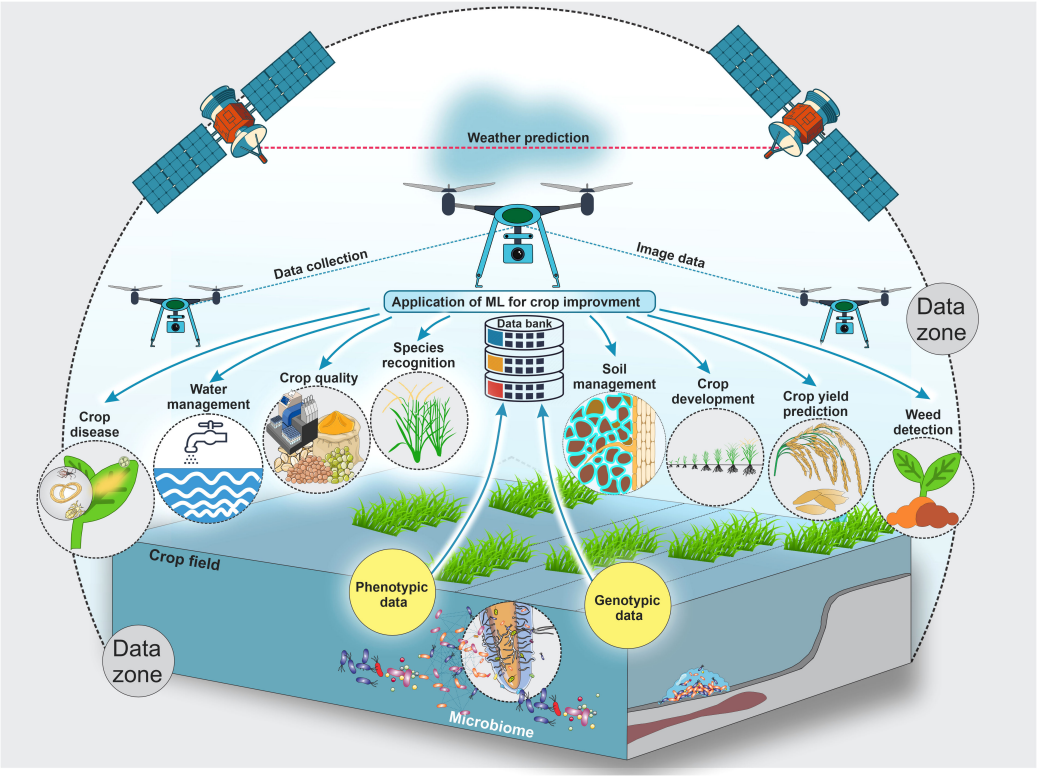


FIGURE 1
This schematic illustrates key applications of artificial intelligence and machine learning for crop development and improvement, including crop diseases, crop quality, crop species recognition, crop development, crop yield prediction, crop-related microbiome improvement, water management, soil management, etc. Farmers and researchers still encounter numerous obstacles due to employing traditional methods in the crop sector. Artificial intelligence and machine learning are used extensively to address these issues. Also, this figure shows possible data types and collection zones from crop fields to feed different machine-learning models to improve and develop different crops.

3.2 Concept of supervised, unsupervised, semi-supervised, and reinforcement learning

Supervised machine learning describes how a model can be fitted to data or part of target data that distinct labels have received for which a ground truth attribute exists; this quality is often determined by experimentation, researchers, or data collectors. In contrast to knowledge derived from inference, ground truth is information verified via direct observation and measurement, thus known to be accurate or real (Kondermann, 2013). Among the examples are high-yield prediction (Panigrahi et al., 2023) and water quality prediction (Ahmed et al., 2019; Ghosh et al., 2023; Chatterjee et al., 2024) using supervised learning for crop improvement. Laboratory or experimental observations ultimately serve as the source of ground truth in both cases. Contrary to supervised learning, patterns in unlabeled data can be found using unsupervised-learning techniques (James et al., 2023). This approach does not require predetermined labels with ground truth information (Sindhu Meena and Suriya, 2020). For example, plant image data can be analyzed using an unsupervised machine-learning technique (Davis et al., 2020; Bah et al., 2023). Semi-supervised learning, in which a significant quantity of unlabeled data is paired with tiny quantities of labeled data, occasionally combines the two methodologies (Ouali et al., 2020; Ahfock and McLachlan, 2023); for example, weed distribution and density estimation (Liu et al., 2024). When obtaining tagged or labeled data is expensive, this can dramatically enhance performance. Another component of machine learning known as reinforcement learning (RL) teaches an agent how to behave and react in a given environment by having it carry out specific tasks and then watching the rewards or outcomes. This technique is already employed in different agricultural domains, such as crop yield prediction (Elavarasan and Vincent, 2020; Iniyan et al., 2023) and a completely autonomous precision agricultural aerial scouting technique (Zhang et al., 2020; Elango et al., 2024).

3.3 Concept of classification, clustering, and regression problems

In machine learning, a task is referred to as a classification challenge when it requires allocating data points to a collection of discrete classes such as varieties emitting high or low methane, and a classifier is any algorithm that carries out this kind of classification (Sen et al., 2020), such as cassava disease detection and classification (Bian and Priyadarshi, 2024). Contrary to classification, regression models produce a collection of values that are continuous

(Pardoe, 2020; Panigrahiet al., 2023), such as the prediction of yield before the harvest of very early potato cultivars by using a regression model (Piekutowska et al., 2021). Regression problems can frequently be reformulated as classification problems since continuous values can be discretized or thresholded (Greener et al., 2022). Typically based on some metric of data point similarity, in a target dataset, clustering algorithms are applied to predict and group similar data points (Ghosal et al., 2020). These techniques are unsupervised and do not necessitate labeling the instances inside a dataset. For example, according to images of soybean, clustering could predict seed weight (Duc et al., 2023).

3.4 Concept of classes and labels

When a classifier returns a discrete collection (set) of mutually exclusive values, such values are referred to as classes. These values are called labels when they do not have to be mutually exclusive. Typically, an encoding is used to represent classes and labels. One essential step in preparing data for machine-learning tasks is encoding categorical variables. It is essential to convert categorical data into a numerical format to make them compatible with machine-learning algorithms. Categorical data are not numerical values, such as categories or text. For example, a place variable with the values first, second and third or a color variable with the values; red, green, and blue is categorical data, which every value denotes a distinct category. There might be an inherent ordering or link between some categories. There is a natural ordering of values for the aforementioned place variable. There is a natural ordering of values for the aforementioned place variable. Due to the fact that the values can be ranked or ordered, this kind of categorical variable is known as an ordinal variable. There are several popular category encoding methods, each combining benefits and drawbacks such as label encoding, ordinal encoding, and one-hot encoding methods. One-hot encoding is one of these techniques, which is most frequently employed (Yu et al., 2022b). When there is no innate link or order among the categories, this encoding performs well with nominal categorical variables (Rodríguez et al., 2018). The distinctiveness of every category is maintained by one-hot encoding. It guarantees that no ordinal link between the categories is assumed by the method. Also, one-hot encoding eliminates the possibility of unintentionally adding biases based on the sequence of categories because each category is represented independently. But when working with categorical variables that have a large number of distinct categories, one-hot encoding can dramatically increase the dataset's dimensionality. This may result in the curse of dimensionality and have an adverse effect on the performance of the model. Ordinal encoding is used when the categorical feature is ordinal. Every distinct category value in ordinal encoding is given an integer value. For example, in the color categorical data, red is 1, green is 2, and blue is 3. Maintaining the order is crucial in this method and encoding should so take the sequence into account. Equal intervals between categories are assumed by ordinal encoding, yet this may not always be the case in real-world situations (Dahouda and Joe, 2021). Unlike one-hot encoding, ordinal encoding does not increase the dimensionality of the dataset and It saves space and processing time by substituting integers for categorical variables. Label encoding assigns a unique integer value to each category in a categorical feature. This is an easy-to-use strategy that can be helpful when the categories' order matters. However, because of the allocated integer values, it could create unintentional linkages between categories. For instance, label encoding could assign the values 0, 1, and 2 correspondingly if a categorical feature is small, medium, and large. This would suggest that "large" is twice as significant as "small", which is probably incorrect. Important point is that the type of categorical variable and the issue those researchers are trying to solve will determine which encoding techniques should be used.

3.5 Concept of cost or loss functions

Machine-learning models never produce perfect results; they always deviate from the ground truth or the real world (Ho and Wookey, 2019). Cost or loss functions are the mathematical functions that compute this deviation or, more broadly, the degree of disagreement between the actual and ideal outputs (Uma et al., 2021). Mean squared error loss for regression problems is one example, and, for classification-related problems, binary cross entropy (Nar et al., 2019). A mean squared error loss function calculates the average squared difference between the anticipated value and ground truth. Binary cross entropy is a binary classification problem that must divide observations into one of two labels according to specific criteria [such as healthy leaf and infected leaf (Sarkar et al., 2023)].

3.6 Concept of parameters and hyperparameters

In essence, models are mathematical functions that take a collection of imported features and return one or several features or values as an output. Models include adaptable and flexible parameters that can be adjusted throughout the training process to optimize the models' performance, allowing them to learn from training data (Yu and Zhu, 2020). In a simple regression model, for instance, each feature has a particular parameter that is being multiplied by the value of the feature; these are then integrated and combined to provide a forecast. Hyperparameters are tunable values that are not changed during training and are thus not regarded as a model component. But this nonetheless affects the performance and training model. The learning rate, which regulates the pace at which the model's parameters are changed during training, is a standard description of a hyperparameter. To simplify it, hyperparameters control a structure and training procedure of machine-learning models, and they might be the number of clusters in K-means clustering, the learning rate in a neural network, or the depth in a decision tree. Hyperparameters, in contrast to model parameters, need to be predefined and cannot be learned during training. A model's ability to perform well or poorly can be determined by selecting the appropriate collection of hyperparameters. Therefore, choosing the set of hyperparameters

that result in the best possible model performance is known as hyperparameter tuning. Depending on the type of model being trained, different sorts of hyperparameters may be employed including learning rate, number of epochs, batch size, number of hidden layers and units, regularization parameters, momentum, and activation function. Several tools are developed for model tuning and hyperparameter optimization such as Ray Tune (Shin et al., 2020), Optuna (Akiba et al., 2019), HyperOpt (Bergstra et al., 2015), and AWS Sage Maker (Das et al., 2020).

3.7 Splitting target data into training, validation, and testing sets

Models need to be trained, which is the process of automatically modifying model parameters to enhance performance before they can be used to generate predictions (Mathai et al., 2020) This means altering the parameters in a supervised learning setting to minimize the average value of the loss or cost function and improves model performance with a training dataset. Typically, a separate validation dataset tracks but does not alter the training process to detect any overfitting (Twomey and Smith, 1997). Even if a cost function does not run on ground truth outputs in unsupervised scenarios, it is nonetheless decreased. After training, the model can be evaluated using data not used during training (Figure 2A) (Eelbode et al., 2021). For a general overview of the training procedure and instructions on how to divide the target dataset into training set and testing set. Figure 2 illustrates the principal notions for the training of models and displays a flowchart to aid in the whole procedure.

3.8 Concept of overfitting and underfitting

For a model to be predictive of unobserved (non-training) data, it must be fitted to training data to grasp the entire connection among all possible variables inside the dataset. The common reasons that a machine-learning model performs poorly are challenges, two key concepts in the field of machine learning (Figure 2B). An overfitted model (often caused by having too

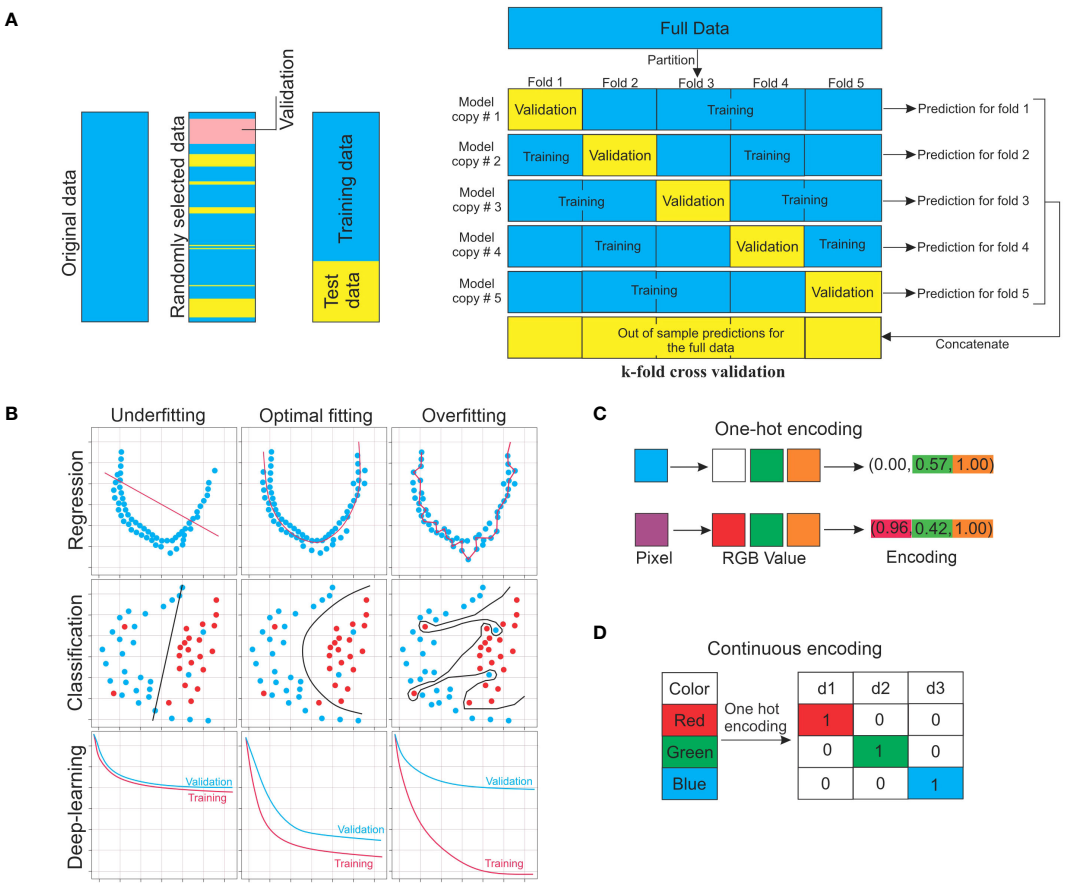


FIGURE 2
ML approaches for training. **(A)** Target data for machine learning should be split into training, validation, and testing sets. The training set is used to train the model directly. With the validation set, the training set is monitored. Test data are used to assess the performance of the model. The *k*-fold cross-validation approach is also used for validation. **(B)** Concept of underfitting and overfitting. **(C)** One-hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. **(D)** Numerical data can be represented in a way that machine-learning algorithms can understand using continuous encoding. In the given example, the RGB (R: red, G: green, and B: blue) are shown as the specific values of pixels in the targeted images.

many parameters) can generate outstanding output on trained data but will produce adverse outcomes on unobserved data. High variance and low bias might lead to overfitting. The training dataset will have zero prediction error since the overfitted model goes through each training point perfectly, as shown in Figure 2B. Conversely, an underfitted model cannot accurately represent the connections between the data variables. This can result from an improper model type selection, inaccurate or inadequate data assumptions, a high bias, or a low variance procedure (Figure 2B).

3.9 Concept of the bias-variance tradeoff

The inductive bias of the model is the collection of assumptions made by the learning algorithm that leads it to prefer one solution to a learning problem over another (Baxter, 2000). This could be understood as the model favoring one learning solution over another. This choice is frequently encoded into the model by using a specific loss function and/or its particular mathematical form. Various model types have distinct inductive biases that make them more appropriate and they often perform better for specific categories of data. The tradeoff between variance and bias is another crucial concept in machine learning. The general argument is that a model with a large bias places more restrictions on the trained model. Conversely, the low-bias model decreases the number of assumptions about the model property, and it is theoretically capable of modeling a large range of function kinds (Neal, 2019). The amount that the trained model varies when it is trained on various training datasets is indicated by the variance of the model. Ideally, models should have low variance and bias, but these goals frequently conflict with each other, given that a model with low bias will learn distinct signals on separate training sets. To prevent either overfitting or underfitting, it is essential to manage the bias-variance tradeoff.

4 Overview of ML procedures and required concepts

This section is a concise survey of the procedures that should be followed for training an ML model (Figure 3). Surprisingly, little advice is given for the selection of specific models and methods of training (Bengio, 2012; Greener et al., 2022). The first step is to understand the problem, the nature of the imported data, and the final goal of the prediction, which should come before writing any ML algorithms. This step is essential to have a comprehensive understanding of the crop improvement aspect of the problem or

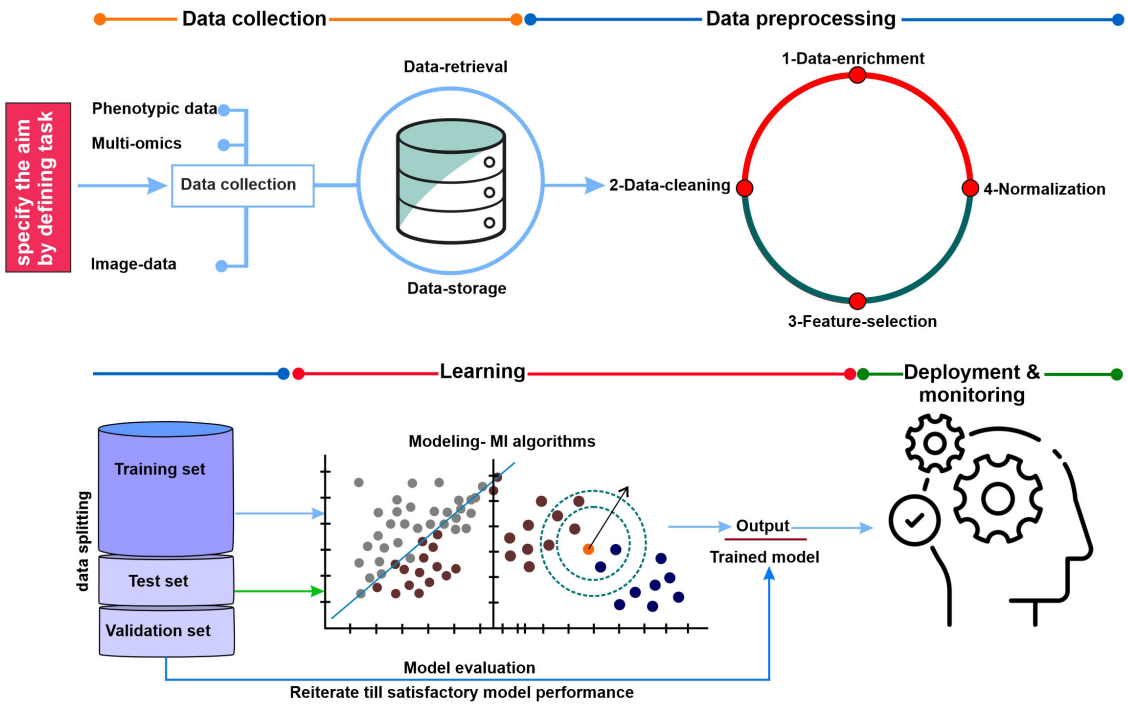


FIGURE 3
The graphic illustrates the general process for data collection, training, testing, and validation of machine-learning and model evaluation methods. Data collection: collecting data from various data sources related to crop improvement and development. Data processing: the most important step in the ML space is data cleaning and pre-processing. Before beginning the analysis of the best algorithm for the provided data set, we must comprehend the data set and select the cleaning or pre-processing procedures to obtain the best possible outcomes. Learning (Model building and selection of ML algorithm): This is a crucial stage that will soon bring the process to an end. Thus, we must choose our models carefully because we will select one as the ultimate model to solve the provided business challenge. Model evaluation: Model evaluation validates a target model with some standardized mathematical formulae or methodology. Deployment and monitoring: the selected ML model is authorized for deployment in the production environment after its performance satisfies the requirements, and the performance and behavior of the selected model in the real world are continuously observed, analyzed, and assessed through the process known as model monitoring.

question: for example, knowing the sources of noise and the origin of the target data. Understanding the computational storage of the inputs and outputs is also crucial. The following questions could be addressed: Are they adjusted (normalized) to avoid an excessively high effect of one attribute on prediction? Do they have continuous or binary encodings? Are some entries repeated? Are some data pieces missing (NaN)?

In the following step, the collected data (target dataset) must be divided into the first training dataset, the second for validation, and finally the testing dataset (Figure 2A). The training dataset is used for training an ML algorithm. In contrast, the test dataset (holdout set) is used to evaluate the resulting model (to estimate how well the model performs on unseen data). This idea is further used in the model selection part of the training procedure, which might lead to allocating a part of the training dataset as a validation set while the rest of the training data are used for training proper. Using the training set, the parameters of the specific model are updated during the training procedure. Usually, 10% of the available data are split and considered validation data to oversee instruction (training) performance, thus avoiding overfitting of the target model, and select hyperparameters (previously explained) based on datasets for training. Frequently, *k*-fold cross-validation is used in this step. Typically, 10% to 20% of the total dataset is dedicated as a test dataset to evaluate the expected real-world performance of the target model by assessing how well it performs on data that were not used for training or validation. To prevent adjusting the model to match the test set, there should be only one-time use of the test set in the later stages, if possible (Hastie et al., 2009; Bzdok et al., 2018). Selecting a model comes next, depending on the dataset type (nature of data) and the kind of anticipation being formed. This is conceptualized and made concise in Figure 3. To raise the overall accuracy of the undertaken model, the ensemble model averages the outputs of several comparable models that could be considered. Finally, evaluating the model's accuracy in the dedicated test dataset is crucial.

5 Conventional machine learning

This section investigates several essential and traditional machine-learning techniques, focusing on their advantages and disadvantages. Table 1 presents a comparison of several machine-learning techniques along with some applications for crop improvement and production. Figure 4 illustrates a few of the conventional machine-learning techniques. To train these models, several software programs have been available, such as Caret in R (Kuhn, 2008; Dege and Brüggemann, 2023), MLJ in Julia (Blaom et al., 2020), and scikit-learn in Python (Pedregosa et al., 2011; Rajamani and Iyer, 2023). When developing machine-learning algorithms for crop improvement-related data, conventional machine learning is typically the first area to investigate to find the most appropriate solution for a given problem. Deep learning is currently prevalent and has the potential to be a robust and valuable method. It is still restricted to the application domains where it performs well, though, such as when a vast quantity of data are accessible, such as extreme data points, when there are several features on each data point or when the features have a lot of structure (Greener et al., 2022). Drone images from crop fields (Killeen et al., 2024; Sahoo et al., 2024) and genotypic data (SNPs) (Uppu et al., 2016) are two examples of agricultural data for which deep learning could be effectively used. Even when the other two conditions are satisfied, deep learning may not be the best option because of the need for vast volumes of data. Technically, conventional approaches build and evaluate solutions for a particular problem far more quickly than deep learning. When compared to more conventional models such as random forests and support vector machines (SVMs) (Hastie et al., 2009), creating the architecture and training a deep neural network might be a computation-intensive and costly process (Sejnowski, 2018). For a given agricultural prediction problem, even if deep learning seems theoretically doable, it is usually wise to train a conventional technique and evaluate it against a model based on neural networks such as ANN (artificial neural network), if at all possible (Smith et al., 2020). Conventional approaches usually assume that every sample in the collection has the same number of characteristics, which is not always feasible. Using SNP data with varying lengths for each case is a clear illustration of this problem. The data can be adjusted using basic techniques such as windowing and padding to make them all the same size and employing standard ways with them. Padding refers to the process that can add zero value to each example up to making the size of each of them equal to the most prominent example in the target dataset. Conversely, the windowing approach condenses each sample to a specific size (Chrysostomou et al., 2011).

5.1 Application of regression and classification models

Regarding regression problems such as those depicted in Figure 4A, ridge regression (a type of linear regression) is frequently a valuable place to start when building and developing a model since it could offer a quick and clear baseline for a particular responsibility. The value of one variable can be predicted by using linear regression analysis according to the value of another variable (Su et al., 2012). On the other hand, when a model relies on as few features as possible from the given data, then other variations of linear regression, such as elastic net regression (Zou and Hastie, 2005) and LASSO regression (Tibshirani, 1996), are also worthy of consideration. Since the correlations between the characteristics in the data are frequently non-linear, using a model such as an SVM is usually a better option in these situations (Noble, 2006), as shown in Figure 4B. SVMs are a practical kind of classification and regression model that convert non-separable problems into easier-to-solve separable problems by using kernel functions. A kernel function is a technique for transforming input data into the format needed for data processing. Both non-linear (a statistical method called non-linear regression is used to model non-linear relationships between independent and dependent variables) and linear regression could be carried out with SVMs based on the kernel function that was applied (Ben-Hur et al., 2008; Ben-Hur and Weston, 2010; Kircher et al., 2014). To quantify, the best idea is to train an SVM through a kernel of a radial basis function and a linear SVM can be used from a non-

TABLE 1 Comparison of different machine-learning methods.

Method	Data types	Advantage	Disadvantage	Agricultural example
Support vector machine (SVM)	- Supervised learning (labeled) - Definite number of features	Capable of doing regression and classification both linearly and non-linearly	Large dataset scaling is frequently challenging	- Land suitability - Crop yield prediction (Lingwal et al., 2024) - Classification of weeds and crops based on digital images (Ahmed et al., 2012; Agarwal, 2024)
Ridge regression	- Supervised learning (labeled) - Definite number of features	- Prevent overfitting - Simple to train - A decent reference point (benchmark)	- Unable to understand the sophisticated relationship between features - Having overfits with an excessive number of features	- Genotype-specific grain yields of wheat (Herrera et al., 2018) - Predicting soil nutrition (Sudha et al., 2022)
LASSO regression	- Supervised learning (labeled) - Definite number of features	-Prevent overfitting -Remove highly inter-correlated features in data	- Chooses just one feature from a set of related features - Certain features might have significant bias	- Forecasting crop yield (Kashyap et al., 2024) - Wheat yield prediction (Shafiee et al., 2021)
Random forest	- Supervised learning (labeled) - Definite number of features	- Effective with big datasets - Discover how crucial each feature is to the forecast - More accessible for training and adjusting since it is less susceptible to feature normalization and scaling	- Not as suitable for regression - Interpreting several decision trees might be challenging.	- Crop yield predictions (Jeong et al., 2016; Basha et al., 2020; Dhillon et al., 2023)
Gradient boosting (such as XGBoost)	- Supervised learning (labeled) - Definite number of features	- Discover how crucial each feature is to the forecast - Easier to train and adjust since it is less susceptible to feature normalization and scaling	- Not as suitable for regression - Might find it difficult to learn information when there is noise	- Yield estimation (Huber et al., 2022) - Maize variable-rate seeding decision (Du et al., 2022)
Clustering	- Unsupervised learning (unlabeled) - Definite number of features	- Performance could be evaluated using accessible cluster validation metrics - Good clustering for low-dimensional data is readily observable	- Results from noisy datasets could occasionally be contradicting - Certain techniques have trouble scaling to huge datasets	- Crop yield predictions (Vani and Rathi, 2023) - Better energy use in crop production (Khoshnevisan et al., 2015; Wu et al., 2024)
Reduction of dimensions	- Unsupervised learning (unlabeled) - Definite number of features	- Gives clear ideas through visualization of datasets - Evaluations of goodness-of-fit are often provided to evaluate performance	- For specific techniques, scaling to vast numbers of samples is challenging - Preserving both local and global data differences is challenging	Dimensional reduction from genotypic data (SNPs) (Heffner et al., 2009; Evamoni et al., 2023)
Multi-layer perceptron	- Supervised learning (labeled) - Definite number of features	- Applies to intricate non-linear issues - Performs well with considerable data input - Quickly makes predictions following training - Even with fewer data, the same accuracy ratio can be attained	- The degree to which the dependent variable impacts each independent variable is unknown - Completing computation takes a lot of effort and time - Training data quality is critical to the correct operation of the model	- Predicting maize yield (Ahmed, 2023) - Predicting soil electrical conductivity (Mosavi et al., 2021)
Convolutional neural network (CNN)	- Grid-based spatial data arrangement	- High precision - Specifically made to handle image datasets - Able to derive spatial characteristics from a hierarchical matter	- Hefty computational expenses - Needs a huge dataset - Huge parameter size makes it challenging to optimize	- Crop classification (Mazzia et al., 2019; Kavitha et al., 2024) - Crop yield prediction (Nevavuori et al., 2019; Kolipaka and Namburu, 2024)
Recurrent neural network (RNN)	Data in sequential format (genotype data or time series)	- Capable of handling input of any length - For lengthier input, the model size would not increase - Sequence data format is seen in many agricultural domains	- Recurrent processing is time-consuming - High memory needs for computing	- Crop improvement (Gopi and Karthikeyan, 2024) - Crop yield prediction (Gopi and Karthikeyan, 2024)
Graph convolutional network	Connections and relationships between entities define the data	- Observes graph connection to identify patterns, allowing the predictor to use the most pertinent links	- More complex designs are challenging to train - High memory needs for computing	- Weed and crop recognition (Jiang et al., 2020; Pandey et al., 2024) - Crop recommendation systems (Ayesha Barvin and Sampradeepraj, 2023)

(Continued)

TABLE 1 Continued

Method	Data types	Advantage	Disadvantage	Agricultural example
Autoencoders	Supervised and unsupervised data (labeled and unlabeled data format)	- Noise identification ability - Effective in extracting features	- Restricted ability - The challenge of interpreting the outcome - Other datasets might not benefit from using latent space unique to the training set's data - Uses more memory resource	- Plant disease detection (Boukhris et al., 2024) -Crop classification (Guo et al., 2020; Cui et al., 2023)

linear model, if any. Numerous models that are often employed in regression could be used in classification as well. Another acceptable default starting point for a classification problem is to train an SVM based on the kernel function and a linear SVM. *k*-nearest neighbors classification (also known as k-NN or KNN) is a further technique that could be used (Bzdok et al., 2018). A non-parametric supervised-learning classifier, the *k*-nearest neighbors method employs closeness to classify or anticipate how a single data point will be grouped (Peterson, 2009). XGBoost (Figure 4C) (Chen and Guestrin, 2016; Olson et al., 2018) and random forests (Wang and Zhang, 2017) are examples of ensemble-based models, which provide another family of resilient non-linear techniques. These techniques are effective non-linear models offering feature significance estimations and frequently just need minor adjustments to the hyperparameters. There are often an overwhelming number of variations among the several models available for regression and classification. It can be misleading to try to forecast how well-suited a specific method will be to a given issue in advance; instead, it is usually wiser to use an empirical approach to identify the optimum model via trial-and-error methods. Swapping out these model versions often involves only one line of code change thanks to a novel and robust machine-learning library such as scikit-learn (Pedregosa et al., 2011), which can efficiently run in a Python environment. To find the best approach overall, it is an excellent strategy to optimize and train several of the previously described techniques, and then compare the results on a different test set to see which method performed the best on the validation set.

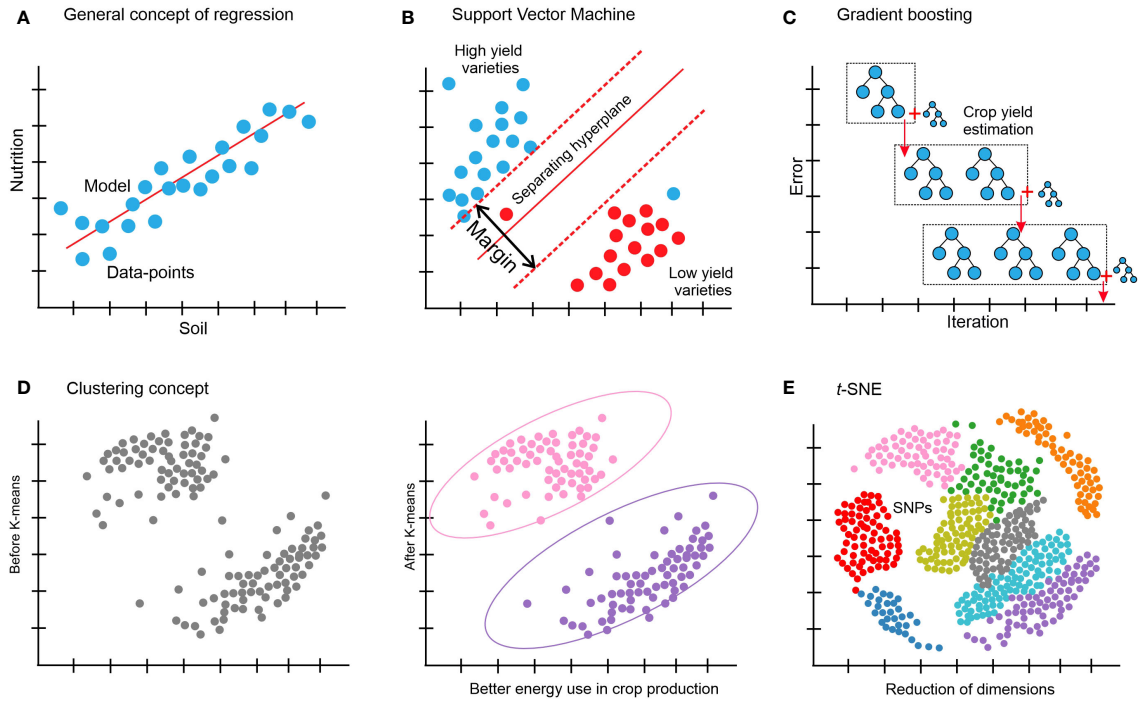


FIGURE 4
(A) Regression is the link between a single and/or several independent variables, also known as features, and a dependent variable (the observable attribute) is determined by using regression. A straightforward example is the prediction of crop yield based on one or some of the phenotypic features. **(B)** SVM: a support vector machine divides the original input data into several categories by creating a gap as large as feasible between the data in each converted version. One example is a prediction of whether a variety of a specific crop is a low- or high-yield variety. **(C)** Gradient boosting makes predictions by combining several weak prediction models, most often decision trees; for example, the prediction of sugarcane yield grade. **(D)** Clustering: using one of several algorithms, based on related objects; for example, better energy use in crop production. **(E)** t-distributed stochastic neighbor embedding (t-SNE), for example, dimensionality reduction of crop genotypic (SNP) data.

5.2 Application of clustering models

Like many other clustering algorithms (Figure 4D), k-means is a powerful multi-purpose clustering technique that requires the number of clusters to be specified as a hyperparameter (Jain, 2010). An alternate method that is not necessary for a predetermined number of clusters is DBSCAN (Ester et al., 1996). For datasets with plenty of features, dimensionality reduction can also be done prior to clustering to enhance performance.

5.3 Dimensionality reduction

High-dimensional data can be transformed into a lower-dimensional format while preserving the different connections and interactions between the data points and pieces using dimensionality reduction techniques. Although more dimensions could be used in machine learning, two or three dimensions are often selected to enable data visualization on several axes. These methods include data transformations that are both linear and non-linear. Principal component analysis (PCA) (Jolliffe and Cadima, 2016) and *t*-distributed stochastic neighbor embedding (*t*-SNE) (Van der Maaten and Hinton, 2008) are some of the examples common in the agriculture domain for dimensionality reduction. The circumstance determines which technique to apply. PCA is based on a linear combination of input features; each component preserves the global connections between the data points and could be explainable, implying that it is simple to identify the characteristics that contribute to data diversity. *t*-SNE is a versatile technique that can uncover structure in complicated datasets and more robustly maintain local links between data points (Figure 4E).

6 Concept of artificial neural networks

The mathematical principle of artificial neural networks (ANN) has been conceptualized by following and understanding the behaviors and connectivity of human neurons in the human brain. It was created initially to study the workings of the brain (Crick, 1989). The significant advances in deep neural network training and architecture over the past few decades have increased interest in neural network models (LeCun et al., 2015). The following section covers the fundamentals of neural networks and common varieties used in research on crop improvement. Figure 5 displays some of these concepts.

6.1 Concept of neural network fundamentals

The capacity of neural networks to approximate functions universally is one of their primary characteristics; this implies that, with minimal presumptions, any mathematical function can be accurately approximated to any degree by a neural network that is set up appropriately. The fundamental units of every neural network model are artificial neurons. A mathematical function that translates (converts) inputs to outputs in a certain way constitutes an artificial neuron (Wu and Feng, 2018). Any number of input values can be fed into a single artificial neuron, which then uses a predetermined mathematical function to produce an output value. Artificial neurons are layered and the output of one layer is the input of the next, which forms a network. In the following subsections, we present several methods for configuring artificial neurons, sometimes called neural network architectures. Combining several architectural styles is also popular. For instance, fully linked layers are typically used to provide the final classification output in a CNN (convolutional neural network) used for classification.

6.2 Concept of multi-layer perceptrons

A feed-forward ANN (artificial neural network) having several layers, comprising an input layer, one or several hidden layers, and an output layer, is called a multi-layer perceptron (MLP) (Figure 5A). Every layer is wholly interconnected with every other layer. The term "perceptron" was initially established by Frank Rosenblatt (Seising, 2018). The fundamental building block of an artificial neural network, the perceptron, specifies the artificial neuron inside the network. Activation functions, node values, inputs, and weights are all used in this supervised learning technique to determine the output. The forward direction is supported by the MLP neural network. Every node has complete network connectivity. Only in the forward direction does each node transmit its value to the next node. Back-propagation is a method used by the MLP neural network to back propagate the error in order to optimize the weights and unit values.

6.3 Concept of convolutional neural networks

CNNs are developed mainly to use image data format, and the fundamental component of a CNN is the convolutional layer (Figure 5B). Three things are needed: a feature map, a filter, and input data (Li et al., 2021). Suppose the input will consist of a color picture, a 3D matrix of pixels. As a result, the input will have three dimensions: height, width, and depth, which match the RGB color space of a picture. CNNs are equipped with a feature detector, which could also be called a kernel or filter. This detector traverses the receptive fields of the image and determines (Bouvrie, 2006). A convolution is the name given to this procedure. CNNs can be set up (configured) to function well with various spatially structured datasets. A 1D CNN, for instance, would contain filters that move in only one way. Data with one spatial dimension would be a perfect fit for this kind of CNN (Tang et al., 2020), such as genotypic (SNP) data from rice varieties. Digital images are examples of data with two spatial dimensions that 2D CNNs can process (Hara et al., 2018). Volumetric data, such as multi-temporal remote-sensing images, are what 3D CNNs use to function (Ji et al., 2018). Significant progress has been made in crop improvement for

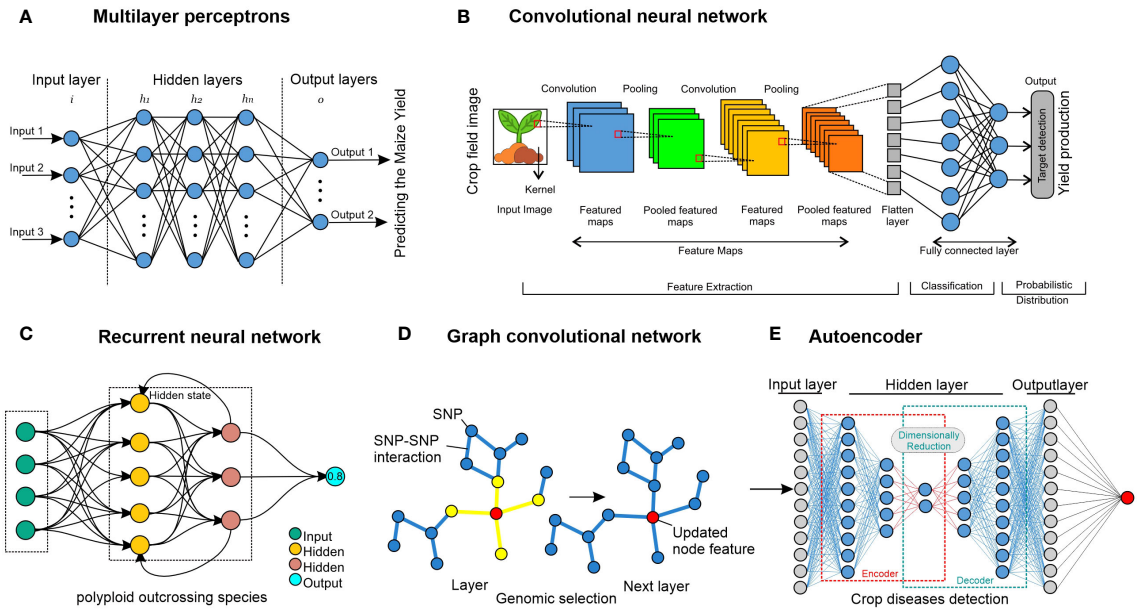


FIGURE 5
Neural network approaches. **(A)** MLP comprise nodes, which are represented by circles and can be either internal (hidden) values or output values. Layers of nodes are created by connecting each node in one layer to every node of other layers, signifying that the links represent learned parameters. For instance, the maize crop has been used for predicting yield (Ahmed, 2023). **(B)** To compute the values in the subsequent layer, a CNN employs filters that traverse the input layer. Since the filters work throughout the layer, parameters are shared, making it possible to identify related things wherever they may be. Although 1D and 3D CNNs are also used in crop improvement, 2D CNNs demonstrated operating on images of crops. 1D and 3D CNNs have been used for crop and crop-land classification (Ji et al., 2018; Liao et al., 2020; Liu et al., 2023). **(C)** An RNN is a deep-learning model trained to interpret and translate a given set of sequential data inputs into a predetermined set of sequential data outputs. It is used for weed detection for crop improvement (Brahim et al., 2021). **(D)** A GNN uses data from linked nodes of graph format data. It can be used for genomic selection in crops, and it has been used for predicting crop variety yield (Yang et al., 2023). **(E)** An autoencoder is composed of an encoder neural network that transforms an input into a latent representation with lower dimensions, and this hidden representation is transformed back into the original input by using a decoder neural network. This method has been used in crop disease detection for crop improvement (Abinaya and Devi, 2022).

various datasets using CNNs (Jiang and Li, 2020). Crop classification (Durrani et al., 2023), crop yield prediction (Nejad et al., 2022), and maize seedling recognition (Diao et al., 2022; Wei et al., 2024) are some examples of CNN models for crop improvement, and they now frequently surpass skilled human performance.

6.4 Concept of recurrent neural networks

RNNs are the most suitable approach with data organized into sequences, where each point in the series has some semblance of dependence or connection with the previous one (at least conceptually) (Greener et al., 2022), as seen in Figure 5C. The primary use of this approach is probably in NLP (natural language processing), which considers text a succession of characters (Medsker and Jain, 2001). One kind of RNN that can retain the outputs of each node for extended periods is called long short-term memory (LSTM) (Goodfellow et al., 2016). In other words, RNNs are modified to build LSTM networks, which provide better recall of previously learned data. Using back-propagation, they train the target model. When dealing with time delays of undetermined length, LSTM is a robust tool for classifying, processing, and predicting time series. Thus, once data are presented in an orderly structure, such as time sentences, LSTM can frequently be employed in different fields such as NLP and time-series analysis (Abdel-Nasser and Mahmoud, 2019). In the crop domain, RNNs are used extensively for crop improvement, such as land cover classification (Sun et al., 2019), prediction of crop biomass (Masjedi et al., 2019), and land cover and crop classification (Mazzia et al., 2019; Abidi et al., 2023; Moharram and Sundaram, 2023).

6.5 Principle of graph neural networks

Graph neural networks (GNNs) are especially well-suited for data that lack a clear apparent structure, such as a picture. Still, they are made up of things connected by randomly determined interactions or relationships (Battaglia et al., 2018). Such applications relevant to crop improvement include weed and crop recognition in smart farming (Jiang et al., 2020; Pandey et al., 2024) and crop recommendation systems (Ayesha Barvin and Sampradeepraj, 2023; Ge et al., 2024). In computer language, a graph is merely a representation of this kind of data, and every graph has a collection of nodes or vertices and a collection of edges that show different types of relationships or connections between the nodes. As seen in Figure 5D, when each feature of the nodes is updated across the network, neighboring nodes are considered. The node features in the final layer are then used as the output or merged to generate an output for the entire graph. Graphs

illustrating various correlations could use data from several sources to make predictions. Graph Nets (Gao and Ji, 2019) and PyTorch Geometric (Fey and Lenssen, 2019) are some of the most popular programs used to train GNNs.

6.6 Autoencoder networks

Autoencoder is used for unsupervised learning or the efficient coding of unlabeled input (Bank et al., 2023). The autoencoder method can learn two tasks: transforming input data by using an encoding function and recreating the input data from the encoded representation by a decoding function (Figure 5E). An alternative perspective is that the encoder attempts to compress the input and the decoder attempts to decompress it. Concurrent training is applied to the encoder, latent representation, and decoder (Doersch, 2016). Predicting the imposing of a structure on the latent space and the degree of similarity between two data points helpful for prediction tasks are two examples of applications. This approach has been used in several domains of crop improvement, such as crop classification (Bhosle and Musande, 2022; Cui et al., 2023) and crop mapping (Hamidi et al., 2021; Madala and Prasad, 2023; Hamidi et al., 2024).

6.7 Neural network improving and training

Several issues are unique to neural networks as they are far more sophisticated than conventional machine-learning techniques. It is frequently a good idea to train a neural network on a single training sample after deciding that it is the best model for the desired application for instance, a single image. The trained model is not helpful in forecasting, whereas it is adequate for exposing programming flaws (errors). As the network retains only the input, the training loss function ought to rapidly approach zero. If not, either the algorithm is not sophisticated enough to represent the input data or there is probably a mistake in the code. The network can begin with training on the whole training set after passing this fundamental debugging test when there is a minimum in the training loss function. It might be necessary to adjust hyperparameters such as the learning rate for this, as shown in Figure 6A. Overfitting of the network can be identified by tracking loss on the training dataset and validation dataset, where loss on the training set starts to rise and loss on the validation set keeps becoming less. At that moment, training is often discontinued, a procedure called early stopping, as shown in Figure 6B. A neural network overfitting indicates that the model's capacity to generalize to new data is beginning to wane as it starts to memorize only the features of the training set. Although early stopping is an intelligent strategy for avoiding this, other training approaches could be employed, such as dropout methods or model regularization. Nodes within the network are arbitrarily disregarded to compel the network to discover a more reliable prediction method incorporating more nodes. TensorFlow (Abadi et al., 2016) and PyTorch (Paszke et al., 2019) are well-liked neural network training programs. Neural network training is computationally intensive and often calls for a tensor processing unit or graphics processing unit with enough RAM (random-access memory) because using these devices could accelerate work 10 to 100 times faster than using a regular CPU (central processing unit). This acceleration is necessary for training massive datasets and for the larger models that have demonstrated success in recent years. Nevertheless, using a model that has already been trained is typically much quicker and this could frequently be accomplished with a simple CPU. For researchers without access to a GPU (a graphics processing unit is on-demand computing services) for training, cloud computing options are available from popular suppliers, and thus it is essential to remember that for simple tasks. Python code could be freely tested on graphics or tensor processing units using Colaboratory (Colab). A practical method to get started with deep learning based on Python is to use the Colab environment.

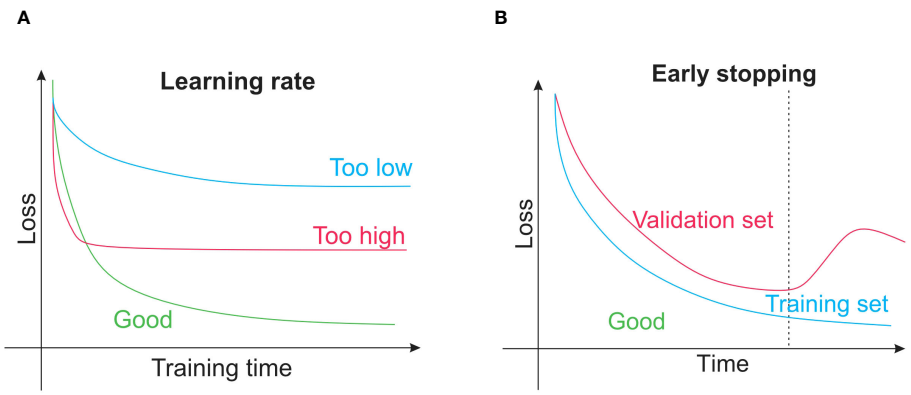


FIGURE 6
(A) The learning rate concept is that, when training a neural network or other conventional techniques such as gradient boosting, the learning rate of the model controls how quickly parameters that are learned are changed. **(B)** Early stopping is a regularization technique that helps prevent overfitting when training learners using gradient descent or other iterative methods.

7 Challenges of machine learning for crop improvement and production-related data

The enormous diversity of agricultural and crop domain data is one of the significant challenges in modeling, and these data are also generated from different nature domains. Crop improvement-related data can be yield-related data, land-related data, crop-development-related data, crop-disease-related data, or even microorganism data. Most of them can be along with genotypic and transcriptomic data such as SNPs or RNA-seq data and/or high-resolution images, 3D structures, or gene expression profiles over time, and different interactions of networks are some examples of these data formats and natures. A summary of recommended techniques and crucial factors for several crop-improvement data kinds is provided in Table 2. Because of the variety of data formats encountered, processing crop-improvement-related data frequently calls for customized solutions. Because of this, it is challenging to provide ready-made solutions or even broad suggestions for applying machine learning in various fields of study. However, for machine learning to be used successfully in crop improvement and agriculture, as well as more broadly, a few common challenges must be considered.

7.1 Availability of high-quality data

Since data quality directly affects the functionality, precision, and dependability of ML models, it is essential to the field of artificial intelligence. Models that use high-quality data are more predictive and yield more consistent results. There are some main challenges for insuring data quality in ML including Data collection; the problem facing crop research institutes is obtaining high-quality data from a variety of sources. Ensuring that every data point followed to the same criteria for collecting data and getting rid of redundant or contradicting data is difficult. Data labeling; for training purposes, machine learning algorithms require labeled data; yet, manual labeling is error-prone and time-consuming. Accurate labels that accurately represent real-world conditions are the difficult part. There

TABLE 2 Suggestive strategies for applying machine-learning techniques to varied datasets related to crop improvement.

Input data format	Recent instances of prediction tasks	Suggested models	Challenges for implementing
Images	- Crop disease monitoring (Bouguettaya et al., 2023; Zhang et al., 2024) - Crop protection (Gauriau et al., 2024) - Yield prediction (Zanella et al., 2024) - Stress detection (Butte et al., 2021; Gholap et al., 2024) - Crop growth (Memon et al., 2021; Attri et al., 2024) - Species detection (Picon et al., 2022) - Water management for crop improvement (Jain et al., 2021; Meenal et al., 2024)	- Autoencoders - 2D CNNs - Conventional techniques based on image features	- Difficult to have reliable dataset - Produces massive amount of data, which are difficult to maintain - Prediction could be affected by systematic variations in data collection - Expensive to provide the dataset - Data collection is an expensive process
Phenotypic data	- Yield prediction (Cao et al., 2021; Dhaliwal and Williams, 2024) - Crop productivity (Mochida et al., 2019) - Species recognition (Chen et al., 2023; Rangarajan et al., 2023) - Crop seed germination (Colmer et al., 2020; Duc et al., 2023)	- SVM - KNN - ANN/SNKs - 1D CNNs - K-means clustering - Conventional machine-learning models - Deep feed-forward multi-layer perceptron	- Lack of access to reliable datasets - Lack of uniform protocol for data collection - High noise
Geographic and climatic data	- Crop production (Alif et al., 2018; Dhillon et al., 2024) - Forecasting crop yield (Veenadhari et al., 2014; Kheir et al., 2024) - Crop yield change projections (Li et al., 2023) - Crop modeling with machine learning (Zhang et al., 2021b; Mousavi et al., 2024) - Crop selection (Yesugade et al., 2018; Kamatchi and Muthukumaravel, 2024) - Crop evapotranspiration (Yamaç and Todorovic, 2020; Du et al., 2024)	- SVM - ANN - 1D CNNs - LSTM RNN	- Different performance of the trained model in unknown regions - High noise
Genotypic data	- Crop improvement (Tong and Nikoloski, 2021; Guo and Li, 2023) - Identifying true single nucleotide polymorphisms (Korani et al., 2019; Sehrawat et al., 2023) - Phenotype prediction (Danilevicz et al., 2022) - Uncovering QTL (Yoosefzadeh-Najafabadi et al., 2022) - Introducing new candidate genes for specific traits (Mora-Poblete et al., 2023)	- Autoencoders - 1D CNNs - SVM - ANN - CNN - GNN - Graph embedding	- Because datasets are dispersed and stored in different places, they are difficult to obtain - Data leaks might make validation challenging

are some tools and software to generate ground truth data for ML specifically for certain domain such as ROOSTER (image labeler and classifier) (Tang et al., 2023), Bounding boxes (Osman et al., 2021), Polygons (Li et al., 2012), and Polylines (Opach and Rød, 2018). Data security and storage; preserving the integrity of data also entails shielding it from potential corruption and unwanted access. Also, it is essential for agricultural research institutions to have reliable and secure data storage. Data governance; it is difficult for many research facilities to put in place data governance structures that adequately handle problems with data quality. Errors, inconsistent data, and segregated data can result from improper data governance. Also, there is a need for more open-access datasets and standardized data collection protocols to facilitate ML research It is essential to develop more reliable and accurate data collection methods to ensure high-quality data for ML research for crop development improvement.

7.2 Accessibility of data

Compared to other domains, agricultural and crop-related data have little publicly available data. The selection of strategies that could be applied successfully is significantly influenced by the amount of data available for a particular import data format. Technically, researchers are effectively compelled to employ more conventional machine-learning techniques when limited quantities of data are available because the accuracy of these approaches is more reliable in these particular cases. Deep neural networks and other highly specified models can be explored once more significant quantities of data are available. For supervised machine-learning approaches, it is essential to take into account the relative quantities of every ground truth label included in the dataset. If some labels are insufficient, more data will be needed for machine learning to function (Wei and Dunbrack, 2013; Alzubaidi et al., 2023).

7.3 Model interpretability

Researchers often aim to determine why a particular model predicts some subjects in a certain way and why this particular model works in certain situations and is not accurate in other conditions. Putting it in another way, rather than focusing just on correct modeling, agri-researchers are typically interested in identifying the mechanisms and causes accountable for modeling output. The machine-learning technique and the input data determine how well a model can be interpreted. Non-neural network approaches typically contain fewer learnable parameters and feature sets that are more accessible to meaningful interpretation, making interpretation easier. For example, in a simple linear regression model, the parameter allotted to every input feature indicates how that variable influences the prediction. Because non-neural network approaches are inexpensive to train, ablation research in which the impact of eliminating certain input features on performance is quantified is recommended. One approach to potentially finding more reliable, effective, and understandable models is through ablation experiments, which can highlight which aspects are most helpful for a particular modeling job. Because a neural network often has many input parameters and features, interpreting one is often significantly more difficult.

7.4 Challenges in transdisciplinary partnerships

The main concern for data-driven crop improvement and production programs is standardized data collection protocol to prevent noisy data and the availability of high-resolution data. On the other hand, it is uncommon for one research organization to be aware of specific resources and knowledge to collect data in machine-learning research and adequately employ the most suitable machine-learning algorithms unless publicly available data are being used. Computer scientists and experimental agri-institutes frequently collaborate, and the outcomes of these collaborations are often outstanding. However, in these kinds of partnerships, each party must understand the other. In particular, agri-institutes and researchers should be aware of the constraints of the machine-learning algorithms being applied, and computer scientists should understand thoroughly the nature of the data, including the anticipated repeatability and level of noise. Developing such awareness takes time and work, but it is crucial for halting the frequent accidental spread of below-standard models and false conclusions.

8 Federated learning and gossip learning as recommendation approaches for global crop improvement and production programs

When leveraging datasets from many institutions, the model could be trained centrally, combining data from silos of various institutions onto a single server. However, different legal, ethical, and administrative restrictions exist on publicly exchanging crop-based data. In many countries, crop-based data must remain in the group, company, or institution. Machine-learning models are trained using a decentralized method called federated learning, often called collaborative learning. Federated learning (FL) is an approach for building machine-learning models where distributed data are used cooperatively by a central server (McMahan et al., 2017; Kairouz et al., 2021), as illustrated in Figure 7. FL allows the data to remain at the original site to protect the safety and intellectual privacy of data, in contrast to centralized training, which transfers data from produced locations to a central server to train the model. Once a new training cycle begins, the most recent version of the model is transmitted to every storage site where the training data are stored (Greener et al., 2022). Each copy of the model is then trained and updated using the data that belong to each unique site. The revised models are then returned to the central server from each site, where they are merged to create a universal model. After that, the freshly revised universal model is released for distribution once more, and the cycle continues until either the model training or convergence is completed. Only those

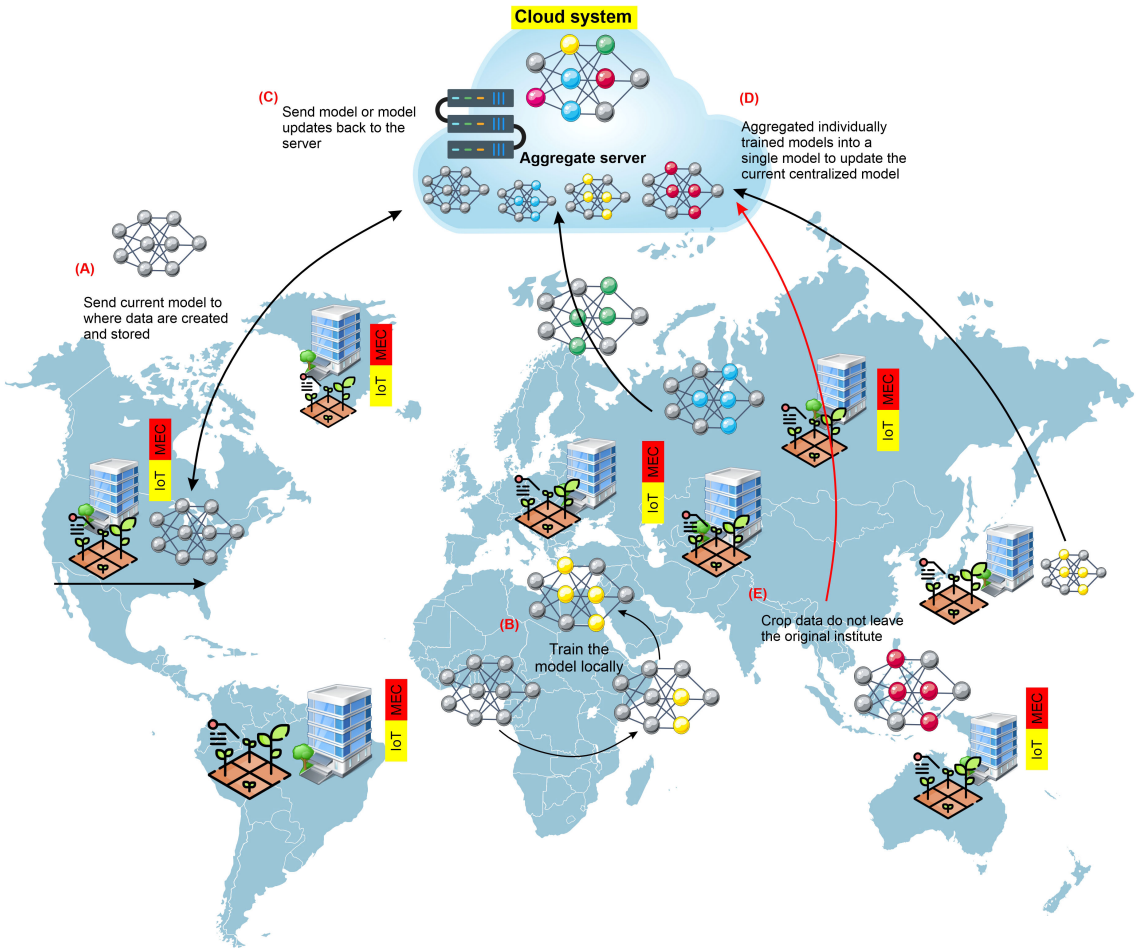


FIGURE 7
General pipeline constructing-silo for crop-improvement data to use in FL approaches. Several universities or institutes cooperatively train an ML model via federated learning (FL). In phase **(A)**, the central server provides the institution with the most recent model version. In phase **(B)**, each organization uses its data to train the model locally. In phase **(C)**, each institution transmits its trained model to the central server. In phase **(D)**, the central server combines all of the models that have been locally trained by the various universities into a single updated model. In phase **(E)**, each training cycle involves repeating this procedure till the training of the model is complete. Crop data never leave the institution during any of the training phases. Institutions need access to essential resources such as powerful hardware and specialists to conduct FL successfully.

people directly related to that institution have direct access to the data, which means that the data are never virtually transported from the originating location or institution. In an FL approach, the risks of data ownership violations are decreased, data aggregation costs are kept to a minimum, and training datasets can quickly increase in size and variety. Optimum use of the FL approach can lay the groundwork for training deep-learning models for universal crop-based data.

8.1 FL taxonomy

The data matrix is the foundation of FL (Li et al., 2022). FL is categorized into three groups according to the various distribution patterns of the sample space and feature space of the data: federated transfer learning (FTL), vertical FL (VFL), and horizontal FL (HFL), which partition datasets non-dimensionally, longitudinally (i.e., dimension of features), and horizontally (i.e., dimension of users), correspondingly, as shown in Figure 8.

8.2 General workflow for employing the FL approach

Data holders and central servers are the usual components of FL systems (Li et al., 2022). Not enough local data or feature counts from individual data holders may be available to enable effective model training. As a result, cooperation from other data owners is needed. The FL procedure for the architecture of the client-server is shown in Figure 7. To safeguard data privacy, the data holders exclusively train their data locally in a standard cooperative modeling procedure of FL. After desensitization, the gradients produced by the iterations are used as interaction information and sent to a trustworthy third-party server in place of local data and, to update the model, the server should return the aggregated parameters. The stages involved in FL can be summed up in detail below. The first step is system initialization. In this step, the central server sends out the modeling work and tries to engage with the client. Local calculation is the second step. Upon opening the joint

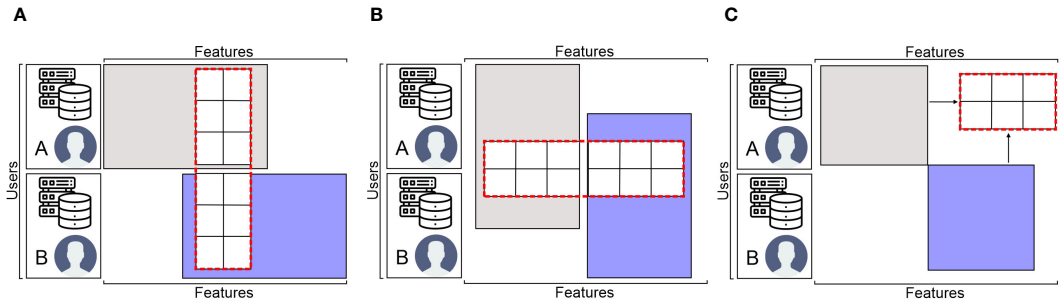


FIGURE 8
The FL data partition categories: **(A)** horizontal FL, **(B)** vertical FL, and **(C)** federated transfer learning.

modeling job and initializing the system settings, it will be necessary for each data owner (holder) to initially carry out local measurements and calculations based on the data locally. Eventually, the third step is central polymerization. The central server compiles the estimated values after obtaining the computation results from various data owners (holders). Security, privacy, efficiency, and other concerns are considered and checked during the aggregation process in this step. Significantly, the FL central server's functioning is comparable to that of a distributed machine-learning server, which gathers each data holder's gradient and then produces a new gradient via server aggregation processes.

8.3 FL applications in agriculture and relevant work in some crops

Since FL allows datasets to be analyzed even when the raw data are either not readily available or the data owners are not ready to share target data, this opens tremendous opportunities to use the mentioned approach in different domains. In the medical field, FL has been used to recognize COVID-19 disease during the pandemic through an image analysis approach from chest-computed tomography (Lai and Yan, 2022). According to their findings, their network's communication cost decreased by using the federated averaging model. Additionally, to lessen Byzantine assaults in their federated learning test bed, researchers suggest a modified federated learning model in which the edge nodes are randomly split into groups, each assigned a separate transmission time slot (Sifaou and Li, 2022). Because edge devices have a wide range of capabilities and resources, researchers have developed a federated learning framework that analyzes the models without jeopardizing data security or privacy while reaching convergence (Kevin et al., 2022). Agri-researchers, agri-institutes, and agri-companies also frequently gather private data and information that they prefer to keep private, as presented in Table 3. FL uses machine learning to train a shared model across several devices without requiring data exchange. It is perfect for agricultural applications. The FL applications in agriculture are categorized below to create a global model based on data-partitioning techniques, architecture, aggregation algorithms, and scale of federation. In one effort, researchers use a horizontally distributed dataset placed on several client devices to train the yield prediction model using FL (Manoj et al., 2022). To demonstrate the efficacy of agricultural data under decentralized learning, the FedAvg algorithm is used to build deep regression models such as ResNet-16. In another effort, to classify crops (chickpea, maize, and rice), the federated averaging approach has been employed (Idoje et al., 2023). Compared to the stochastic gradient descent (SGD) optimizer, the Adam optimizer model converged more quickly in this research. The study using the farm dataset has shown that decentralized models outperform centralized network models in terms of accuracy and convergence speed.

8.4 Federated learning challenges and limitation

Like other systems, FL also has some limitation and challenges for users which can categorized in four main groups include high-cost communication, heterogeneity of systems, heterogeneity in statistics, and privacy issues (Mammen, 2021; Moshawrab et al., 2023). The first challenges are raised in FL system is high-cost communication. Network communication in federated systems can be many orders of magnitude slower than local computing because these models consist of a large number computing devices. Compared to traditional data center facilities, communication in these networks can be substantially more expensive. It is also required to design communication-efficient approaches that iteratively send short messages or model updates as part of the training process, instead of sending the complete dataset over the network, in order to fit a model to data supplied by the devices in a federated network. The second challenge is heterogeneity of systems. Due to variations in hardware (memory, CPU), power, and network connectivity, each device in federated networks may have different computing, storage, and communication capabilities. Furthermore, only a small percentage of the devices are usually active at any given time due to the scale of the network and limits imposed by individual systems on each device. For instance, in a network with millions of devices, only hundreds of devices might be in use. It is also possible for any device to be unreliable, and it happens frequently for an active device to stop working during a particular cycle. Problems like stragglers and fault tolerance are far

TABLE 3 Agricultural applications of the federated learning method in some crops.

Target area in agriculture	Issue	Number of customers	Challenges	Data used	Aggregation approaches	Trained model	Ref
Smart farming and crop classification	Data security in intelligent farming	6	Usage of FL in intelligent agriculture	The dataset included rainfall, pH, humidity, and temperature of independent variables	Model of federated averaging	CNN	(Idoje et al., 2023)
Production from the agricultural sector	Directing the production of agriculture	10	Inexpensive transmission, quick convergence rate, and precise modeling with limited resources	Soybean iron deficiency chlorosis (IDC) photos from the real world	A greedy algorithm and suggested a collaborative FL framework for the Edge-IoAT (Internet of Agriculture Things) framework to identify the best course of action	GA (greedy algorithm)	(Yu et al., 2022a)
Detection of various pests and diseases	To prevent imbalanced and inadequate orchard data, expensive data storage and transmission, various pests and diseases, and challenging detection situations for typical cloud-based deep-learning solutions	6	Prevent the communication costs that arise from uploading a lot of data to address the problem of imbalanced and inadequate data	445 images of orchard apples, of which only 152 images include five diseases	FedAvg approach	Improved faster region convolutional neural network (R-CNN)	(Deng et al., 2022)
Using FL for amendable multi-function control method for smart sensors for enhanced agricultural production	Enhancing efficiency	47	FL is derived from sensor information	Soil and crop data	Amendable multi-function sensor control method (AMFSC)	AMFSC	(Abu-Khadrah et al., 2023)
Disease detection in food crops	Anticipating leaf diseases	4	Privacy of data	Data from plant-village	FedAvg approach	Five CNNs: ShuffleNet, SqueezeNet, AlexNet, VGG-11, and ResNet-18	(Antico et al., 2022)

more common due to these system-level features than they are in standard data center settings. The third challenge od FL system is heterogeneity in statistics in the system. Across the FL network, devices typically produce and gather data in non-identically dispersed ways. Furthermore, there may be a large variation in the quantity of data points amongst devices. And finally, the last challenge of FL system is privacy concerns. In contrast to learning in data centers, privacy is frequently a primary problem in FL systems. FL only shares model updates rather than raw data, which is a step in the right direction towards preserving user data. Sensitive and important information may still be revealed to the central server or a third party by sharing model changes during the training phase. Although there have been efforts recently to improve FL privacy through the use of techniques like differential privacy or secure multiparty computing, these strategies frequently sacrifice system efficiency or model performance in order to achieve privacy (Zhang et al., 2021a; Wen et al., 2023).

9 Gossip learning can be alternative to federated learning

To tackle the same issue, gossip learning has also been suggested as an alternative to federated learning (Ormándi et al., 2013; Hegedűs et al., 2016, 2019). There is no need for a parameter server because this method is completely decentralized. Nodes immediately share and combine models. Undoubtedly, there are seveal advantages to using gossip learning because there is no single

point of failure and gossip learning has far cheaper scalability and better resilience because no infrastructure is needed. The term "gossip" describes the information-sharing process that occurs over the network in a manner akin to that of gossip within a social group. In this approach, through information sharing with other nodes in the network, each node in the network updates its model parameters in this distributed machine learning technique. The theory is that any node can rapidly converge to the global optimum by exchanging information with other nodes. In large-scale distributed systems where node-to-node communication is unreliable or expensive, gossip learning is very helpful.

10 Conclusions and future direction

Future predictions display significantly greater use of AI and ML approaches in crop science, which could open a new horizon for integrated and valuable solutions in this area. We have undertaken a thorough review of the essential elements, concepts, applications, and machine-learning definitions required for agri-crop improvement. Nowadays, crop science is leveraging tons of available data to obtain deeper insights through AI and ML and offer the best suggestions for following actions and decisions for enhancing crop productivity or for other necessary tasks. Crop improvement and forecasting are made more accessible by combining computer science and agriculture. Offering broad recommendations and guidance for machine learning in agriculture is challenging because of the diversity of agricultural data. Therefore, our article aimed to provide agricultural and crop science researchers with an overview of the many accessible approaches, as well as some suggestions for conducting efficient machine learning through available data. It is vital to recognize that machine learning is inappropriate for all problems and to know when to avoid it: when the available data are insufficient, when it is necessary to comprehend rather than anticipate, or when it is not apparent how to fairly evaluate performance. Also, here we highlighted the application of federated learning in agriculture along with the definition, procedures, and structure, which can be beneficial for researchers in the agricultural sector. Even though there has been huge progress in machine learning in agriculture, many challenges still need to be addressed to mark ML territory in agricultural science. There is no denying that machine learning has influenced and will continue to influence agricultural research significantly.

Author contributions

SMHK: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. JA: Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was funded by the International Rice Research Institute (IRRI)-Hybrid Rice Development Consortium (HRDC) and the AGGRi Alliance project, "Accelerated Genetic Gains in Rice Alliance" by the Bill and Melinda Gates Foundation through the grant ID OPP1194925- INV 008226.

Acknowledgments

The authors thank members of the IRRI Rice Breeding Innovation platform (RBI) for valuable discussions and comments, especially Dr. Dmytro Chebotarov and Dr. Giovanny Covarrubias-Pazaran, along with Dr. Amir-Hossein Karimi from the University of Waterloo, Canada.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016) *{TensorFlow}: a system for {Large-Scale} machine learning, 12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 265–283.

Abdel-Nasser, M., and Mahmoud, K. (2019). Accurate photovoltaic power forecasting models using deep LSTM-RNN. *Neural computing Appl.* 31, 2727–2740. doi: 10.1007/s00521-017-3225-z

Abidi, A., Ienco, D., Abbes, A. B., and Farah, I. R. (2023). Combining 2D encoding and convolutional neural network to enhance land cover mapping from Satellite Image Time Series. *Eng. Appl. Artif. Intell.* 122, 106152. doi: 10.1016/j.engappai.2023.106152

Abinaya, S., and Devi, M. K. (2022). Enhancing crop productivity through autoencoder-based disease detection and context-aware remedy recommendation system. *Appl. Mach. Learn. Agric.* (Cambridge, MA, USA: Academic Press), 239–262. doi: 10.1016/B978-0-323-90550-3.00014-X

Abu-Khadrah, A., Ali, A. M., and Jarrah, M. (2023). An amendable multi-function control method using federated learning for smart sensors in agricultural production improvements. *ACM Trans. Sensor Networks*. doi: 10.1145/3582011

Agarwal, D. (2024). A machine learning framework for the identification of crops and weeds based on shape curvature and texture properties. *Int. J. Inf. Technol.* 16, 1261–1274. doi: 10.1007/s41870-023-01598-9

Aguilar-Zambrano, J., Mambuscay, C. A. A., and Jaramillo-Botero, A. (2023). Omics sciences in agriculture: crop phenomes and microbiomes *Sello Editorial Javeriano-Pontificia Universidad Javeriana, Cali.*

Ahfock, D., and McLachlan, G. J. (2023). Semi-supervised learning of classifiers from a statistical perspective: A brief review. *Econometrics Stat* 26, 124–138. doi: 10.1016/j.ecosta.2022.03.007

Ahmed, F., Al-Mamun, H. A., Bari, A. H., Hossain, E., and Kwan, P. (2012). Classification of crops and weeds from digital images: A support vector machine approach. *Crop Prot.* 40, 98–104. doi: 10.1016/j.cropro.2012.04.024

Ahmed, S. (2023). A software framework for predicting the maize yield using modified multi-layer perceptron. *Sustainability* 15, 3017. doi: 10.3390/su15043017

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., and García-Nieto, J. (2019). Efficient water quality prediction using supervised machine learning. *Water* 11, 2210. doi: 10.3390/w11112210

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. Anchorage. doi: 10.1145/3292500.3330701

Alif, A. A., Shukanya, I. F., and Afee, T. N. (2018). *Crop prediction based on geographical and climatic data using machine learning and deep learning* (BRAC University).

Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamaría, J., Albahri, A. S., Al-dabbagh, B. S. N., et al. (2023). A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *J. Big Data* 10, 46. doi: 10.1186/s40537-023-00727-2

Antico, T. M., Moreira, L. F. R., and Moreira, R. (2022). "Evaluating the potential of federated learning for maize leaf disease prediction," in *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional* (SBC). doi: 10.5753/eniac.2022

Attri, I., Awasthi, L. K., and Sharma, T. P. (2024). Machine learning in agriculture: a review of crop management applications. *Multimedia Tools Appl.* 83, 12875–12915. doi: 10.1007/s11042-023-16105-2

Ayesha Barvin, P., and Sampradeepraj, T. (2023). Crop recommendation systems based on soil and environmental factors using graph convolution neural network: A systematic literature review. *Eng. Proc.* 58, 97. doi: 10.3390/ecsa-10-16010

Bah, M. D., Hafiane, A., and Canals, R. (2023). Hierarchical graph representation for unsupervised crop row detection in images. *Expert Syst. Appl.* 216, 119478. doi: 10.1016/j.eswa.2022.119478

Bank, D., Koenigstein, N., and Giryes, R. (2023). *Autoencoders. Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook* (NY, USA: Springer New York).

Basha, S. M., Rajput, D. S., Janet, J., Somula, R. S., and Ram, S. (2020). Principles and practices of making agriculture sustainable: crop yield prediction using Random Forest. *Scalable Computing: Pract. Exp.* 21, 591–599. doi: 10.12694/scpe.v21i4.1714

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., et al. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint*. arXiv:1806.01261. doi: 10.48550/arXiv.1806.01261

Baxter, J. (2000). A model of inductive bias learning. *J. Artif. Intell. Res.* 12, 149–198. doi: 10.1613/jair.731

Bengio, Y. (2012). *Practical recommendations for gradient-based training of deep architectures,* Neural networks: Tricks of the trade: Second edition (Springer), 437–478.

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., and Rätsch, G. (2008). Support vector machines and kernels for computational biology. *PloS Comput. Biol.* 4, e1000173. doi: 10.1371/journal.pcbi.1000173

Ben-Hur, A., and Weston, J. (2010). "A user's guide to support vector machines," in *Data mining techniques for the life sciences*, 223–239.

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., and Cox, D. D. (2015). Hyperopt: a python library for model selection and hyperparameter optimization. *Comput. Sci. Discovery* 8, 014008. doi: 10.1088/1749-4699/8/1/014008

Bhosle, K., and Musande, V. (2022). Evaluation of CNN model by comparing with convolutional autoencoder and deep neural network for crop classification on hyperspectral imagery. *Geocarto Int.* 37, 813–827. doi: 10.1080/10106049.2020.1740950

Bian, K., and Priyadarshi, R. (2024). Machine learning optimization techniques: a Survey, classification, challenges, and Future Research Issues. *Arch. Comput. Methods Eng.*, 1–25. doi: 10.1007/s11831-024-10110-w

Blaom, A. D., Kiraly, F., Lienart, T., Simillides, Y., Arenas, D., and Vollmer, S. J. (2020). MLJ: A Julia package for composable machine learning. *arXiv preprint* arXiv:2007.12285. doi: 10.48550/arXiv.2007.12285

Bouguettaya, A., Zarzour, H., Kechida, A., and Taberkit, A. M. (2023). A survey on deep learning-based identification of plant and crop diseases from UAV-based aerial images. *Cluster Computing* 26, 1297–1317. doi: 10.1007/s10586-022-03627-x

Boukhris, A., Jilali, A., and Asri, H. (2024). Deep learning and machine learning based method for crop disease detection and identification using autoencoder and neural network. *Rev. d'Intelligence Artificielle* 38, 459–472. doi: 10.18280/ria

Bouvrie, J. (2006). *Notes on convolutional neural networks*.

Brahim, J., Loubna, R., and Noureddine, F. (2021). RNN-and CNN-based weed detection for crop improvement: An overview. *Foods Raw materials* 9, 387–396. doi: 10.21603/2308-4057-2021-2-387-396

Butte, S., Vakanski, A., Duellman, K., Wang, H., and Mirkouei, A. (2021). Potato crop stress identification in aerial images using deep learning-based object detection. *Agron. J.* 113, 3991–4002. doi: 10.1002/agj2.20841

Bzdok, D., Krzywinski, M., and Altman, N. (2018). Machine learning: supervised methods. *Nat. Methods* 15, 5. doi: 10.1038/nmeth.4551

Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Zhang, J., et al. (2021). Integrating multi-source data for rice yield prediction across China using machine learning and deep learning approaches. *Agric. For. Meteorology* 297, 108275. doi: 10.1016/j.agrformet.2020.108275

Chang, H.-X., Haudenshield, J. S., Bowen, C. R., and Hartman, G. L. (2017). Metagenome-wide association study and machine learning prediction of bulk soil microbiome and crop productivity. *Front. Microbiol.* 8, 519. doi: 10.3389/fmicb.2017.00519

Chatterjee, T., Gogoi, U. R., Samanta, A., Chatterjee, A., Singh, M. K., and Pasupuleti, S. (2024). Identifying the most discriminative parameter for water quality prediction using machine learning algorithms. *Water* 16, 481. doi: 10.3390/w16030481

Chen, T., and Guestrin, C. (2016). "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. doi: 10.1145/2939672

Chen, Y., Huang, Y., Zhang, Z., Wang, Z., Liu, B., Liu, C., et al. (2023). Plant image recognition with deep learning: A review. *Comput. Electron. Agric.* 212, 108072. doi: 10.1016/j.compag.2023.108072

Chrysostomou, C., Seker, H., and Aydin, N. (2011). "Effects of windowing and zero-padding on complex resonant recognition model for protein sequence analysis," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Boston, MA, USA. doi: 10.1109/IEMBS.2011.6091228

Colmer, J., O'Neill, C. M., Wells, R., Bostrom, A., Reynolds, D., Websdale, D., et al. (2020). SeedGerm: a cost-effective phenotyping platform for automated seed imaging and machine-learning based phenotypic analysis of crop seed germination. *New Phytol.* 228, 778–793. doi: 10.1111/nph.16736

Crick, F. (1989). The recent excitement about neural networks. *Nature* 337, 129–132. doi: 10.1038/337129a0

Cui, S., Su, Y. L., Duan, K., and Liu, Y. (2023). Maize leaf disease classification using CBAM and lightweight Autoencoder network. *J. Ambient Intell. Humanized Computing* 14, 7297–7307. doi: 10.1007/s12652-022-04438-z

Dahouda, M. K., and Joe, I. (2021). A deep-learned embedding technique for categorical features encoding. *IEEE Access* 9, 114381–114391. doi: 10.1109/ACCESS.2021.3104357

Danilevicz, M. F., Gill, M., Anderson, R., Batley, J., Bennamoun, M., Bayer, P. E., et al. (2022). Plant genotype to phenotype prediction using machine learning. *Front. Genet.* 13, 822173. doi: 10.3389/fgene.2022.822173

Das, P., Ivkin, N., Bansal, T., Rouesnel, L., Gautier, P., Karnin, Z., et al. (2020). "Amazon SageMaker Autopilot: a white box AutoML solution at scale," in *Proceedings of the fourth international workshop on data management for end-to-end machine learning*. doi: 10.1145/3399579

Davis, R. L., Greene, J. K., Dou, F., Jo, Y.-K., and Chappell, T. M. (2020). A practical application of unsupervised machine learning for analyzing plant image data collected using unmanned aircraft systems. *Agronomy* 10, 633. doi: 10.3390/agronomy10050633

Dege, D., and Brüggemann, P. (2023). *Marketing analytics with RStudio: a software review* (Springer). doi: 10.1057/s41270-023-00264-0

Deng, F., Mao, W., Zeng, Z., Zeng, H., and Wei, B. (2022). Multiple diseases and pests detection based on federated learning and improved faster R-CNN. *IEEE Trans. Instrumentation Measurement* 71, 1–11. doi: 10.1109/TIM.2022.3201937

Dhaliwal, D. S., and Williams, M. M. (2024). Sweet corn yield prediction using machine learning models and field-level data. *Precis. Agric.* 25, 51–64. doi: 10.1007/s11119-023-10057-1

Dhillon, M. S., Dahms, T., Kuebert-Flock, C., Rummler, T., Arnault, J., Steffan-Dewenter, I., et al. (2023). Integrating random forest and crop modeling improves the crop yield prediction of winter wheat and oil seed rape. *Front. Remote Sens.* 3, 1010978. doi: 10.3389/frsen.2022.1010978

Dhillon, R., Takoo, G., Sharma, V., and Nagle, M. (2024). Utilizing machine learning framework to evaluate the effect of climate change on maize and soybean yield. *Comput. Electron. Agric.* 221, 108982. doi: 10.1016/j.compag.2024.108982

Diao, Z., Yan, J., He, Z., Zhao, S., and Guo, P. (2022). Corn seedling recognition algorithm based on hyperspectral image and lightweight-3D-CNN. *Comput. Electron. Agric.* 201, 107343. doi: 10.1016/j.compag.2022.107343

Doersch, C. (2016). Tutorial on variational autoencoders. *arXiv preprint*. arXiv:1606.05908. doi: 10.48550/arXiv.1606.05908

Du, C., Jiang, S., Chen, C., Guo, Q., He, Q., and Zhan, C. (2024). Machine learning-based estimation of daily cropland evapotranspiration in diverse climate zones. *Remote Sens.* 16, 730. doi: 10.3390/rs16050730

Du, Z., Yang, L., Zhang, D., Cui, T., He, X., Xiao, T., et al. (2022). Corn variable-rate seeding decision based on gradient boosting decision tree model. *Comput. Electron. Agric.* 198, 107025. doi: 10.1016/j.compag.2022.107025

Duc, N. T., Ramlal, A., Rajendran, A., Raju, D., Lal, S., Kumar, S., et al. (2023). Image-based phenotyping of seed architectural traits and prediction of seed weight using machine learning models in soybean. *Front. Plant Sci.* 14, 1206357. doi: 10.3389/fpls.2023.1206357

Durrani, A. U. R., Minallah, N., Aziz, N., Frnda, J., Khan, W., and Nedoma, J. (2023). Effect of hyper-parameters on the performance of ConvLSTM based deep neural network in crop classification. *PloS One* 18, e0275653. doi: 10.1371/journal.pone.0275653

Eelbode, T., Sinonquel, P., Maes, F., and Bisschops, R. (2021). Pitfalls in training and validation of deep learning systems. *Best Pract. Res. Clin. Gastroenterol.* 52, 101712. doi: 10.1016/j.bpg.2020.101712

Elango, E., Hanees, A., Shanmuganathan, B., and Kareem Basha, M. I. (2024). "Precision Agriculture: A Novel Approach on AI-Driven Farming," in *Intelligent Robots and Drones for Precision Agriculture* (Springer), 119–137.

Elavarasan, D., and Vincent, P. D. (2020). Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. *IEEE Access* 8, 86886–86901. doi: 10.1109/Access.6287639

Elbasi, E., Zaki, C., Topcu, A. E., Abdelbaki, W., Zreikat, A. I., Cina, E., et al. (2023). Crop prediction model using machine learning algorithms. *Appl. Sci.* 13, 9288. doi: 10.3390/app13169288

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). kdd.*, 226–23.

Evamoni, F. Z., Nulit, R., Yap, C. K., Ibrahim, M. H., and Sidek, N. B. (2023). Assessment of germination performance and early seedling growth of Malaysian indica rice genotypes under drought conditions for strategic cropping during water scarcity. *Chilean J. Agric. Res.* 83, 281–292. doi: 10.4067/S0718-58392023000300281

Fey, M., and Lenssen, J. E. (2019). Fast graph representation learning with PyTorch Geometric. *arXiv preprint* arXiv:1903.02428. doi: 10.48550/arXiv.1903.02428

Fu, X., Ma, Q., Yang, F., Zhang, C., Zhao, X., Chang, F., et al. (2023). Crop pest image recognition based on the improved ViT method. *Inf. Process. Agric.* 11 (2), 249–259. https://doi.org/10.1016/j.inpa.2023.02.007

Gafurov, A., Mukharamova, S., Saveliev, A., and Yermolaev, O. (2023). Advancing agricultural crop recognition: the application of LSTM networks and spatial generalization in satellite data analysis. *Agriculture* 13, 1672. doi: 10.3390/agriculture13091672

Gano, B., Bhadra, S., Vilbig, J. M., Ahmed, N., Sagan, V., and Shakoor, N. (2024). Drone-based imaging sensors, techniques, and applications in plant phenotyping for crop breeding: A comprehensive review. *Plant Phenome J.* 7, e20100. doi: 10.1002/ppj2.20100

Gao, H., and Ji, S. (2019). "Graph u-nets," in *International conference on machine learning*.

Gauriau, O., Galárraga, L., Brun, F., Termier, A., Davadan, L., and Joudelat, F. (2024). Comparing machine-learning models of different levels of complexity for crop protection: A look into the complexity-accuracy tradeoff. *Smart Agric. Technol.* 7, 100380. doi: 10.1016/j.atech.2023.100380

Ge, W., Zhou, J., Zheng, P., Yuan, L., and Rottok, L. T. (2024). A recommendation model of rice fertilization using knowledge graph and case-based reasoning. *Comput. Electron. Agric.* 219, 108751. doi: 10.1016/j.compag.2024.108751

Gholap, P. S., Sharma, G., Deepak, A., Madan, P., Sharma, R., Sharma, M., et al. (2024). IoT enabled stress detection based on image processing with ensembling machine learning approach. *Int. J. Intelligent Syst. Appl. Eng.* 12, 760–768.

Ghosal, A., Nandy, A., Das, A. K., Goswami, S., and Panday, M. (2020). "A short review on different clustering techniques and their applications," in *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph*, Vol. 2018. 69–83.

Ghosh, H., Tusher, M. A., Rahat, I. S., Khasim, S., and Mohanty, S. N. (2023). "Water quality assessment through predictive machine learning," in *International Conference on Intelligent Computing and Networking*.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning* (MIT press).

Gopi, P., and Karthikeyan, M. (2024). Red fox optimization with ensemble recurrent neural network for crop recommendation and yield prediction model. *Multimedia Tools Appl.* 83, 13159–13179. doi: 10.1007/s11042-023-16113-2

Greener, J. G., Kandathil, S. M., Moffat, L., and Jones, D. T. (2022). A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* 23, 40–55. doi: 10.1038/s41580-021-00407-0

Guo, J., Li, H., Ning, J., Han, W., Zhang, W., and Zhou, Z.-S. (2020). Feature dimension reduction using stacked sparse auto-encoders for crop classification with multi-temporal, quad-pol SAR Data. *Remote Sens.* 12, 321. doi: 10.3390/rs12020321

Guo, T., and Li, X. (2023). Machine learning for predicting phenotype from genotype and environment. *Curr. Opin. Biotechnol.* 79, 102853. doi: 10.1016/j.copbio.2022.102853

Hamidi, M., Homayouni, S., Safari, A., and Hasani, H. (2024). Deep learning based crop-type mapping using SAR and optical data fusion. *Int. J. Appl. Earth Observation Geoinformation* 129, 103860. doi: 10.1016/j.jag.2024.103860

Hamidi, M., Safari, A., and Homayouni, S. (2021). An auto-encoder based classifier for crop mapping from multitemporal multispectral imagery. *Int. J. Remote Sens.* 42, 986–1016. doi: 10.1080/01431161.2020.1820619

Hara, K., Kataoka, H., and Satoh, Y. (2018). "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. doi: 10.1109/CVPR.2018.00685

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction, 2 (Springer). doi: 10.1007/978-0-387-84858-7

Heffner, E. L., Sorrells, M. E., and Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Sci.* 49, 1–12. doi: 10.2135/cropsci2008.08.0512

Hegedűs, I., Berta, Á., Kocsis, L., Benczúr, A. A., and Jelasity, M. (2016). Robust decentralized low-rank matrix decomposition. *ACM Trans. Intelligent Syst. Technol. (TIST)* 7, 1–24. doi: 10.1145/2854157

Hegedűs, I., Danner, G., and Jelasity, M. (2019). "Gossip learning as a decentralized alternative to federated learning," in *Distributed Applications and Interoperable Systems*.

Herrera, J. M., Häner, L. L., Holzkämper, A., and Pellet, D. (2018). Evaluation of ridge regression for country-wide prediction of genotype-specific grain yields of wheat. *Agric. For. meteorology* 252, 1–9. doi: 10.1016/j.agrformet.2017.12.263

Ho, Y., and Wookey, S. (2019). The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access* 8, 4806–4813. doi: 10.1109/Access.6287639

Hu, K., Wang, Z., Coleman, G., Bender, A., Yao, T., Zeng, S., et al. (2021). Deep learning techniques for in-crop weed identification: A review. *arXiv preprint* arXiv:2103.14872. doi: 10.1007/s11119-023-10073-1

Huber, F., Yushchenko, A., Stratmann, B., and Steinhage, V. (2022). Extreme Gradient Boosting for yield estimation compared with Deep Learning approaches. *Comput. Electron. Agric.* 202, 107346. doi: 10.1016/j.compag.2022.107346

Idoje, G., Dagiuklas, T., and Iqbal, M. (2023). Federated Learning: Crop classification in a smart farm decentralised network. *Smart Agric. Technol.* 5, 100277. doi: 10.1016/j.atech.2023.100277

Iniyan, S., Varma, V. A., and Naidu, C. T. (2023). Crop yield prediction using machine learning techniques. *Adv. Eng. Software* 175, 103326. doi: 10.1016/j.advengsoft.2022.103326

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition Lett.* 31, 651–666. doi: 10.1016/j.patrec.2009.09.011

Jain, T., Garg, P., Tiwari, P. K., Kuncham, V. K., Sharma, M., and Verma, V. K. (2021). "Performance prediction for crop irrigation using different machine learning approaches," in *Examining the Impact of Deep Learning and IoT on Multi-Industry Applications* (IGI Global), 61–79.

James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). "Unsupervised learning," in *An Introduction to Statistical Learning: with Applications in Python* (Springer), 503–556. doi: 10.1007/978-3-031-38747-0

Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., et al. (2016). Random forests for global and regional crop yield predictions. *PloS One* 11, e0156571. doi: 10.1371/journal.pone.0156571

Ji, S., Zhang, C., Xu, A., Shi, Y., and Duan, Y. (2018). 3D convolutional neural networks for crop classification with multi-temporal remote sensing images. *Remote Sens.* 10, 75. doi: 10.3390/rs10010075

Jiang, Y., and Li, C. (2020). Convolutional neural networks for image-based high-throughput plant phenotyping: a review. *Plant Phenomics*. doi: 10.34133/2020/4152816

Jiang, H., Zhang, C., Qiao, Y., Zhang, Z., Zhang, W., and Song, C. (2020). CNN feature based graph convolutional network for weed and crop recognition in smart farming. *Comput. Electron. Agric.* 174, 105450. doi: 10.1016/j.compag.2020.105450

Jolliffe, I. T., and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc. A: Mathematical Phys. Eng. Sci.* 374, 20150202. doi: 10.1098/rsta.2015.0202

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., et al. (2021). Advances and open problems in federated learning. *Foundations trends® Mach. Learn.* 14, 1–210. doi: 10.1561/2200000083

Kamatchi, C. B., and Muthukumaravel, A. (2024). "Machine learning in agriculture: A land data approach to optimize crop choice with the LAGNet model," in *2024 International Conference on Emerging Smart Computing and Informatics (ESCI)*, Pune, India. doi: 10.1109/ESCI59607.2024.10497261

Kashyap, G. R., Sridhara, S., Manoj, K. N., Gopakkali, P., Das, B., Jha, P. K., et al. (2024). Machine learning ensembles, neural network, hybrid and sparse regression approaches for weather based rainfed cotton yield forecast. *Int. J. Biometeorol.* 68, 1179–1197. doi: 10.1007/s00484-024-02661-1

Kavitha, E., Jadhav, H. M., Goyal, V., Deepak, A., Pokhariya, H. S., Sharma, B. D., et al. (2024). Utilizing convolutional neural networks for image-based crop classification system. *Int. J. Intelligent Syst. Appl. Eng.* 12, 685–694.

Kevin, I., Wang, K., Ye, X., and Sakurai, K. (2022). "Federated learning with clustering-based participant selection for IoT applications," in *2022 IEEE International Conference on Big Data (Big Data)*. Osaka, Japan. doi: 10.1109/BigData55660.2022.10020575

Kheir, A., Nangia, V., Elnashar, A., Devakota, M., Omar, M., Feike, T., et al. (2024). Developing automated machine learning approach for fast and robust crop yield prediction using a fusion of remote sensing, soil, and weather dataset. *Environ. Res. Commun.* doi: 10.1088/2515-7620/ad2d02

Khoshnevisan, B., Bolandnazar, E., Barak, S., Shamshirband, S., Maghsoudlou, H., Altameem, T. A., et al. (2015). A clustering model based on an evolutionary algorithm for better energy use in crop production. *Stochastic Environ. Res. Risk Assess.* 29, 1921–1935. doi: 10.1007/s00477-014-0972-6

Killeen, P., Kiringa, I., Yeap, T., and Branco, P. (2024). Corn grain yield prediction using UAV-based high spatiotemporal resolution imagery, machine learning, and spatial cross-validation. *Remote Sens.* 16, 683. doi: 10.3390/rs16040683

Kircher, M., Witten, D. M., Jain, P., O'roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310–315. doi: 10.1038/ng.2892

Kolipaka, V. R. R., and Namburu, A. (2024). An automatic crop yield prediction framework designed with two-stage classifiers: a meta-heuristic approach. *Multimedia Tools Appl.* 83, 28969–28992. doi: 10.1007/s11042-023-16612-2

Kondermann, D. (2013). "Ground truth design principles: an overview," in *Proceedings of the International Workshop on Video and Image Ground Truth in Computer Vision Applications*. ACM: New York, NY, USA. doi: 10.1145/2501105

Korani, W., Clevenger, J. P., Chu, Y., and Ozias-Akins, P. (2019). Machine learning as an effective method for identifying true single nucleotide polymorphisms in polyploid plants. *Plant Genome* 12, 180023. doi: 10.3835/plantgenome2018.05.0023

Kuhn, M. (2008). Building predictive models in R using the caret package. *J. Stat. software* 28, 1–26. doi: 10.18637/jss.v028.i05

Kulkarni, P., and Shastri, S. (2024). Rice leaf diseases detection using machine learning. *J. Sci. Res. Technol.*, 17–22. doi: 10.61808/jsrt81

Lai, W., and Yan, Q. (2022). "Federated learning for detecting COVID-19 in chest CT images: a lightweight federated learning approach," in *2022 4th International Conference on Frontiers Technology of Information and Computer (ICFTIC)*. Qingdao, China. doi: 10.1109/ICFTIC57696.2022.10075165

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Li, D., Han, D., Weng, T.-H., Zheng, Z., Li, H., Liu, H., et al. (2022). Blockchain for federated learning toward secure distributed machine learning systems: a systemic survey. *Soft Computing* 26, 4423–4440. doi: 10.1007/s00500-021-06496-5

Li, Z., Huffman, T., Zhang, A., Zhou, F., and McConkey, B. (2012). Spatially locating soil classes within complex soil polygons–Mapping soil capability for agriculture in Saskatchewan Canada. *Agriculture Ecosyst. Environ.* 152, 59–67. doi: 10.1016/j.agee.2012.02.007

Li, Z., Liu, F., Yang, W., Peng, S., and Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Trans. Neural Networks Learn. Syst.* 33, 6999–7019. doi: 10.1109/TNNLS.2021.3084827

Li, L., Zhang, Y., Wang, B., Feng, P., He, Q., Shi, Y., et al. (2023). Integrating machine learning and environmental variables to constrain uncertainty in crop yield change projections under climate change. *Eur. J. Agron.* 149, 126917. doi: 10.1016/j.eja.2023.126917

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., and Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors* 18, 2674. doi: 10.3390/s18082674

Liao, C., Wang, J., Xie, Q., Baz, A. A., Huang, X., Shang, J., et al. (2020). Synergistic use of multi-temporal RADARSAT-2 and VENμS data for crop classification based on 1D convolutional neural network. *Remote Sens.* 12, 832. doi: 10.3390/rs12050832

Lingwal, S., Bhatia, K. K., and Singh, M. (2024). A novel machine learning approach for rice yield estimation. *J. Exp. Theor. Artif. Intell.* 36, 337–356. doi: 10.1080/0952813X.2022.2062458

Liu, J., Wang, T., Skidmore, A., Sun, Y., Jia, P., and Zhang, K. (2023). Integrated 1D, 2D, and 3D CNNs enable robust and efficient land cover classification from hyperspectral imagery. *Remote Sens.* 15, 4797. doi: 10.3390/rs15194797

Liu, T., Zhai, D., He, F., and Yu, J. (2024). Semi-supervised learning methods for weed detection in turf. *Pest Manage. Sci.* doi: 10.1002/ps.7959

Madala, K., and Prasad, M. S. G. (2023). Crop mapping through hybrid capsule transient auto-encoder technique based on radar features. *Multimedia Tools Appl.* 8, 1–31. doi: 10.1007/s11042-023-17327-0

Mammen, P. M. (2021). Federated learning: Opportunities and challenges. *arXiv preprint arXiv*:2101.05428. doi: 10.48550/arXiv.2101.05428

Manoj, T., Makkithaya, K., and Narendra, V. (2022). "A federated learning-based crop yield prediction for agricultural production risk management," in *2022 IEEE Delhi Section Conference (DELCON)*. New Delhi, India. doi: 10.1109/DELCON54057.2022.9752836.

Masjedi, A., Carpenter, N. R., Crawford, M. M., and Tuinstra, M. R. (2019). "Prediction of sorghum biomass using UAV time series data and recurrent neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. doi: 10.1109/CVPRW47913.2019

Mathai, N., Chen, Y., and Kirchmair, J. (2020). Validation strategies for target prediction methods. *Briefings Bioinf.* 21, 791–802. doi: 10.1093/bib/bbz026

Mazzia, V., Khaliq, A., and Chiaberge, M. (2019). Improvement in land cover and crop classification based on temporal features learning from Sentinel-2 data using recurrent-convolutional neural network (R-CNN). *Appl. Sci.* 10, 238. doi: 10.3390/app10010238

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. (2017). "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. Seattle, WA 98103 USA.

Medsker, L. R., and Jain, L. (2001). Recurrent neural networks. *Design Appl.* 5, 2.

Meenal, R., Jala, P. K., Samundeswari, R., and Rajasekaran, E. (2024). Crop water management using machine learning-based evapotranspiration estimation. *J. Appl. Biol. Biotechnol.* 12, 198–203. doi: 10.7324/JABB.2024.155791

Memon, R., Memon, M., Malioto, N., and Raza, M. O. (2021). "Identification of growth stages of crops using mobile phone images and machine learning," in *2021 International conference on computing, Electronic and Electrical Engineering (ICE Cube)*. Quetta, Pakistan. doi: 10.1109/ICECube53880.2021.9628197

Mochida, K., Koda, S., Inoue, K., Hirayama, T., Tanaka, S., Nishii, R., et al. (2019). Computer vision-based phenotyping for improvement of plant productivity: a machine learning perspective. *GigaScience* 8, giy153. doi: 10.1093/gigascience/giy153

Modi, R. U., Kancheti, M., Subeesh, A., Raj, C., Singh, A. K., Chandel, N. S., et al. (2023). An automated weed identification framework for sugarcane crop: a deep learning approach. *Crop Prot.* 173, 106360. doi: 10.1016/j.cropro.2023.106360

Moharram, M. A., and Sundaram, D. M. (2023). Land Use and Land Cover Classification with Hyperspectral Data: A comprehensive review of methods, challenges and future directions. *Neurocomputing*. doi: 10.1016/j.neucom.2023.03.025

Morales, A., and Villalobos, F. J. (2023). Using machine learning for crop yield prediction in the past or the future. *Front. Plant Sci.* 14, 1128388. doi: 10.3389/fpls.2023.1128388

Mora-Poblete, F., Maldonado, C., Henrique, L., Uhdre, R., Scapim, C. A., and Mangolim, C. A. (2023). Multi-trait and multi-environment genomic prediction for flowering traits in maize: a deep learning approach. *Front. Plant Sci.* 14, 1153040. doi: 10.3389/fpls.2023.1153040

Mosavi, A., Samadianfard, S., Darbandi, S., Nabipour, N., Qasem, S. N., Salwana, E., et al. (2021). Predicting soil electrical conductivity using multi-layer perceptron integrated with grey wolf optimizer. *J. Geochemical Explor.* 220, 106639. doi: 10.1016/j.gexplo.2020.106639

Moshawrab, M., Adda, M., Bouzouane, A., Ibrahim, H., and Raad, A. (2023). Reviewing federated learning aggregation algorithms; strategies, contributions, limitations and future perspectives. *Electronics* 12, 2287. doi: 10.3390/electronics12102287

Mousavi, S. R., Jahandideh Mahjenabadi, V. A., Khoshru, B., and Rezaei, M. (2024). Spatial prediction of winter wheat yield gap: agro-climatic model and machine learning approaches. *Front. Plant Sci.* 14, 1309171. doi: 10.3389/fpls.2023.1309171

Nar, K., Ocal, O., Sastry, S. S., and Ramchandran, K. (2019). Cross-entropy loss and low-rank features have responsibility for adversarial examples. *arXiv preprint arXiv*:1901.08360. doi: 10.48550/arXiv.1901.08360

Neal, B. (2019). On the bias-variance tradeoff: Textbooks need an update. *arXiv preprint arXiv*:1912.08286. doi: 10.48550/arXiv.1912.08286

Nejad, S. M. M., Abbasi-Moghadam, D., Sharifi, A., Farmonov, N., Amankulova, K., and László, M. (2022). Multispectral crop yield prediction using 3D-convolutional neural networks and attention convolutional LSTM approaches. *IEEE J. Selected Topics Appl. Earth Observations Remote Sens.* 16, 254–266. doi: 10.1109/JSTARS.2022.3223423

Nevavuori, P., Narra, N., and Lipping, T. (2019). Crop yield prediction with deep convolutional neural networks. *Comput. Electron. Agric.* 163, 104859. doi: 10.1016/j.compag.2019.104859

Noble, W. S. (2006). What is a support vector machine? *Nat. Biotechnol.* 24, 1565–1567. doi: 10.1038/nbt1206-1565

Olson, R. S., Cava, W. L., Mustahsan, Z., Varik, A., and Moore, J. H. (2018). "Data-driven advice for applying machine learning to bioinformatics problems," in *Pacific Symposium on Biocomputing 2018: Proceedings of the Pacific Symposium*. doi: 10.1142/10864

Opach, T., and Rød, J. K. (2018). Augmenting the usability of parallel coordinate plot: The polyline glyphs. *Inf. Visualization* 17, 108–127. doi: 10.1177/1473871617693041

Ormándi, R., Hegedűs, I., and Jelasity, M. (2013). Gossip learning with linear models on fully distributed data. *Concurrency Computation: Pract. Exp.* 25, 556–571. doi: 10.48550/arXiv.1109.1396

Osman, Y., Dennis, R., and Elgazzar, K. (2021). Yield estimation and visualization solution for precision agriculture. *Sensors* 21, 6657. doi: 10.3390/s21196657

Ouali, Y., Hudelot, C., and Tami, M. (2020). An overview of deep semi-supervised learning. *arXiv preprint arXiv*:2006.05278. doi: 10.48550/arXiv.2006.05278

Pandey, S., Yadav, P. K., Sahu, R., and Pandey, P. (2024). "Improving crop management with convolutional neural networks for binary and multiclass weed recognition," in *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*. Bengaluru, India. doi: 10.1109/IDCIoT59759.2024.10467501

Panigrahi, B., Kathala, K. C. R., and Sujatha, M. (2023). A machine learning-based comparative approach to predict the crop yield using supervised learning with regression models. *Proc. Comput. Sci.* 218, 2684–2693. doi: 10.1016/j.procs.2023.01.241

Pardoe, I. (2020). *Applied regression modeling* (John Wiley & Sons). doi: 10.1002/9781119615941

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* 32, 8026–8037. doi: 10.48550/arXiv.1912.01703

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia* 4, 1883. doi: 10.4249/scholarpedia.1883

Picon, A., San-Emeterio, M. G., Bereciartua-Perez, A., Klukas, C., Eggers, T., and Navarra-Mestre, R. (2022). Deep learning-based segmentation of multiple species of

weeds and corn crop using synthetic and real image datasets. *Comput. Electron. Agric.* 194, 106719. doi: 10.1016/j.compag.2022.106719

Piekutowska, M., Niedbała, G., Piskier, T., Lenartowicz, T., Pilarski, K., Wojciechowski, T., et al. (2021). The application of multiple linear regression and artificial neural network models for yield prediction of very early potato cultivars before harvest. *Agronomy* 11, 885. doi: 10.3390/agronomy11050885

Rajamani, S. K., and Iyer, R. S. (2023). "Machine Learning-Based Mobile Applications Using Python and Scikit-Learn," in *Designing and developing innovative mobile applications* (IGI Global), 282–306.

Rangarajan, A. K., Purushothaman, R., Prabhakar, M., and Szczepański, C. (2023). Crop identification and disease classification using traditional machine learning and deep learning approaches. *J. Eng. Res.* 11, 228–252. doi: 10.36909/jer.11941

Rodríguez, P., Bautista, M. A., Gonzalez, J., and Escalera, S. (2018). Beyond one-hot encoding: Lower dimensional target embedding. *Image Vision Computing* 75, 21–31. doi: 10.1016/j.imavis.2018.04.004

Sahoo, R. N., Rejith, R., Gakhar, S., Ranjan, R., Meena, M. C., Dey, A., et al. (2024). Drone remote sensing of wheat N using hyperspectral sensor and machine learning. *Precis. Agric.* 25, 704–728. doi: 10.1007/s11119-023-10089-7

Sarkar, C., Gupta, D., Gupta, U., and Hazarika, B. B. (2023). Leaf disease detection using machine learning and deep learning: Review and challenges. *Appl. Soft Computing* 22, 110534. doi: 10.1016/j.asoc.2023.110534

Sehrawat, S., Najafian, K., and Jin, L. (2023). Predicting phenotypes from novel genomic markers using deep learning. *Bioinf. Adv.* 3, vbad028. doi: 10.1093/bioadv/vbad028

Seising, R. (2018). The emergence of fuzzy sets in the decade of the perceptron—Lotfi A. Zadeh's and frank rosenblatt's research work on pattern classification. *Mathematics* 6, 110.

Sejnowski, T. J. (2018). *The deep learning revolution* (MIT press). doi: 10.7551/mitpress/11474.001.0001

Sen, P. C., Hajra, M., and Ghosh, M. (2020). "Supervised classification algorithms in machine learning: A survey and review," in *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018*.

Shafiee, S., Lied, L. M., Burud, I., Dieseth, J. A., Alsheikh, M., and Lillemo, M. (2021). Sequential forward selection and support vector regression in comparison to LASSO regression for spring wheat yield prediction based on UAV imagery. *Comput. Electron. Agric.* 183, 106036. doi: 10.1016/j.compag.2021.106036

Sharma, A., Jain, A., Gupta, P., and Chowdary, V. (2020). Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access* 9, 4843–4873. doi: 10.1109/Access.6287639

Shin, A., Kim, D. Y., Jeong, J. S., and Chun, B.-G. (2020). Hippo: Taming hyperparameter optimization of deep learning with stage trees. *arXiv preprint arXiv*:2006.11972.

Sifaou, H., and Li, G. Y. (2022). "Robust federated learning via over-the-air computation," in *2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*. doi: 10.1109/MLSP55214.2022.9943401

Sindhu Meena, K., and Suriya, S. (2020). "A survey on supervised and unsupervised learning techniques," in *Proceedings of international conference on artificial intelligence, smart grid and smart city applications: AISGSC 2019*.

Smith, A. M., Walsh, J. R., Long, J., Davis, C. B., Henstock, P., Hodge, M. R., et al. (2020). Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinf.* 21, 1–18. doi: 10.1186/s12859-020-3427-8

Srinivas, L., Bharathy, A. V., Ramakuri, S. K., Sethy, A., and Kumar, R. (2024). An optimized machine learning framework for crop disease detection. *Multimedia Tools Appl.* 83, 1539–1558. doi: 10.1007/s11042-023-15446-2

Su, X., Yan, X., and Tsai, C. L. (2012). Linear regression. *Wiley Interdiscip. Reviews: Comput. Stat* 4, 275–294.

Sudha, M. K., Manorama, M., and Aditi, T. (2022). Smart agricultural decision support systems for predicting soil nutrition value using IoT and ridge regression. *AGRIS on-line Papers Economics Inf.* 14, 95–106. doi: 10.7160/aol.2022.140108

Sun, Z., Di, L., and Fang, H. (2019). Using long short-term memory recurrent neural network in land cover classification on Landsat and Cropland data layer time series. *Int. J. Remote Sens.* 40, 593–614. doi: 10.1080/01431161.2018.1516313

Tang, Z., Hu, Y., and Zhang, Z. (2023). ROOSTER: An image labeler and classifier through interactive recurrent annotation. *F1000Research* 12, 137. doi: 10.12688/f1000research

Tang, W., Long, G., Liu, L., Zhou, T., Jiang, J., and Blumenstein, M. (2020). Rethinking 1d-cnn for time series classification: A stronger baseline. *arXiv preprint arXiv:2002.10061*, 1–7.

Tian, Y., Yang, C., Huang, W., Tang, J., Li, X., and Zhang, Q. (2021). Machine learning-based crop recognition from aerial remote sensing imagery. *Front. Earth Sci.* 15, 54–69. doi: 10.1007/s11707-020-0861-x

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B: Stat. Method.* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tong, H., and Nikoloski, Z. (2021). Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. *J. Plant Physiol.* 257, 153354. doi: 10.1016/j.jplph.2020.153354

Twomey, J. M., and Smith, A. E. (1997). Validation and verification. *Artif. Neural Networks civil engineers: Fundamentals Appl.* (New York: ASCE), 44–64.

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. (2021). Learning from disagreement: A survey. *J. Artif. Intell. Res.* 72, 1385–1470. doi: 10.1613/jair.1.12752

Uppu, S., Krishna, A., and Gopalan, R. P. (2016). A deep learning approach to detect SNP interactions. *J. Softw* 11, 965–975. doi: 10.17706/jsw.11.10.965-975

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9.

Vani, P. S., and Rathi, S. (2023). Improved data clustering methods and integrated A-FP algorithm for crop yield prediction. *Distributed Parallel Database* 41, 117–131.

Van Klompenburg, T., Kassahun, A., and Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.* 177, 105709. doi: 10.1016/j.compag.2020.105709

Veenadhari, S., Misra, B., and Singh, C. (2014). "Machine learning approach for forecasting crop yield based on climatic parameters," in *2014 International Conference on Computer Communication and Informatics*. doi: 10.1109/ICCCI.2014.6921718

Venkataraju, A., Arumugam, D., Stepan, C., Kiran, R., and Peters, T. (2023). A review of machine learning techniques for identifying weeds in corn. *Smart Agric. Technol.* 3, 100102. doi: 10.1016/j.atech.2022.100102

Wang, C., and Zhang, Y. (2017). Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. *J. Comput. Chem.* 38, 169–177. doi: 10.1002/jcc.24667

Wei, Q., and Dunbrack, R. L.Jr. (2013). The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PloS One* 8, e67863. doi: 10.1371/journal.pone.0067863

Wei, J., Zhang, M., Wu, C., Ma, Q., Wang, W., and Wan, C. (2024). Accurate crop row recognition of maize at the seedling stage using lightweight network. *Int. J. Agric. Biol. Eng.* 17, 189–198. doi: 10.25165/j.ijabe.20241701.7051

Wen, J., Zhang, Z., Lan, Y., Cui, Z., Cai, J., and Zhang, W. (2023). A survey on federated learning: challenges and applications. *Int. J. Mach. Learn. Cybernetics* 14, 513–535. doi: 10.1007/s13042-022-01647-y

Wu, Y.-C., and Feng, J.-W. (2018). Development and application of artificial neural network. *Wireless Pers. Commun.* 102, 1645–1656. doi: 10.1007/s11277-017-5224-x

Wu, D., Yang, Z., Li, T., and Liu, J. (2024). JOCP: A jointly optimized clustering protocol for industrial wireless sensor networks using double-layer selection evolutionary algorithm. *Concurrency Computation: Pract. Exp.* 36, e7927.

Yamaç, S. S., and Todorovic, M. (2020). Estimation of daily potato crop evapotranspiration using three different machine learning algorithms and four scenarios of available meteorological data. *Agric. Water Manage.* 228, 105875. doi: 10.1016/j.agwat.2019.105875

Yang, F., Zhang, D., Zhang, Y., Zhang, Y., Han, Y., Zhang, Q., et al. (2023). Prediction of corn variety yield with attribute-missing data via graph neural network. *Comput. Electron. Agric.* 211, 108046. doi: 10.1016/j.compag.2023.108046

Yesugade, K., Kharde, A., Mirashi, K., Muley, K., and Chudasama, H. (2018). Machine learning approach for crop selection based on agro-climatic conditions. *Mach. Learn.* 7. doi: 10.17148/IJARCCE

Yoosefzadeh-Najafabadi, M., Eskandari, M., Torabi, S., Torkamaneh, D., Tulpan, D., and Rajcan, I. (2022). Machine-learning-based genome-wide association studies for uncovering QTL underlying soybean yield and its components. *Int. J. Mol. Sci.* 23, 5538. doi: 10.3390/ijms23105538

Yu, C., Shen, S., Zhang, K., Zhao, H., and Shi, Y. (2022a). "Energy-aware device scheduling for joint federated learning in edge-assisted internet of agriculture things," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*. doi: 10.1109/WCNC51071.2022.9771547

Yu, L., Zhou, R., Chen, R., and Lai, K. K. (2022b). Missing data preprocessing in credit classification: One-hot encoding or imputation? *Emerging Markets Finance Trade* 58, 472–482. doi: 10.1080/1540496X.2020.1825935

Yu, T., and Zhu, H. (2020). Hyper-parameter optimization: A review of algorithms and applications. *arXiv preprint arXiv*:2003.05689.

Zanella, M. A., Martins, R. N., da Silva, F. M., Carvalho, L. C. C., de Carvalho Alves, M., and Rosas, J. T. F. (2024). Coffee yield prediction using high-resolution satellite imagery and crop nutritional status in Southeast Brazil. *Remote Sens. Applications: Soc. Environ.* 33, 101092.

Zhang, Z., Boubin, J., Stewart, C., and Khanal, S. (2020). Whole-field reinforcement learning: A fully autonomous aerial scouting method for precision agriculture. *Sensors* 20, 6585. doi: 10.3390/s20226585

Zhang, T., Cai, Y., Zhuang, P., and Li, J. (2024). Remotely sensed crop disease monitoring by machine learning algorithms: A review. *Unmanned Syst.* 12, 161–171. doi: 10.1142/S2301385024500237

Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., and Gao, Y. (2021a). A survey on federated learning. *Knowledge-Based Syst.* 216, 106775. doi: 10.1016/j.knosys.2021.106775

Zhang, L., Zhang, Z., Tao, F., Luo, Y., Cao, J., Li, Z., et al. (2021b). Planning maize hybrids adaptation to future climate change by integrating crop modelling with machine learning. *Environ. Res. Lett.* 16, 124043. doi: 10.1088/1748-9326/ac32fd

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B: Stat. Method.* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 10 September 2024
DOI 10.3389/fpls.2024.1460038

OPEN ACCESS

EDITED BY
Yuri Shavrukov,
Flinders University, Australia

REVIEWED BY
Sławomir Sowa,
Plant Breeding and Acclimatization Institute, Poland
Theo Prins,
Wageningen University and Research, Netherlands

*CORRESPONDENCE
Feiwu Li
lifeiwu3394@sina.com

RECEIVED 05 July 2024
ACCEPTED 20 August 2024
PUBLISHED 10 September 2024

CITATION
Long L, Zhao N, Li C, He Y, Dong L, Yan W, Xing Z, Xia W, Ma Y, Xie Y, Liu N and Li F (2024) Development and collaborative validation of an event-specific quantitative real-time PCR method for detection of genetically modified CC-2 maize. *Front. Plant Sci.* 15:1460038. doi: 10.3389/fpls.2024.1460038

Development and collaborative validation of an event-specific quantitative real-time PCR method for detection of genetically modified CC-2 maize

Likun Long, Ning Zhao, Congcong Li, Yuxuan He, Liming Dong, Wei Yan, Zhenjuan Xing, Wei Xia, Yue Ma, Yanbo Xie, Na Liu and Feiwu Li*

Institute of Agricultural Quality Standard and Testing Technology, Jilin Academy of Agricultural Sciences, Changchun, China

As one of the developed genetically modified (GM) maize varieties in China, CC-2 has demonstrated promising commercial prospects during demonstration planting. The establishment of detection methods is a technical prerequisite for effective supervision and regulation of CC-2 maize. In this study, we have developed an event-specific quantification method that targets the junction region between the exogenous gene and the 5′ flanking genomic DNA (gDNA) of CC-2. The accuracy and precision of this method were evaluated across high, medium, and low levels of CC-2 maize content, revealing biases within ±25% and satisfactory precision data. Additionally, we determined the limits of quantification of the method to be 0.05% (equivalent to 20 copies) of the CC-2 maize. A collaborative trial further confirmed that our event-specific method for detecting CC-2 produces reliable, comparable, and reproducible results when applied to five different samples provided by various sources. Furthermore, we calculated the expanded uncertainty associated with determining the content level of CC-2 in these samples.

KEYWORDS

genetically modified maize CC-2, event-specific PCR, quantification, real-time quantitative PCR, detection, validation

1 Introduction

Maize (*Zea mays L.*) occupies a significant position in the domain of genetically modified organisms (GMOs), with more than 25% of global maize varieties undergoing genetic alteration (ISAAA, 2021). As of 2023, it holds the record for the highest number of approved GM crop events, totalling 69.32 million hectares and the number of authorized

maize events has reached 421 (GM AgbioInvestor Monitor, 2024). Maize accounts for 146 of these events (Meng et al., 2022). The cultivation and utilization of genetically modified maize for food or feed purposes is becoming increasingly widespread on a global scale (Turkec et al., 2015; Avsar et al., 2020).

Prior to the commercial release of a new GM event, it is widely acknowledged that regulatory must be conducted to assess their potential impacts on human, animal and environmental health (Akinbo et al., 2021; Giraldo et al., 2019). The incorporation of tracking and tracing tools for transgenic insertion is considered an indispensable component of the deregulation process (European Commission Regulation (EC) No.1830/2003). Furthermore, the development of detection methods for GM identification and quantification is not only pivotal for ensuring legality and traceability but also for compliance with GM labelling regulations (Gruère and Rao, 2007). Moreover, method validation plays a crucial role in standardizing GM testing methods to ensure that GM testing laboratories can generate reliable analytical results (Meng et al., 2022).

The GM maize CC-2, developed by the Chinese Agricultural University (CN Patent No. CN105331725A), is a transgenic maize event containing modified *EPSPS* genes linking with the the chloroplast signaling peptide of sorghum named *maroACC* gene (Chen et al., 2019; Fan et al., 2018). It is one of the first batch of GM maize varieties in China obtaining production safety certificates, and it will have great commercial potential. Therefore, developing highly specific event-specific PCR methods for GM maize CC-2 and its derivatives is of significant importance for promoting the commercialization of GM maize in China. At the same time, the application of this method will also provide necessary technical support for safety testing, intellectual property protection, and product supervision (Li et al., 2020). Therefore, there is an urgent need for more specific, accurate, and standardized methods to meet the technical requirements of GM regulation in various countries' food trade regulations (Carzoli et al., 2018). Real-time PCR is widely regarded as the gold standard method for GM quantification in food or feed products (Li et al., 2022; Turkec et al., 2015; Avsar et al., 2020), though many other new GMO detection techniques such as microarray, digital PCR, re-sequencing, and biosensor, etc., were also reported (Yi et al., 2022; Li et al., 2017; Fraiture et al., 2021). The reliability of inter-laboratory results relies on method comparisons, validation, and harmonization. The inter-laboratory validation of methods is a crucial step in standardizing GM detection procedures, as it empowers the GM detection laboratory to generate dependable analytical outcomes.

This study established an event-specific real-time PCR method based on the molecular characteristics of CC-2 for the detection and quantification of GM maize CC-2. For this method, we organized a collaborative ring trial, which confirmed the specificity, applicability and viability of quantitative determination and the range of quantitative uncertainty. The development and application of the method will provide technical support for CC-2 maize commercialization supervision and implementation of quantitative labelling system.

2 Materials and methods

2.1 Plant material

Seeds of homozygous GM maize CC-2 and its non-GM maize ZD958 were provided by the Agricultural University of China. For specific testing, a total of 42 varieties of other GM crop events were graciously supplied by their respective developers, encompassing 13 transgenic maize (*Zea mays L.*) events (MON810, MON863, MON88017, MON89034, MON87460, MON87427, NK603, GA21, Bt176, Bt11, MIR604, MIR162, 3272, DAS-40278, 59122, 5307, 4114, T25, DBN9936, C0010.3.7 and TC1507), 7 transgenic soybean (*Glycine max L.*) events (GTS 40-3-2, A2704-12, MON89788, DP-356043, A5547-127, CV-127 and DP-305423), 7 transgenic rapeseed (*Brassica napus L.*) events (MS1, Topas19/2, Oxy-235, MS8, RF1/RF2/RF3 and T45) and 6 transgenic cotton (*Gossypium hirsutum L.*) events (MON531, MON88913, MON1445, MON15985, LLCotton25 and GHB614). All these seeds served as a source of genomic DNA for the purpose of this study.

2.2 Sample preparation

Seeds of CC-2 and non-GM maize were planted in the greenhouse. Genomic DNA extracted from leaves was used for quantitative DNA calibrant. Blind matrix samples containing different CC-2 event mass fractions (5%, 2%, 1%, 0.5%, and 0.1%) were created using ground seed powder from both types of maize, provided by the Development Center of Science and Technology, Ministry of Agriculture and Rural Affairs, China.

2.3 DNA extraction

The DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) was used to extract genomic DNA (gDNA) from plant or seed material following the manufacturer's instructions. The quality of the gDNA was evaluated by measuring the OD260/OD280 ratio with a Nanodrop 8000 (Thermo Scientific™ NanoDrop™, USA). To analyze low initial DNA concentrations, CC-2 maize DNA samples were diluted with water and prepared at concentrations of 10, 5, 0.4, 0.08, and 0.016 ng/μl using QubitR 2.0 (Life Technologies, United States). The copy number of gDNA was estimated based on the haploid genome size of maize being 2500 Megabasepairs (Arumuganathan and Earle, 1991), corresponding to a weight of 2.74 pg.

2.4 Primers and probes design

The design of primers and probes at the mutation positions was carried out using Primer Express Software 3.0 following the manufacturer's instructions. The design principles involved

placing one set of primers/probe on either the 5' or 3' side of the exogenous gene insertion locus in the genome, as well as specifying the region for the 5' end of probes to maintain sensitivity. Candidate primer pairs were additionally confirmed through traditional endpoint PCR to ensure generation of a single PCR product of correct size. Endogenous gene probes were labelled with 5'HEX, while specific probes were labelled with 5' FAM, both quenched with BHQ or MGB at the 3'end (Sangon BioTech, China). The details of the primers and probes used in this study are provided in Table 1.

2.5 Quantitative real-time PCR

The 7,500 Real-Time PCR System from Life Technologies AB (USA) was employed for quantitative real-time PCR analysis. The amplification system followed the instructions provided by Roche (Switzerland) for FastStart Universal Probe Master and ROX reference dye. qPCR analysis was conducted according to the methodology described in previous research (Guertler et al., 2019), with a minimum of three biological replicate samples included in each experiment. The maize endogenous gene *zSSIIb* (maize starch synthase IIb) and the event CC-2 specific fragment were separately amplified following the thermal cycle protocol: 95 °C for 5 min, 40 cycles at 95 °C for 15 s (denaturation), and 60 °C for 1 min (annealing and extension). Fluorescent signals were read out during the extension steps, and analyzed using the software Option Monitor 2 version 2.02 (MJ Research, Waltham, MA, USA).

2.6 Method verification

The validation of this method adheres to the standards outlined in the " Verification of analytical methods for GMO testing when implementing interlaboratory validated methods" (Hougs et al., 2017). Key parameters including dynamic range, accuracy, precision, limit of detection (LOD) and limit of quantification (LOQ) were assessed. Furthermore, the specificity of the method was investigated by analyzing DNA samples from diverse species encompassing GM maize events as well as soybean, cotton, sugar beet, and rapeseed. Additionally, sensitivity in detecting target sequences in other DNA samples was evaluated by combining an equal amount of DNA from three GM maize samples with a positive control (CC-2 maize).

2.7 Collaborative trials

The ring trial comprised eight GMO detection laboratories, all of which were affiliated with the Ministry of Agriculture, China. Each laboratory was provided with seven genomic samples: one sample labelled as CC-2 was utilized for constructing standard curves through serial dilution; one sample designated as a negative control; and five blind samples labelled S1, S2, S3, S4 and S5 representing CC-2 content levels of 5%, 2%, 1%, 0.5% and 0.1% respectively. Each sample had a volume of 100 µL at a concentration of 50 ng/µL. The genomic DNA samples along with the primers/probe were stored in an enclosed container filled with dry ice and dispatched to each participating laboratory.

3 Results and discussion

3.1 Establishment of CC-2 event-specific PCR

The 5' end of the insert and the flanking sequence of maize genomic DNA were provided and licensed by China Agricultural University for reference and method development (Patent No. CN201510856966). Blastn analysis against sequences in the GenBank database confirmed that the isolated junctions indeed spanned the integration border between the genome and integrated construct. Multiple primer-probe combinations specific to CC-2 were designed based on the 5' end-boundary genome sequence and CC-2 insertion, utilizing online software Primer3 (http://primer3.ut.ee/). These primers and probes were screened through qPCR amplification using CC-2 genomic DNA as a template. Among them, CC-2-F/R combined with QP primer probes exhibited a characteristic "S" shaped curve with stronger amplification signal and lower quantification cycle (Cq) value, making it a potential candidate for further analysis. The amplified fragment length of this combination was determined to be 107 bp through sequencing verification, which matched expectations. As an endogenous reference gene in maize, *zSSIIb* gene was employed

TABLE 1 Primer/probe information of CC-2 and zSSIIb.

Purpose	Name	Sequence (5'-3')	Amplicon size (bp)	Specificity	Source
PCR analysis of *zSSIIb* gene	zSSIIb-F	CGGTGGATGCTAAGGCTGATG	88	Maize genome	Yang et al., 2005
	zSSIIb-R	AAAGGGCCAGGTTCATTATCCTC			
	zSSIIb-P	HEX - TAAGGAGCACTCGCCGCCGCATCTG -BHQ1			
Event-specific PCR analysis of CC-2	CC-2-F	TGCAATGGGCCAGATCTAGTTA	107	5' junction of CC-2 event	this study
	CC-2-R	GCTCACTGAATTAACGCCGA			
	CC-2-P	FAM - CCAGTACTAAAATCCAGATCCCCCGA -BHQ1			

for quantitative PCR assay due to its confirmed single copy status in maize genome; moreover, real-time PCR protocol for *zSSIIb* gene had been previously validated (Yang et al., 2005). Specific information regarding primer probes is presented in Figure 1 while details about primers and probes are provided in Table 1.

In order to optimize the real-time fluorescence PCR reaction system, concentration gradients of six primers (0 μmol/L, 0.1 μmol/L, 0.2 μmol/L, 0.4 μmol/L, 0.6 μmol/L and 0.8 μmol/L) were set respectively, with the probe concentration being half of the primer concentration. The concentrations of primers and probes were determined based on fluorescence curves and relative Cq values. The results demonstrated that the smallest Cq value and higher fluorescence intensity were achieved when using a primer concentration of 0.4 μmol/L and a probe concentration of 0.2 μmol/L; there was no significant difference compared to amplification with high-concentration primer-probe pairs (Supplementary Figure S1). Considering its consistency with the internal standard gene zSSIIb and minimal impact on annealing temperature in real-time fluorescent PCR reactions, this method adopted the conventional qPCR reaction procedure.

3.2 Specific test of CC-2 event-specific PCR method

The specific test of the CC-2 primer combination CC-2-uQF/uQR/uQP was conducted on different crops and their main commercial events. The samples consisted of 6 mixed samples of common GM crops and 6 mixed samples of non-GM maize. Supplementary Table S1 provides detailed information on the events and results. The obtained results demonstrated that only the CC-2 sample exhibited the expected amplification curve, while other samples did not show such curves. These findings indicate that the method exhibits excellent specificity for detecting the CC-2 maize event.

3.3 LOD

The concentration of homozygous genomic DNA template was quantified and the copy numbers were calculated based on an estimated haploid maize genome size of 2500 Mbp (Arumuganathan and Earle, 1991). Six samples of the homozygous genomic DNA template were serially diluted to achieve copy number concentrations of 40, 20, 10, 5, 2.5 and 0.5 copies/μL. Subsequently, 2 μL test samples were introduced into the system. Utilizing a total DNA template amount of 100 ng, the mass-percentage of CC-2 event corresponded to approximately 0.20%, 0.10%, 0.05%, 0.25%, 0.125% and 0.025%. Each dilution was subjected to analysis in 12 parallel reactions to identified LOD.

The amplification test results are presented in Figure 2, and the statistical findings are summarized in Table 2. The data indicate that at a template concentration of 0.5 copies/μL, corresponding to 1 copy of substrate in the reaction system, 5 out of 12 parallel reactions yielded positive results, validating the accuracy of the template concentration dilution. At a template volume of 5 copies, 9 out of 12 parallel reactions were positive. When using templates with 2 copies, 7 out of 12 parallel reactions produced a positive result, but mean copy number was not available. Notably, typical amplification curves were obtained for all twelve parallel reactions when utilizing templates with copy numbers ranging from 10 to 40 copies. These observations suggest that the sample containing a fraction as low as 0.05% of CC-2 event (equivalent to 20 copies of CC-2 maize) could be reliably detected.

3.4 Dynamic range

The concentration of genomic DNA was diluted to 50 ng/μl, followed by gradient dilution with water or 0.1×TE buffer to prepare multiple calibrators with varying contents. These calibrators were utilized as templates for real-time fluorescent PCR reactions, with the amount of DNA template in the PCR system being 2 μL. The corresponding DNA quality and copy number were then determined and are presented in Table 3. Standard curves for CC-2 and *zSSIIb* genes in maize were plotted based on the Cq value of PCR reaction of standard DNA solution and the logarithm of initial template copy number. The results indicated that when the calibrator ranged from 40 to 40000 copies, the slopes of the standard curves for CC-2 and *zSSIIb* in maize were -3.403 and -3.342 respectively (Figure 3). The coefficient of determination R^2 was 0.999 for both cases, exceeding the minimum acceptable value of 0.98. The PCR amplification efficiencies were 96.70% and 99.20%, respectively, falling within the permissible range of 90% to 110%. More than three replicates were conducted, and all data parameters of standard curves met the requirements for

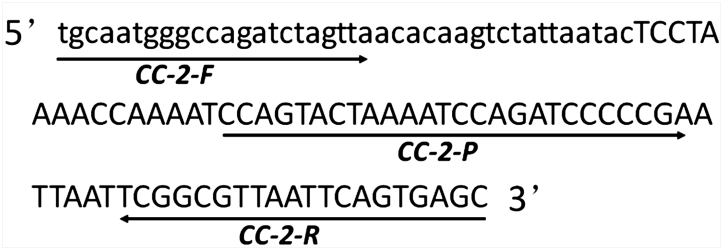


FIGURE 1
Amplicon sequence of GM maize CC-2 event-specific PCR. The junction fragment of CC-2 event covered 40 bp of flanking genomic DNA in lowercase letters and 67 bp exogenous insertion fragment in uppercase letters and the sequence of forward primer CC-2-F, reverse primer CC-2-R and probe CC-2-P were underlined.

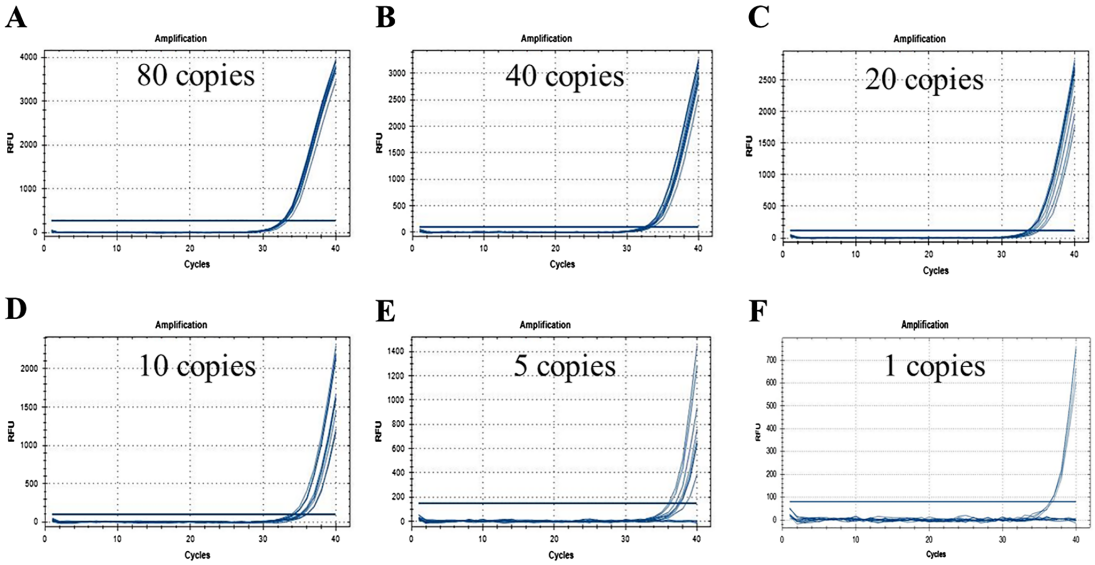


FIGURE 2
LOD test of CC-2 event specific real-time PCR method. is the amplification of 10 replicates when the amount of substrate in the reaction system is 80 **(A)**, 40 **(B)**, 20 **(C)**, 10 **(D)**, 5 **(E)** and 1 **(F)** copy, respectively.

quantitative GM detection methods. The PCR reaction system exhibited a good linear relationship between the PCR Cq value and the copy number detection of CC-2 specific fragment.

3.5 LOQ

The initial determination of the limit of quantification (LOQ) is established based on the concentration range of the low content calibrator as measured by the limit of detection (LOD) (see Table 2). The findings indicate that only when the substrate concentration in the sample reaches 40 copies, both the relative bias and relative standard deviation are ≤25%, falling within an acceptable range. Therefore, it can be concluded that a minimum substrate amount of 40 copies is required for accurate determination.

For the samples with the copy number fraction of 0.1% of CC-2 event that passed the preliminary test, a total of 60 quantitative tests were conducted to calculate the relative bias (biasR) and relative standard deviation (RSD) of the detection data (Supplementary Figure S2). The results showed that both the relative bias and relative standard deviation for all test samples fell within ±25% of the acceptable range, indicating that the quantitative limit of CC-2 fluorescent quantitative PCR method could be determined to reach 0.1%, thus meeting EU standards requirements (European Network of GMO Laboratories (ENGL), 2015).

3.6 Quantification of blind samples by real-time quantitative PCR

Samples containing genomic DNA with copy number fractions of 5%, 1%, and 0.1% of CC-2 were subjected to testing, with three parallel samples set for each level and the experiment repeated thrice. The proportion of CC-2 DNA to total maize DNA (%) was computed as (mean copies of GM maize of three parallel assays)/(mean copies of total maize DNA of three parallel assays)×100% (Kuribara et al., 2002). Table 4 presents results indicating values of 5.23%, 1.00%, and 0.10% for the three respective samples. The

TABLE 2 LOD and LOQ of CC-2 event-specific realtime PCR method.

Amount of DNA (copies/ reaction)	Signal rate (positive signals)	Mean Copy Number	SD of the Copy Number	RSD(%)	Bias(%)
40	12/12	36.31	7.3	23.10	-9.22
20	12/12	23.70	9.6	31.14	18.50
10	12/12	6.81	4.2	42.77	31.90
5	9/12	3.8	[b]	[b]	[b]
2	7/12	[b]	[b]	[b]	[b]
1	5/12	[b]	[b]	[b]	[b]

[b]Not available.

10.3389/fpls.2024.1460038

TABLE 3 Repeatability of real-time PCR assays employing CC-2 DNA as reference.

DNA amount (ng)	CC-2 copy number[a]	Repeat	Cq			Mean of Cq Values	SD_r	RSD_r (%)[c]	Mean of all Cq Values	SD_R	RSD_R (%)[c]
			1	2	3						
100	40000	1	24.24	24.31	24.22	24.26	0.05	0.19	24.39	0.11	0.45
		2	24.41	24.47	24.39	24.42	0.04	0.17			
		3	24.5	24.53	24.4	24.48	0.07	0.28			
20	8000	1	26.38	26.49	26.62	26.50	0.12	0.45	26.65	0.15	0.56
		2	26.67	26.58	26.68	26.64	0.06	0.21			
		3	26.8	26.86	26.73	26.80	0.07	0.24			
4	1600	1	29.02	28.97	28.96	28.98	0.03	0.11	29.16	0.17	0.59
		2	29.16	29.16	29.18	29.17	0.01	0.04			
		3	29.27	29.51	29.24	29.34	0.15	0.50			
0.8	320	1	31.33	31.27	31.28	31.29	0.03	0.10	31.47	0.16	0.50
		2	31.57	31.47	31.55	31.53	0.05	0.17			
		3	31.53	31.49	31.75	31.59	0.14	0.44			
0.1	40	1	34.41	34.76	34.16	34.44	0.30	0.88	34.65	0.49	1.41
		2	35.17	34.29	34.79	34.75	0.44	1.27			
		3	34.54	34.14	35.61	34.76	0.76	2.19			
0.0125	5	1	36.25	[b]	36.01	36.96	[b]	[b]	36.91	0.97	2.62
		2	[b]	37.32	35.67	38.45	[b]	[b]			
		3	37.30	38.27	35.89	37.15	0.72	1.20			
0.0025	1	1	[b]	37.59	37.29	38.4	[b]	[b]	37.73	0.67	1.77
		2	[b]	[b]	38.71	38.71	[b]	[b]			
		3	37.33	[b]	[b]	37.33	[b]	[b]			

[a]Calculated based on a haploid maize genome size of 2500 Mbp.
[b]Not available.
[c]RSDr, Repeatability relative standard deviation; RSD_R, Reproducibility relative standard deviation.

relative bias (biasR) between the measured total average value and the nominal value of the three subsamples ranges from 0.37% to 4.51%, falling within the acceptable range of ±25%, indicating that the measurement results obtained using the CC-2 quantitative PCR method exhibit good accuracy. The relative standard deviation (RSD_r) for repeatability of the three horizontal samples ranged from 1.06% to 14.38%, all of which were below the specified threshold of 25%. The results of repeatability analysis demonstrate that the CC-2 converter fluorescence quantitative PCR method exhibits excellent precision.

3.7 Collaborative validation of the quantitative PCR method for GM event CC-2 detection

The new qualitative detection concept would be useful for ensuring robust and reproducible results among laboratories, particularly for detecting low-copy-number DNA samples. Samples of different CC-2 content were performed an interlaboratory evaluation of the developed quantitative method as a blind test performed by 8 laboratories. Blind samples with the CC-2 content levels of 5%, 2%, 1%, 0.5% and 0.1% were prepared and provided to measure in each lab. Samples of one content has 3 sub-samples.

As the result submitted of 8 labs shown (Supplementary Table S2; Figure S3), the slope of standard curves of CC-2 specific and *zSSIIb* gene in eight laboratories ranged from -42.63 to -38.29 and -42.95 to -38.091, respectively. The determination coefficients R^2 ranged from 0.995 to 1.000 and from 0.992 to 1.000, respectively, PCR amplification efficiency ranged from 91.70 to 106.14% and from 90.09 to 101.4%, respectively. All the tests of the standard curves were within the acceptable range.

For the different content samples measured data, Cochran's test ($p<0.025$) and Grubb's test ($p<0.025$) were carried according to the harmonized guidelines of AOAC to remove the outlier data and analyze the validated result. The results show that there are no outliers or deviations in 8 laboratory data. the data of 15 samples with 3 different content from 8 laboratories were statistically summarized, and the average value from 8 laboratories was

10.3389/fpls.2024.1460038

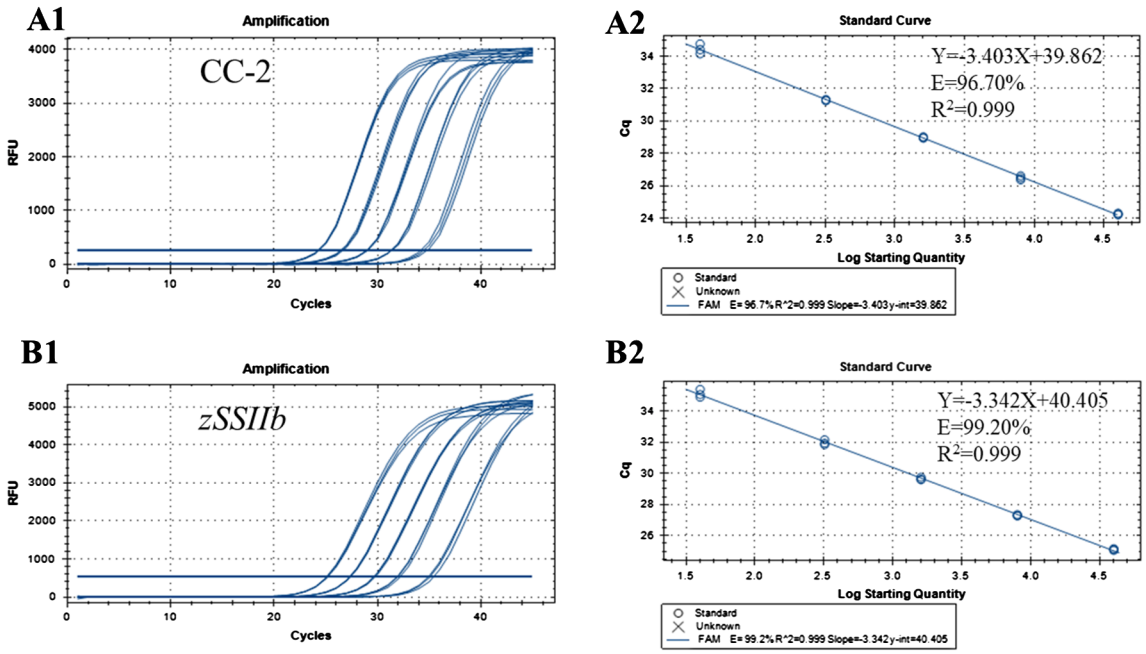


FIGURE 3
Amplification and standard curves for the event-specific quantitative PCR method using gradient-diluted CC-2 genomic DNA as the template analyzed using CFX96 System. **(A1)** Amplification graph for the CC-2 event-specific assay. **(A2)** Standard curve for the CC-2 event-specific assay. The copy numbers of the CC-2 event in each dilution were 40000,10000,1000,200 and 20 copies per reaction, respectively. **(B1)** Amplification graph for the endogenous gene *zSSIIb* assay. **(B2)** Standard curve for the gene *zSSIIb* assay. The quantities of maize genome in each dilution were 40000, 10000, 1000, 200, and 20 copies per reaction, respectively.

calculated (Table 5). The average values of 5 content samples in laboratories were 4.81%, 2.18%, 0.96%, 0.51% and 0.095%, respectively. The quantitative values deviated slightly from the expected values for all tested samples with bias (%) ranging from -5.0% and 9.0%. In the ENGL method acceptance criterion, the trueness should be within ±25% (European Network of GMO Laboratories (ENGL), 2015). It indicated that the CC-2 PCR specific method in quantitative measurement was credible.

After that, we conducted further statistical analyses for the values of quantification. The trueness and precision were determined as previously described. The mean, bias, repeatability of RSD (RSD_r) and reproducibility of RSD (RSD_R) of blind samples were measured (Table 6). The RSD_r values for samples were 11.43%, 8.44%, 15.73%, 3.38% and 8.75%, respectively; all RSD^r values were below 25%. The RSD_R values were within the range of 3.23% to 12.19%, all below 35% across the entire dynamic range. Both the RSD of test samples were similar to or within a narrower range than those in previously reported method of GMO events (Wang et al., 2013; Mano et al., 2013; Jacchia et al., 2015). The repeatability and reproducibility of the method meet the acceptance criteria and performance requirements (European Network of GMO Laboratories (ENGL), 2015), indicating that the CC-2 specific method is stable, reliable, and suitable for quantifying CC-2. The analysis results demonstrate that the established event-specific real-time PCR system for CC-2 can generate accurate, repeatable, and comparable results across different laboratories.

TABLE 4 Trueness and precision data for the CC-2 event-specific realtime PCR method.

Theoretical contents	Assay	Experimental (copies)			Mean(copies)	RSD_r (%)	Experimental (%)	Bias (%)	RSD_R (%)
		1	2	3					
5.00%	CC-2	2001	2093	1944	2013	3.74	5.23	4.51	4.97
	zSSIIb	36597	39867	39237	38567	4.50			
1.00%	CC-2	371	389	411	390	5.13	1.00	0.37	1.06
	zSSIIb	36483	39197	41050	38910	5.90			
0.10%	CC-2	41	41	31	38	15.33	0.10	2.44	14.38
	zSSIIb	35763	37867	35693	36441	3.39			

RSD_r, Repeatability relative standard deviation; RSD_r, Reproducibility relative standard deviation.

10.3389/fpls.2024.1460038

TABLE 5 Determined GM% values of the eight participants for the five unknown samples.

Labs		GMO content (GM% = GM copy number/genome copy number × 100)(%)				
		Level 1	Level 2	Level 3	Level 4	Level 5
Lab 1	Rep 1	3.90	2.26	0.66	0.49	0.09
	Rep 2	4.05	2.44	0.71	0.52	0.08
	Rep 3	4.06	2.27	0.71	0.51	0.08
Lab 2	Rep 1	4.67	2.49	0.84	0.38	0.08
	Rep 2	4.45	2.31	0.81	0.6	0.07
	Rep 3	4.29	2.45	0.85	0.51	0.08
Lab 3	Rep 1	5.15	2.06	0.97	0.49	0.10
	Rep 2	5.24	2.10	1.00	0.5	0.10
	Rep 3	5.07	1.99	0.95	0.5	0.09
Lab 4	Rep 1	4.92	2.02	0.88	0.53	0.10
	Rep 2	4.99	2.07	0.87	0.52	0.11
	Rep 3	4.93	2.00	0.93	0.5	0.10
Lab 5	Rep 1	5.66	1.83	1.17	0.47	0.11
	Rep 2	5.90	1.99	1.08	0.50	0.10
	Rep 3	5.53	1.93	1.13	0.48	0.10
Lab 6	Rep 1	5.40	2.37	1.09	0.55	0.12
	Rep 2	4.29	2.30	1.13	0.51	0.10
	Rep 3	4.43	2.47	1.14	0.48	0.08
Lab 7	Rep 1	4.96	2.03	1.02	0.53	0.10
	Rep 2	4.59	2.01	0.97	0.53	0.10
	Rep 3	5.00	2.02	1.07	0.54	0.10
Lab 8	Rep 1	4.53	2.41	0.961	0.53	0.11
	Rep 2	4.51	2.35	1.036	0.47	0.11
	Rep 3	4.97	2.28	1.123	0.55	0.10
Mean value		**4.81**	**2.18**	**0.96**	**0.51**	**0.095**

TABLE 6 Summary of validation results for the CC-2-specific method.

blind Samples	Expected value(%)				
	5.0	2.0	1.0	0.5	0.1
Mean value	4.81	2.18	0.96	0.51	0.095
repeatability standard deviation *Sr*	0.558	0.179	0.151	0.017	0.009
relative repeatability standard deviation, RSD_r (%)	11.43	8.44	15.73	3.38	8.75
reproducibility standard deviation, S_R	0.59	0.18	0.15	0.017	0.009
relative reproducibility standard deviation, RSD_R (%)	12.19	8.50	15.81	3.23	8.77
bias (absolute value)	-0.19	0.18	-0.04	0.01	-0.005
biasR (%)	-3.80	9.00	-4.00	2.00	-5.00

3.8 Measurement uncertainty of the tested results

According to the guidance documents (Trapmann et al., 2007; Standardization ISO/TS 21748:2004), an estimation of measurement uncertainty (MU) was conducted for the quantitative results. This can be achieved by plotting a chart correlating repeatability standard deviation (sR value) in collaborative trials with the average number of blind samples tested (c), and calculating linear regression to estimate absolute standard uncertainty (u0) and relative standard uncertainty (RSU). The value of u0 is a constant equal to the intercept of linear regression (u0 = 0.0037), while RSU is equal to the slope of linear regression (RSU = 0.0384). The critical value (LC = 2 × u0) corresponds to a measurement result of 0.1% CC-2, indicating that if the estimated value is below 0.1%, it can be concluded with a confidence level of 95% that target CC-2 does not exist in

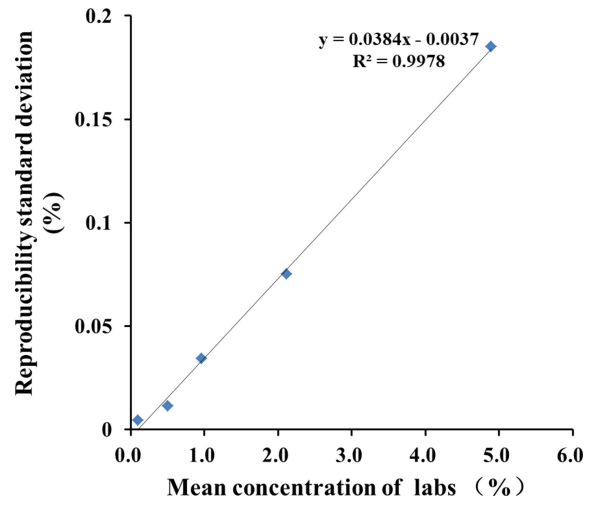


FIGURE 4
Linear regression produced by plotting mean measurement concentration (c) against reproducibility standard deviation (sR).

the tested sample. The standard uncertainty associated with measurement results (u) is calculated using formula $u = (0.0037^2 + (0.0384 \times c)^2)^{1/2}$) (Figure 4). Typically, when reporting test results, an accompanying measurement uncertainty is provided as expanded uncertainty (U= 2 × u), which is derived from standard uncertainty using a coverage factor of 2.This corresponds approximately to a confidence level of about 95%. For blind samples S1-S5, respective values for c in measurement results are 5.083%, 2.071%, 1.049%, 0.503% and 0.102% (w/w).The U values for expanded uncertainties were calculated as follows: S1-0.144%, S2- 0.059%, S3-0.030%, S4-0.013% and S5-0.007%(w/w). Here, the uncertainty formula provides a closely approximate representation of the true distribution when laboratories utilize the CC-2 quantitative PCR method, offering a suitable level of precision for scientific research and analysis.

4 Conclusion

In this study, a novel real-time PCR-based analytical method was developed for the event-specific quantification of a genetically modified (GM) maize event CC-2. The specificity, sensitivity of these methods were determined with different concentrations of GM mixing samples. The LODs of these methods for CC-2 segment calculated as the amount of CC-2 were 0.05% or less. The limit of quantitation for the method was estimated to be 0.1% indicating that the LOQ of CC-2 was lower than 40 copies of maize haploid genome. The quantitative method was evaluated by means of blind tests in multi-laboratory trials. The trueness and precision were evaluated as the bias and reproducibility of relative standard deviation (RSD), and the determined bias and RSD values for the method were each less than 25%. These results suggest that the developed method would be suitable for practical analyses for the detection and quantification of CC-2. Furthermore, The uncertainty evaluation equation of the CC-2 method was established by the results from inter-laboratory verification to model the uncertainty arising from the relative repeatability standard deviation of inter-laboratory test values.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

LL: Data curation, Methodology, Writing – original draft. NZ: Investigation, Visualization, Writing – original draft. ZX: Investigation, Visualization, Writing – original draft. YH: Investigation, Methodology, Writing – original draft. WY: Validation, Writing – original draft. LD: Validation, Writing – original draft. YM: Supervision, Writing – original draft. CL: Supervision, Writing – original draft. YX: Resources, Writing – original draft. NL: Writing – review & editing. WX: Resources, Writing – original draft. FL: Conceptualization, Methodology, Software, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Biological Breeding - Major Project (grant no. 2022ZX0402010); Jilin Scientific and Technological Development Program, China (grant no. 20230508091RC); Agricultural Quality Standards and Testing Technology Institute Innovation fund Project (grant no. CXGC2023SJ107).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2024.1460038/full#supplementary-material

References

Akinbo, O., Obukosia, S., Ouedraogo, J., Sinebo, W., Savadogo, M., Timpo, S., et al. (2021). Commercial release of genetically modified crops in Africa: interface between biosafety regulatory systems and varietal release systems. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.605937

Arumuganathan, K., and Earle, E. D. (1991). Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* 9(3), 208–219. doi: 10.1007/BF02672069

Avsar, B., Sadeghi, S., Turkec, A., and Lucas, S. J. (2020). Identification and quantitation of genetically modified (GM) ingredients in maize, rice, soybean and wheat-containing retail foods and feeds in Turkey. *J. Food Sci. Technol.* 57, 787–793. doi: 10.1007/s13197-019-04080-2

Carzoli, A. K., Aboobucker, S. I., Sandall, L. L., Lübberstedt, T. T., and Suza, W. P. (2018). Risks and opportunities of GM crops: Bt maize example. *Global Food Secur.* 19, 84–91. doi: 10.1016/j.gfs.2018.10.004

Chen, L., Zhong, R., Zhang, L., and Zhang, H. (2019). The Chronic Effect of Transgenic Maize Line with mCry1Ac or maroACC Gene on Ileal Microbiota Using a Hen Model. *Microorganisms* 7, 92. doi: 10.3390/microorganisms7030092

European Commission Regulation(EC)No 1830/2003 (2003). Concerning the traceability and labelling of genetically modified organisms and the traceability of food and feed products produced from genetically modified organisms and amending Directive 2001/18/EC 268, 24–28.

European Network of GMO Laboratories (ENGL) (2015). *Definition of Minimum Performance Requirements for Analytical Methods of GMO Testing*. (European Network of GMO Laboratories (ENGL)). doi: 10.13140/RG.2.1.2060.5608

Fan, C. M., Wang, B. F., and Song, X. Y. (2018). Effects of transgenic herbicide-resistant corn CC-2 with EPSPS gene cultivation on soil Collembola. *J. Agro-Environment Sci.* 37, 1203–1210. doi: 10.11654/jaes.2017-1423

Fraiture, M. A., Gobbo, A., Marchesi, U., Verginelli, D., Papazova, N., and Roosens, N. H. C. (2021). Development of a real-time PCR marker targeting a new unauthorized genetically modified microorganism producing protease identified by DNA walking. *Int. J. Food Microbiol.* 354, 109330. doi: 10.1016/j.ijfoodmicro.2021.109330

Giraldo, P. A., Shinozuka, H., Spangenberg, G. C., Cogan, N. O. I., and Smith, K. F. (2019). Safety assessment of genetically modified feed: is there any difference from food? *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01592

GM AgbioInvestor Monitor (2024). *Global GM Crop Area 2023 Review*. Available online at: https://gm.agbioinvestor.com. (Accessed date June 15, 2024).

Gruère, G. P., and Rao, S. R. (2007). A review of international labeling policies of genetically modified food to evaluate India's proposed rule. *Agbioforum* 10, 51–64.

Guertler, P., Grohmann, L., Naumann, H., Pavlovic, M., and Busch, U. (2019). Development of event-specific qPCR detection methods for genetically modified alfalfa events J101, J163 and KK179. *Biomol Detect Quantif* 17, 100076. doi: 10.1016/j.bdq.2018.12.001

Hougs, L., Gatto, F., Goerlich, O., Grohmann, L., Lieske, K., Mazzara, M., et al. (2017). *Verification of analytical methods for GMO testing when implementing interlaboratory validated methods* (Luxembourg: EUR 29015 EN, Publication Office of the European Union), ISBN: . JRC 109940. doi: 10.2760/645114

ISAAA (2021). *Global status of commercialized biotech/GM crops in 2019: Biotech Crops Drive Socio-Economic Development and Sustainable Environment in the New Frontier* (Ithaca, NY: ISAAA Brief No.55. ISAAA). Available at: https://www.isaaa.org/resources/publications/briefs/55/default.asp.

Jacchia, S., Nardini, E., Savini, C., Petrillo, M., Angers-Loustau, A., Shim, J.-H., et al. (2015). Development, optimization, and single laboratory validation of an event-specific real-time PCR method for the detection and quantification of golden rice 2 using a novel taxon-specific assay. *J. Agric. Food Chem.* 63, 1711–1721. doi: 10.1021/jf505516y

Li, Y., Hallerman, E. M., and Peng, Y. (2020). Excessive Chinese concerns over genetically engineered food safety are unjustified. *Nat. Plants* 6, 590–590. doi: 10.1038/s41477-020-0685-4

Li, Z., Li, X., Wang, C., Song, G., Pi, L., Zheng, L., et al. (2017). One novel multiple-target plasmid reference molecule targeting eight genetically modified canola events for genetically modified canola detection. *J. Of Agric. And Food Chem.* 65, 8489–8500. doi: 10.1021/acs.jafc.7b02453

Li, Y., Xiao, F., Zhai, C., Li, X., Wu, Y., Gao, H., et al. (2022). Qualitative and quantitative real-time PCR methods for assessing false-positive rates in genetically modified organisms based on the microbial-infection-linked HPT gene. *Int. J. Mol. Sci.* 23, 10000. doi: 10.3390/ijms231710000

Mano, J., Masubuchi, T., Hatano, S., Futo, S., Koiwa, T., Minegishi, Y., et al. (2013). Development and validation of event-specific quantitative PCR method for genetically modified maize LY038. *Journa Food Hygienic Soc. Japan* 54, 25–30. doi: 10.3358/shokueishi.54.25

Meng, Y., Wang, S., Guo, J., and Yang, L. (2022). Collaborative ring trial of the applicability of a reference plasmid DNA Calibrant in the quantitative analysis of GM maize event MON810. *Foods* 11, 1538. doi: 10.3390/foods11111538

Trapmann, S., Burns, M., Broll, H., Macarthur, R., Wood, R., and Zel, J. (2007). *Guidance Document on Measurement Uncertainty for GMO Testing Laboratories. EUR 22756 EN. 2007. JRC37201*. Available online at: https://publications.jrc.ec.europa.eu/repository/handle/JRC37201. (Accessed date June 01, 2024).

Turkec, A., Lucas, S. J., and Karlık, E. (2015). Monitoring the prevalence of genetically modified maize in commercial animal feeds and food products in Turkey. *J. Sci. Food Agric.* 96, 3173–3179. doi: 10.1002/jsfa.7496

Wang, H., Qian, C., Su, C., Duan, Y., and Bai, H. (2013). Rapid real-time PCR detection of transgenic cry1C rice using plasmid molecule as calibrator. *Eur. Food Res. Technol.* 237, 101–107. doi: 10.1007/s00217-013-1957-2

Yang, L., Xu, S., Pan, A., Yin, C., Zhang, K., Wang, Z., et al. (2005). Event specific qualitative and quantitative polymerase chain reaction detection of genetically modified MON863 maize based on the 5'-transgene integration sequence. *J. Of Agric. And Food Chem.* 53, 9312–9318. doi: 10.1021/jf051782o

Yi, H., Liang, Z., Ge, J., Zhang, H., Liu, F., Ren, X., et al. (2022). A multiplex PCR system for the screening of genetically modified (GM) maize and the detection of 29 GM maize events based on capillary electrophoresis. *Agriculture* 12, 413. doi: 10.3390/agriculture12030413

frontiers | Frontiers in Plant Science

TYPE Review
PUBLISHED 13 September 2024
DOI 10.3389/fpls.2024.1411859

OPEN ACCESS

EDITED BY
Dmitri Voronine,
University of South Florida, United States

REVIEWED BY
Stefan James Hill,
New Zealand Forest Research Institute Limited (Scion), New Zealand
Bo-Fang Yan,
Guangdong Academy of Agricultural Sciences (GDAAS), China
Moisés Roberto Vallejo-Pérez,
Autonomous University of San Luis Potosí, Mexico

*CORRESPONDENCE
Dmitry Kurouski
dkurouski@tamu.edu

RECEIVED 03 April 2024
ACCEPTED 08 August 2024
PUBLISHED 13 September 2024

CITATION
Juárez ID and Kurouski D (2024) Contemporary applications of vibrational spectroscopy in plant stresses and phenotyping.
Front. Plant Sci. 15:1411859.
doi: 10.3389/fpls.2024.1411859

Contemporary applications of vibrational spectroscopy in plant stresses and phenotyping

Isaac D. Juárez[1,2] and Dmitry Kurouski[1,2]*

[1]Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX, United States, [2]Interdisciplinary Faculty of Toxicology, Texas A&M University, College Station, TX, United States

Plant pathogens, including viruses, bacteria, and fungi, cause massive crop losses around the world. Abiotic stresses, such as drought, salinity and nutritional deficiencies are even more detrimental. Timely diagnostics of plant diseases and abiotic stresses can be used to provide site- and doze-specific treatment of plants. In addition to the direct economic impact, this "smart agriculture" can help minimizing the effect of farming on the environment. Mounting evidence demonstrates that vibrational spectroscopy, which includes Raman (RS) and infrared spectroscopies (IR), can be used to detect and identify biotic and abiotic stresses in plants. These findings indicate that RS and IR can be used for in-field surveillance of the plant health. Surface-enhanced RS (SERS) has also been used for direct detection of plant stressors, offering advantages over traditional spectroscopies. Finally, all three of these technologies have applications in phenotyping and studying composition of crops. Such non-invasive, non-destructive, and chemical-free diagnostics is set to revolutionize crop agriculture globally. This review critically discusses the most recent findings of RS-based sensing of biotic and abiotic stresses, as well as the use of RS for nutritional analysis of foods.

KEYWORDS

digital farming, non-invasive phenotyping, nutrient content assessment, plant disease diagnostics, Raman spectroscopy, optical sensing, infrared spectroscopy, surface enhanced Raman spectroscopy

1 Introduction

Most of economically important plants, such as corn, wheat and rice, can be infected with a large number of pathogens. Although a progression of plant diseases directly depends on the pathogen and weather conditions, in most cases, infected plants will decay within several weeks. This process can be decelerated if plant protection chemistry is timely utilized. Satellite- or drone-based RGB imaging can be used to identify such problem areas. However, these techniques lack specificity since the diagnostics is based on the color change. Both of these RGB images are laborious and expensive. These and other factors

largely limit their broad application in modern farming. To overcome the lack of specificity, several molecular methods, such as PCR and ELISA, can be used. These methods directly rely on the presence of the pathogen in the analyzed sample. Although provide very high sensitivity and specificity, both methods are highly laborious, which limits cost-per-sample minimization. On average, one ELISA sample cost around $15, whereas one PCR test is ~$25. The major drawback of both methods is false-negative outcomes in the case of a lack of a pathogen or a pathogen nucleic acid in the sample. For instance, PCR can be efficiently used to diagnose citrus greening disease, also known as Huanglongbing (HLB) (Sanchez et al., 2019b). This disease is caused by bacteria that infect citrus trees. Infected trees exhibit chlorosis and premature fruit drop. Since the bacteria are vectored by psyllids, leaves in once tree branch may possess the pathogen, whereas leaves of the next branch on the same tree will be pathogen-free. In the former case, PCR provides a confirmatory pathogen identification. However, in the latter case, false-positive results will be delivered by this molecular assay. Furthermore, in hot summer seasons, bacteria move to the stem and roots of the trees. Consequently, analysis of plant leaves in these months will indicate falls pathogen-free status of the plants.

Abiotic stresses, such as drought, salinity, and heat, are far more detrimental to the crop yield. On average, these stresses are accountable for ~70% of the crop losses worldwide. Their diagnostic is far more challenging than the detection and identification of biotic stresses. Primarily because both PCR and ELISA cannot be used in such cases. RGB-based imaging is also limited because visual symptoms of biotic and abiotic stresses are very similar. Traditionally, induced coupled plasma mass spectroscopy (ICP-MS) is used to quantify macro- and micronutrients in both soil and plants. This information can be used to alter the dosage of plant fertilizers to mitigate abiotic stresses caused by the lack of nutrients. ICP-MS can be also used to probe plant contamination with heavy and toxic metals, such as lead and arsenic. However, ICP-MS is not portable, which requires sample shipment to analytical laboratories. This technique is also laborious and expensive.

One can expect that timely diagnostics of both biotic and abiotic stresses require new techniques that must be 1) unexpensive; 2) portable, 3) fast, and 4) accurate. During the past years, a growing number of studies demonstrated that both IR and RS fit these strict requirements. IR and RS are label-free methods that use light to probe the chemical structure of analyzed samples. Therefore, the direct cost of both IR- and RS-based analyses is zero. Several companies came to the market with excellent hand-held IR and RS instruments, Figure 1. Although their costs remain high ($20,000-$70,000), these instruments are easy to use. Furthermore, the direct time of spectral acquisition is typically around 1-2 s. In most cases, 10-20 s are required to process the data. Since most of the spectrometers are equipped with a display and a chemometric algorithms, the researcher can see the outcome of the spectral analysis within 15-25 s. The question remains unclear is whether such instruments can be used for an accurate, robust and reliable diagnostics of biotic and abiotic stresses in plants.

FIGURE 1
Commercially available hand-held Raman spectrometer with 830 nm excitations (top) and a bench-top home-built confocal Raman microscope (bottom).

Robust and reliable plant phenotyping is highly important for plant breeders. Currently, such expertise requires years of training and experience. Since different plant species and plant varieties have distinctly different biochemical profiles, one can expect that RS could be used to detect these biochemical differences and, consequently, assist in plant breading.

This review will critically discuss the most recent reports on IR- and RS-based diagnostics of fungal, viral, and bacterial diseases, as well as on the use of both techniques for the quality control of fruits and vegetables. We also briefly discuss the most recent advances in RS-based plant breeding, prediction of the optimal harvest date and genotyping.

2 Instrumentation and imaging approaches

2.1 Raman spectroscopy

Both hand-held and bench-top Raman spectrometers share a very similar engineering concept. In most cases, continuous wavelength (CW) lasers are used to generate light that is directed

towards a beam splitter by a set of mirrors, Figure 2A. Next, the light is focused on the sample either by a simple achromatic lens or by a microscope objective. The scattered light is collected by the same optical setup; Long-pass filter is then used to cut off elastically scattered light, whereas inelastically scattered photons are directed towards a spectrograph where they are split on a grating based on their energies. Finally, a CCD camera is used to collect the inelastically scattered photons. Any Raman spectrometer has several critically important parameters such as laser power, excitation wavelength, laser spot size and a spectral resolution. The first and second parameters are determined by the laser. Although currently available lasers can generate light with pretty much any wavelength ranging from deep UV (~190 nm) to far IR (1064 nm), electromagnetic radiation in the visible areas of the light spectrum is most commonly used. Our group showed that the use of blue and green light in Raman spectroscopy provided the advantage in the detection of carotenoids due to the resonance Raman effect. Plants possess ~20 different carotenoids with have slightly different Raman spectra. Thus, Raman spectrometers with blue and green laser sources provide the advantage of sensing these biological molecules in the plant leaves. We also showed that yellow and red lasers are not useful in optical sensing of plants due to a high chlorophyll fluorescence in this part of electromagnetic spectrum. At the same time, near IR lasers (785 nm, 830 nm and 1064 nm) can be used to overcome this limitation. Utilization of these lasers allows for sensing of a large number of biological molecules in plant samples. It should be noted that silicon-based CCD cameras cannot be used for collection of phonons with wavelength above 1.4 eV (880 nm). To overcome this limitation, InGaAs CCDs are used in the instruments with 1064 nm excitation. However, photon-to-electron conversion on these CCDs is not as good as in silicon-based CCDs. These heterostructure-based detectors also have much greater dark current noise, which results in much more nosier spectra compared to those acquired on silicon-based CCDs.

2.2 Surface-enhanced Raman spectroscopy

SERS is based on a phenomenon of strong (10^6-10^8) amplification of Raman scattering by metal nanostructures. If illuminated by light at or around their absorption maxima, metal nanostructures exhibit coherent oscillations of conductive electrons, also known as localized surface plasmon resonances (LSPRs). LSPRs enhance Raman scattering from molecules located in the close proximity to the metal surfaces allowing for single-molecule detection. Over the past decade, a large number of synthetic appraisals have been reported which could be used to fabricate nanostructures with the desired optical properties. Furthermore, nanostructures can be decorated with molecular analytes, also known as capture layer, to enable targeted sensing of the molecular species of interest. Although silver nanoparticles exhibit high cell toxicity, gold nanoparticles are proudly utilized for SERS sensing in life systems. Several groups demonstrated that utilization of copper or magnesium for the nanostructure fabrication allows for minimization of costs of the final product of syntheses.

2.3 Infrared spectroscopy

Rapid development of mid-IR and QCL lasers allowed for a substantial minimization of the IR spectrometers. In this instruments, IR light (3-13 μm) is directed to the interferometer to enable a Fourier transformed spectral acquisition. Next, the IR light is directed towards the sample (transmittance IR) or a crystal (attenuated total reflectance (ATR-IR). Finally, a MCT detector is used to collect the IR light. In the case of transmittance IR instruments, calcium fluoride or similar substrates are used as a sample support. In ATR-based IR systems, the sample can be directly pressed against the crystal for spectral analysis. If transmittance IR instruments require transparent or translucent samples, ATR-IR based setups can be used to analyze any type of the sample. ATR modality also allows for the development of hand-held IR instruments that can be used directly in the field. It should be noted that IR can be also coupled with atomic force microscopy (AFM). This instrumental setup is known as AFM-IR or photothermal IR. AFM-IR offers ~2-4 nm spatial resolution, which can be highly beneficial in the analysis of biological molecules, and single molecule sensitivity which cannot be achieved using conventional IR instruments.

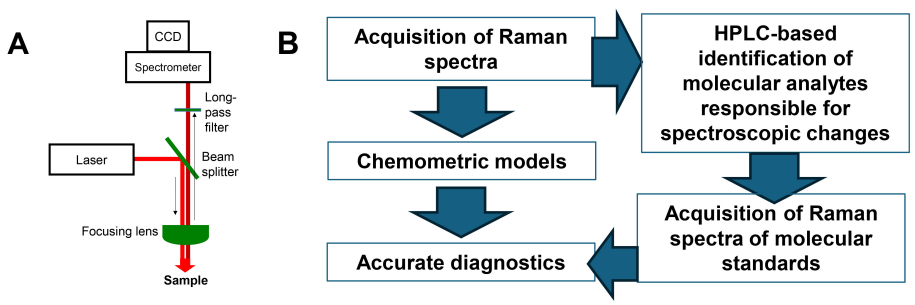


FIGURE 2
Schematic representation of **(A)** Raman spectrometer and **(B)** workflow of used to develop robust and reliable diagnostics of biotic and abiotic stresses in plants.

It is important to emphasize that a term "near-IR spectroscopy" is commonly used to describe electronic absorbance or reflectance spectroscopy. In this case, electronic rather than vibronic properties of samples are probed.

2.4 Spectral analysis and interpretation

In the X axis of Raman spectra, a "Raman Shift" is used to describe the energy change between the incident and acquired inelastically scattered photons. Therefore, Raman spectra collected with different excitations will have the same spectra (except resonance Raman). Vibrations of a vast majority of biological molecules are in 300-1800 cm^{-1}, as well as 2,200-3,500 cm^{-1}. Since 2,200-3,500 cm^{-1} spectral window is primarily dominated by CH, CH_2 and OH vibrations, most of the reported Raman spectra are within 300-1800 cm^{-1}. The Y axis reflects the intensity of inelastically scattered photons. Spectral intensity is primarily dependent on the laser power and spectral acquisition time. Therefore, intensity in counts over laser power (mW) and seconds (s) can be used to describe the intensity of the acquired spectra. In IR spectra, a change in the absorbance intensity over a certain wavenumber is reported. Although less common, a change in transmittance over a certain wavenumber can be presented. It is important to remember that a relative intensity of bands in the spectra acquired in the transmittance and ATR modalities will change. Therefore, ATR correction should be applied to all IR spectra recorded in ATR mode.

If the overall IR spectral intensity directly depends on the amount of the material, intensity of Raman spectra depends on two factors: sample color and Raman cross-section. Since Raman is a scattering phenomenon, the darker is the sample color, the less intense Raman spectrum will be produced. Therefore, if the same type of materials with different colors, such as corn kernels or plant leaves, are analyzed, spectral normalization should be performed. Our group proposed to use 1440 and 1458 cm^{-1} bands for the spectral normalization. These bands originate from aliphatic (CH_2) vibrations that present in nearly all classes of biological molecules. Therefore, this type of normalization is least biased for a comparison of intensities of vibrational bands that originate from biologically important molecules, such as carotenoids or sugars. Raman cross-section directly depends on the excitation wavelength. Since Raman scattering depends on the fourth power of the frequency of light, utilization of UV or near-UV light is far more beneficial compared to the IR or near IR light.

Interpretation of vibrational bands in the IR and Raman spectra of plant material is a challenging process. In the Raman spectra collected form plant leaves, vibrational bands originating from pectin, cellulose, phenylpropanoids, proteins, and carotenoids can be detected, Tables 1, 2.

In the IR spectra, a vast majority of vibrational bands originates from CH_2 vibrations of alkanes. Such spectra also possess the vibrational bands that can be assigned to carbonyl-containing compounds, such as acids, aldehydes, ketones and esters, as well

TABLE 1 Vibrational bands and their assignments for the Raman spectra collected from plant leaves and seeds.

Band (cm-1)	Vibrational mode	Assignment
480	C-C-O and C-C-C Deformations; Related to glycosidic ring skeletal deformations δ(C-C-C)+τ(C-O) Scissoring of C-C-C and out-of-plane bending of C-O	Carbohydrates (Almeida et al., 2010)
520	ν(C-O-C) Glycosidic	Cellulose (Edwards et al., 1997; Pan et al., 2017)
747	γ(C–O-H) of COOH	Pectin (Synytsya et al., 2003)
849-853	(C_6–C_5–O_5–C_1–O_1)	Pectin (Engelsen and Nørgaard, 1996)
917	ν(C-O-C) In plane, symmetric	Cellulose, phenylpropanoids (Edwards et al., 1997)
964-969	δ(CH_2)	Aliphatics (Yu et al., 2007; Cabrales et al., 2014)
1000-1005	In-plane CH_3 rocking of polyene aromatic ring of phenylalanine	Carotenoids (Schulz et al., 2005); protein
1048	ν(C-O)+ν(C-C)+δ(C-O-H)	Cellulose, phenylpropanoids (Edwards et al., 1997)
1080	ν(C-O)+ν(C-C)+δ(C-O-H)	Carbohydrates (Almeida et al., 2010)
1115-1119	Sym ν(C-O-C), C-O-H bending	Cellulose (Edwards et al., 1997)
1155	C-C Stretching; ν(C-O-C), ν(C-C) in glycosidic linkages, asymmetric ring breathing	Carotenoids (Schulz et al., 2005),carbohydrates (Wiercigroch et al., 2017)
1185	ν(C-O-H) Next to aromatic ring+σ(CH)	Carotenoids (Schulz et al., 2005)
1218	δ(C-C-H)	Carotenoids (Schulz et al., 2005), xylan (Agarwal, 2014)
1265	Guaiacyl ring breathing, C-O stretching (aromatic); -C=C-	Phenylpropanoids (Cao et al., 2006), unsaturated fatty acids (Jamieson et al., 2018)
1286	δ(C-C-H)	Aliphatics (Yu et al., 2007)
1301	δ(C-C-H)+δ(O-C-H)+δ(C-O-H)	Carbohydrates (Cael et al., 1975; Almeida et al., 2010)
1327	δCH_2 Bending	Aliphatics, cellulose, phenylpropanoids (Edwards et al., 1997)
1339	ν(C-O); δ(C-O-H)	Carbohydrates (Almeida et al., 2010)
1387	δCH_2 Bending	Aliphatics (Yu et al., 2007)
1443-1446	δ(CH_2)+δ(CH_3)	Aliphatics (Yu et al., 2007)
1515-1535	-C=C- (in plane)	Carotenoids (Rys et al., 2014; Adar, 2017; Devitt et al., 2018)

(Continued)

TABLE 1 Continued

Band (cm-1)	Vibrational mode	Assignment
1606-1632	ν(C-C) Aromatic ring+σ(CH)	Phenylpropanoids (Agarwal, 2006; Kang et al., 2016)
1654-1660	-C=C-, C=O Stretching, amide I	Unsaturated fatty acids (Jamieson et al., 2018), proteins (Devitt et al., 2018)
1682	COOH	Carboxylic acids (Sanchez et al., 2020c)
1748	C=O Stretching	Esters, aldehydes, carboxylic acids and ketones (Colthup et al., 1990)

as C-O vibrations that originate from carbohydrates. In our previous study, we demonstrated that IR and RS are complementary techniques in the characterization of the plant materials. For instance, IR spectra acquired from plant wax were highly rich with C-O-C, C-O-H and C=O vibrations that could be assigned to carbohydrates, alcohols, aliphatic and aromatic acids, aldehydes, ketones and esters. However, Raman spectra acquired from the same plant material only exhibited C-C, C-H and CH_2 vibrations. Theoretical calculations revealed that with an increase in the length of carbon chain, the intensity of CH_2 vibrations in Raman spectra increases in the fourth power, whereas only a linear increase was expected in the corresponding IR spectra. Since a vast majority of alcohols, acids, aldehydes, ketones and esters in plants have more than C18 carbon atoms, the intense CH_2 vibrations obscure the appearance of other vibrational bands in the Raman spectra acquired from such materials. At the same time, theoretical calculations revealed that RS could be used to reveal conformations of aliphatic chains in such molecules that was not accessible by the IR spectroscopy.

TABLE 2 Vibrational bands and their assignments for the IR spectra collected from plant leaves and seeds.

Wavenumber (cm^{-1})	Vibration	Assignment
668	Out of plane ring bending	Aromatic ring (Colthup et al., 1990)
720	CH_2 in-phase rocking	Alkanes (Colthup et al., 1990)
730	CH_2 in-phase rocking	Alkanes (Colthup et al., 1990)
890	CH_2 wag	Alkanes (Colthup et al., 1990)
947	C-O Stretch	Carbohydrates (Colthup et al., 1990)
958	C-O Stretching	Carbohydrates (Colthup et al., 1990)
1027	C-O Stretching	Alcohols (Colthup et al., 1990)
1093	Substituted Benzene	Aromatic ring (Colthup et al., 1990)
1155	C-O Stretching	Alcohol (Colthup et al., 1990)
1292	CH_2 Twisting	Alkane (Colthup et al., 1990)
1305	C-O stretching	Alcohol (Colthup et al., 1990)
1378	CH_3 Symmetric Deformation; OH deformation of carboxyl monomer	Alkane or carboxylic acid (Colthup et al., 1990)
1438	CH_2 Stretching	Alkane (Colthup et al., 1990)
1463	CH_2 Scissoring	Alkane (Colthup et al., 1990)
1472	CH_2 Scissoring	Alkane (Colthup et al., 1990)
1696	C=O Stretching	Carbonyl compound (Colthup et al., 1990)
1710	C=O Stretching	Carbonyl compound (Colthup et al., 1990)

3 Recent literature

3.1 Applications of Raman spectroscopy

3.1.1 Phytopathology

Phytopathogens pose a significant threat to crops worldwide, with reports indicating that 10% to 40% of crops are lost annually due to these pathogens (Ristaino et al., 2021). Climate change exacerbates this issue, enabling the spread of pests to previously unaffected regions (Bebber et al., 2013). Therefore, timely disease detection is crucial for farmers to mitigate losses. During the past decade, Raman spectroscopy (RS) emerged as a valuable tool for addressing this challenge, outperforming traditional methods like qPCR in direct costs and detection limits (Sanchez et al., 2020e). Previous studies have explored its effectiveness across crops and diseases, including Huanglongbing (HLB) in oranges, bacterial and viral infections in tomatoes, and various fungal diseases in corn (Egging et al., 2018; Sanchez et al., 2019a, Sanchez et al., 2020b). Recent studies have continued expanding the application of RS to various crops and pathogens, providing new insights into the molecular origins of Raman diagnostics.

HLB also known as citrus greening disease, is a devastating bacterial disease that is caused by *Candidatus Liberibacter*). HLB affects citrus production globally with no known cure and an annual cost estimated at $3.6 billion just in the US alone, early detection is critical (Ghosh et al., 2022). Previous studies by the Kurouski lab demonstrated RS was capable of early detection and differentiation of HLB from other biotic and abiotic stresses like nutrient deficiency and blight (Sanchez et al., 2019a, Sanchez et al., 2019a). To identify the underlying molecular nature of RS-based sensing of HLB, Dou and co-workers performed HPLC analysis of plant leaves collected from healthy and infected plants (Dou et al., 2021), Figure 3A. The researchers found drastic differences in the concentrations of

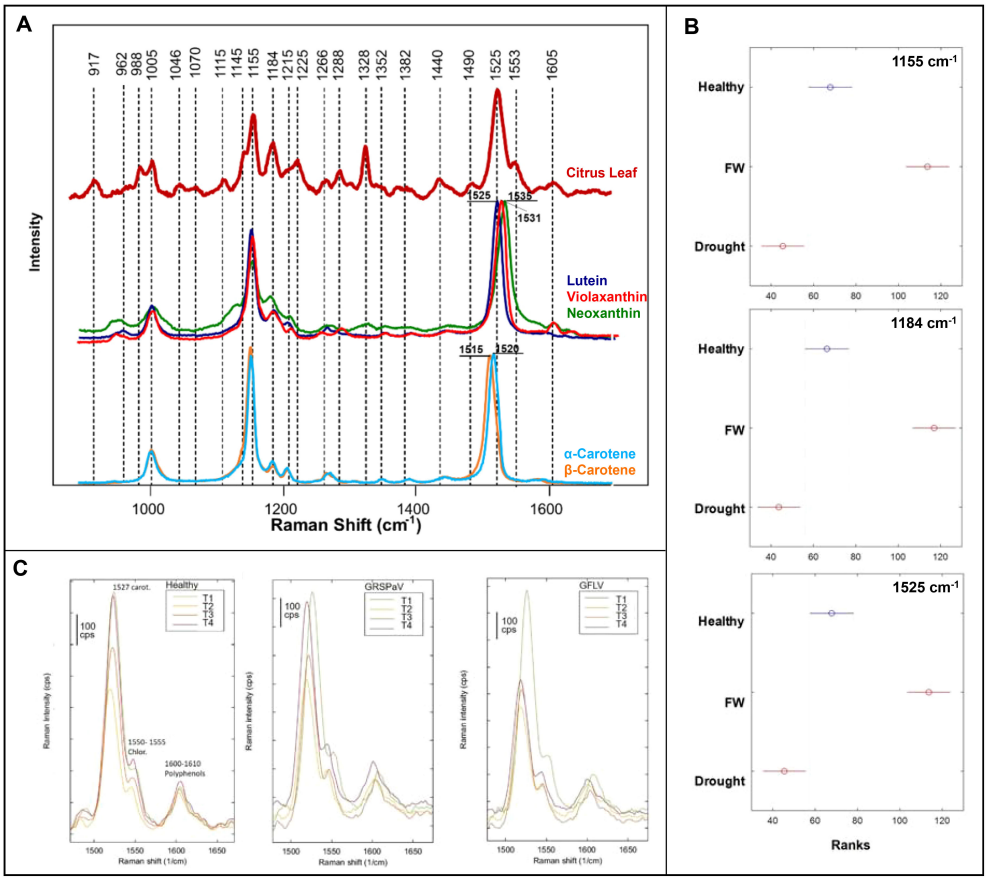


FIGURE 3
(A) Raman spectra of a citrus leaf compared to metabolite standards; **(B)** Kruskal-Wallis of Raman intensities at carotenoid peaks, comparing healthy banana leaves, fusarium wilt infected banana leaves, and drought stressed banana leaves; **(C)** Raman spectra of grapevine leaves at four timepoints (T1-T4) comparing the healthy crop with the two viral infections. Reproduced with permissions from (Dou et al., 2021; Mandrile et al., 2022; Parlamas et al., 2022).

several major carotenoids, including lutein, α and β-carotenes. Dou and co-workers also found a major decrease in the concentration of chlorophyll in the leaves of HLB-infected plants. Next, Raman spectra were acquired from the carotenoids identified by HPLC. It was found that spectroscopic signature of lutein matched the vibrational fingerprint observed in Raman spectra acquired form plant leaves. Based on these results, the researchers concluded that upon the spectroscopic analysis of plant leaves, RS primarily detected changes in lutein. This marks a breakthrough in connecting Raman's diagnostic capabilities with traditional methods of phytopathogen diagnosis.

Fungi are a major crop threat, comprising the largest group of phytopathogens with nearly 8,000 species linked to plant disease (Hariharan and Prasannath, 2021). Notably, *Colletotrichum graminicola* causes anthracnose leaf blight and stalk rot in maize, costing the US economy nearly $420 million annually (Belisário et al., 2022). Farber and co-workers demonstrated that RS coupled with PLS-DA could diagnose stalk rot in maize at the early and late stages (Farber et al., 2021a). This conclusion was made based on the differences in the vibrational bands that could be assigned to cellulose, lignin, and carotenoids in the spectra acquired from healthy and infected plants. Furthermore, these spectroscopic changes correlated with lesion sizes in plant stalks. These results showed that RS could be used to predict cultivar resistance to *C. graminicola*. These experiments were performed with a handheld spectrophotometer equipped with an 830 nm laser. *Fusarium oxysporum*, is another fungus that causes fusarium wilt in bananas (Ploetz, 2021). Parlamas and co-workers were able to detect and identify tropical race 4 strain of *F. oxysporum* infection using RS and PLS-DA (Parlamas et al., 2022). This technique detected the fungal disease in bananas 40 days post-inoculation due to changes in carotenoid-associated vibrations, Figure 3B. They also showcased its robustness in distinguishing from drought stress. Both studies highlight RS's relevance in addressing key fungal pathogens with significant economic implications in global agriculture.

Viruses, with their high mutation rates, present a critical challenge to traditional crop disease management. Rapid identification of infected crops is key, and RS can play a vital role in this solution (Rubio et al., 2020). Using RS coupled with PCA and PLS-DA, Mandrile and co-workers were able to detect grapevine rupestris stem pitting-associated virus and grapevine fanleaf virus in grapevine, a crop central to the economies of several southern European countries (Mandrile et al., 2022). Spectral changes,

mainly in carotenoid content, were confirmed through quantitative chemical analysis, Figure 3C. Combining this with expression levels of enzymes involved in the carotenoid pathway, the researchers revealed that RS was detecting a metabolic cascade leading to the accumulation of abscisic acid and strigalactones. The study employed a dispersive Raman spectrophotometer equipped with a 785 nm laser. Importantly, this study connected enzyme expression data with changes within alterations observed Raman spectra.

Across bacteria, fungi, and viruses, RS consistently relies on changes in carotenoid quantity to differentiate diseased crops from healthy ones. These studies not only broaden the scope of crops and viruses tested but also enhance our understanding of the activated metabolic pathways during pathogenic stress and the specific metabolites responsible for Raman spectral changes.

3.1.2 Environmental stress

Climate change's escalating environmental disasters, coupled with on-going arable farmland losses and population growth, pose the greatest threat to food supply (Zandalinas et al., 2021). Drought is already the leading cause of agricultural loss, incurring approximately $37 billion in annual losses (Fao, 2018). Before 2020, early papers on RS's application to abiotic stresses were limited, but recent advancements demonstrate its effectiveness in studying both environmental and anthropogenic stressors like pesticides and microplastics. This intersection of environmental toxicology and agriculture is critical, given crops' ability to bioaccumulate harmful elements and compounds from the soil.

Breeding resistant crop cultivars is a key strategy against environmental stressors. Altangerel and co-workers demonstrated that RS could assess drought tolerance in maize by studying carotenoid degradation (Altangerel et al., 2021). The researchers also found that only chloroplast carotenoids were depleted during osmotic stress. Finally, Altangerel and co-workers showed that RS could distinguish drought-resistant phenotypes within two weeks in seedlings by tracking carotenoid degradation, greatly expediting the development and testing of new maize cultivars, Figure 4A. This study was performed using a confocal Raman spectrophotometer equipped with a 532 nm laser. Similar breakthroughs were achieved by Higgins and co-workers that used RS to differentiate biotic and abiotic stresses in wheat and maize (Higgins et al., 2022b, Higgins et al., 2023). The first study focused on the analysis of wheat responses to nitrogen deficiency, drought, aphid infestation, and combined viral infections. Despite morphologically similar symptoms, RS coupled with PLS-DA successfully distinguished all stresses with above 90% accuracy. It has been also shown that changes in the Raman spectra (1185 cm^{-1}) have direct relationship with the changes in the concentration of lutein, as was revealed by HPLC. The second study, Higgins and co-workers explored RS's ability to identify individual and co-occurring stresses caused by salinity stress and stalk rot disease in maize, Figure 4B. It has been shown that RS was able to differentiate between individual and combined biotic and abiotic stresses in maize with greater than 90% accuracy. In both studies by Higgins and co-workers, the researchers utilized a handheld spectrophotometer equipped with an 830 nm laser. These studies contribute significantly to the application of RS in detecting diverse stress patterns and highlight RS's ability to differentiate between biotic and abiotic stresses.

Nitrogen is a vital plant nutrient as a major chlorophyll component, and its deficiency significantly reduces plant productivity. While aerial imaging may be used to diagnose nitrogen deficiency, RS is more advantageous in distinguishing the condition from other causes of chlorosis (Mu and Chen, 2021). The Ram lab investigated the accuracy of RS in diagnostics of nitrogen deficiency in three different plant species: Arabidopsis, bok choy, and choy sum (Huang et al., 2020). The researchers found that nitrogen deficiency could be diagnosed by a decrease in the 1046 nm^{-1} peak, Figure 4C. It was also shown that changes in this band were not affected by phosphorus or potassium deficiencies. The second study validated these findings *in situ*, demonstrating the efficacy of the hand-held Raman device for field settings, while expanding to a wider range of crops (Gupta et al., 2020). Both studies used an 830 nm laser, with the first study employing a benchtop Raman spectrophotometer while the second employed a portable Raman clip. Besides environmental conditions, other abiotic stresses result from anthropogenic impacts on the environment.

Pesticides are considered indispensable in modern agriculture. However, the process of bioaccumulation, where plants absorb pesticides from the soil, poses a significant health threat to both humans and livestock. This is attributed to the highly toxic nature of many routine pesticides, even at minute concentrations (Li and Ai, 2023). In a recent study, Sanaeifar and colleagues used RS and electronic nose to assess the concentration of chlorpyrifos residues on tea leaves (Sanaeifar et al., 2021). The researchers employed artificial neural networks to construct a predictive model with a calibration curve measuring residues up to 0.35 mg/kg. The developed model relied on peaks associated with chlorpyrifos moieties at 631, 678, and 1240 cm^{-1}. The study utilized a confocal Raman spectrophotometer equipped with a 532 nm laser. Herbicides also have toxic effects on crops and ornamental plants, as demonstrated by Farber and co-workers. In the recently reported study, RS was used to detect stress in roses induced by common lawn herbicides (Roundup and Weed-B-Gon) with 90% accuracy just a day after application (Farber et al., 2023), Figure 4D. The constructed PLS-DA model relied on major changes at the 1610, 1669, and 1720 cm^{-1} peaks, corresponding to phenylpropanoids, proteins, and carboxyl containing compounds. The study employed a handheld spectrophotometer equipped with an 830 nm laser. These studies, while proof-of-concept, showcase promising techniques applicable to a wide range of pesticides, emphasizing their potential impact.

Plastic pollution is an ongoing environmental and agricultural challenge. As large plastic pieces degrade into microplastics (smaller than 5 microns) and even smaller nanoplastics (less than 1 micron), these particles can be absorbed by crops, contributing to our eventual consumption of microplastics (Programme, 2022). Tympa and co-workers demonstrated the use of RS to sense microplastics in radishes through the direct detection of plastic peaks (Tympa et al., 2021). The microplastics were detected using a confocal Raman spectrophotometer with a 532 nm laser. Since there

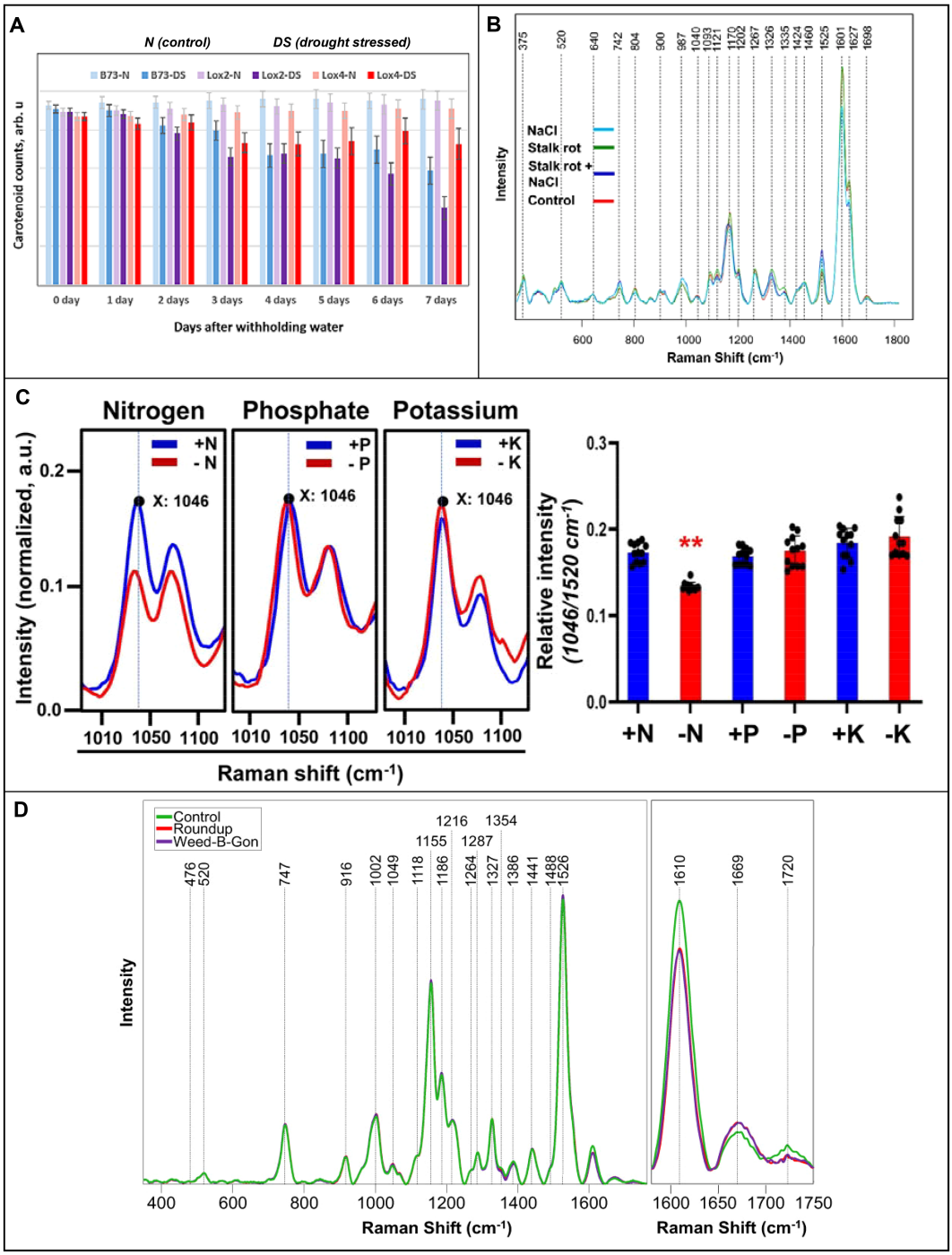


FIGURE 4
(A) Raman spectra of corn stalks exposed to biotic and abiotic stress at experimental day 2; **(B)** Histogram of carotenoid content in drought stressed and control maize lines; **(C)** Raman spectra of Arabidopsis leaves under macronutrient deficiencies with the corresponding histogram on the right; **(D)** Raman spectra of rose leaves one day after herbicide application. Reproduced with permissions from (Huang et al., 2020; Altangerel et al., 2021; Farber et al., 2023; Higgins et al., 2023).

are limited spectroscopic studies on microplastics in agriculture, this research lays a solid foundation for future studies.

The myriad of environmental factors affecting plant growth challenges any diagnostic methods developed, however, RS has proven versatile in diagnosing stress from climate, nutrients, and anthropogenic threats. These studies lay a robust foundation for exploring additional abiotic stresses and enhancing our understanding of RS's molecular detection capabilities.

3.1.3 Phenotyping

Phenotyping is the foundation of plant breeding; therefore, the agriculture sector's progress relies on developing new cultivars

(Costa et al., 2019). Traits are usually selected for reasons like stress resistance and optimized crop features. Spectroscopy, particularly RS, excels in high-throughput plant phenotyping, which has been a limitation of traditional methods. Previous applications of RS include assessing crop quality in tomatoes and olives (Muik et al., 2004; Nikbakht et al., 2011), identifying crop genotypes (Farber et al., 2020; Morey et al., 2020), and distinguishing plants like hemp from cannabis (Sanchez et al., 2020a, Sanchez et al., 2020c). While current studies have continued exploring these areas, novel applications of Raman mapping in plant science are at the forefront.

Accurate prediction of the optimal harvest date (OHD) is essential for successful crop harvesting (Xu et al., 2019). OHD relies on several metabolites such as phytochemical concentration, starch, and sugar content. While many crops show visual cues, vegetative crops often lack them, and certain fruits may not visibly indicate changes in sugar content. In addressing this challenge, Li and co-workers used RS to address this very problem in spearmint and found that the 1600 cm^{-1} peak had a linear increase with concentration of rosmarinic acid in the spearmint leaves as measured by HPLC (Li et al., 2021). The researchers also separated their spectra based on whether the portion of the leaf scanned contained trichomes or not, and how old the leaf was. Based on these results, the researchers found different leaf structures influence rosmarinic acid concentration, while leaf age did not, Figure 5A. Thus, these results showed that any leaf could be used for RS to determine OHD. In this study, the researchers used a confocal Raman spectrophotometer equipped with a 785 nm laser. Carotenoids, natural pigments responsible for vibrant colors in plants, exhibit varying concentrations that can either increase or decrease during ripening. As a proof of principle, Hara and co-workers optimized RS for the determination of carotenoid content in tomatoes, finding that for their specific system, 10 seconds of exposure correlated best with carotenoid concentration (Hara et al., 2021). In another study, Dhanani and co-workers used RS over a period of 50 days to track carotenoid concentration changes in watermelon rind to determine its ripeness (Dhanani et al., 2022), Figure 5B. The reported results demonstrated that RS could be used to determine a ripening stage of four different varieties of watermelons with around 85% accuracy. They also used resonant Raman to determine that lutein and β-carotene were the predominant molecular species in the rind responsible for the accuracy of RS-based sensing of fruit ripeness. Both studies used a handheld Raman spectrophotometer, with Hara using a 785 nm laser and Dhanani using an 830 nm laser. These detailed studies highlight RS's robustness in providing precise information for harvesting decisions.

Most flowering plants have both male and female reproductive organs within each flower, but approximately 6% are dioecious, exclusively producing male or female gametes (Käfer et al., 2017). Dioecious plants are traditionally sexed through visual examination, which is extremely time-consuming. To this end, the Kurouski lab showed that RS could distinguish female hemp from male hemp. Specifically, Higgins and co-workers achieved over 90% accuracy in sex identification in both young and mature crops (Higgins et al., 2022a). Separately, Goff and co-workers used RS to differentiate dioecious from hermaphrodite hemp (Goff et al., 2022), Figure 5C. HPLC unveiled that RS distinguishes plants based on carotenoid concentration differences, with females having significantly higher lutein levels compared to both male and hermaphrodite hemp. Both studies employed an 830 nm laser-equipped handheld Raman spectrophotometer. These findings highlight RS's future potential to automate sexing in hemp and other dioecious crops.

The fundamental science behind phenotyping is genotyping. While many physical traits are observable, some traits are not distinguishable like nutrient content and resistances. Furthermore, traditional genotyping with primers and qPCR can incur high costs. The Kurouski lab has applied RS as a cost-effective alternative in several crops. Notably, the researchers differentiated nutrient components in 15 rice genotypes using RS, identifying protein, polyphenol, and oil peaks for differentiation (Farber et al., 2021b). RS could also determine total starch content in rice grains. Genotyping experiments on peanut plants leaves achieved 94% accuracy for six genotypes when coupled with PLS-DA (Payne et al., 2022). Furthermore, the same model could identify resistance to nematodes with 83% accuracy. This study underscores RS's potential in identifying multiple valuable crop traits after a model is trained on a sizable database. In another study, the Kurouski lab evaluated glyphosate-resistant Palmer amaranth (Singh et al., 2021), Figure 5D. After one day of exposure to glyphosate, RS could phenotype crops with 80% accuracy, while achieving around 70% accuracy for unexposed crops. All these studies used a handheld-Raman spectrophotometer equipped with an 830 nm laser. These studies have significantly advanced the field, particularly in seed genotyping. However, additional research is needed to fully assess the limits of RS in genotyping, particularly regarding resistance determination.

Raman mapping is a combined technique of RS and imaging where a laser is used across discrete section of an area, providing spatially resolved molecular information about the sample (Gordon and McGoverin, 2011). Images can be formed from the spectra based on the distribution of intensity for a selected spectral region. While not new in plant science, it has recently gained more exploration. The Gierlinger group broadened Raman mapping applications to various plants, particularly in recent projects focusing on imaging plant cuticles. These efforts bridged the gap between structural insights from electron microscopy and chemical analyses. In one study, the researchers used Raman mapping analyze cuticle layers in Arabidopsis stem, tomato peel, and spruce needle (Bock et al., 2021), Figure 6A. The images provided 300 nm spatial resolution on the composition of each layer, allowing the researchers to develop models of the cuticles. The researchers also determined molecular orientation using polarized laser light. Finally, they observed that higher concentrations of aromatic compounds resulted in better signal-to-noise ratios with a 785 nm laser compared to a 532 nm laser, albeit at the expense of spatial resolution. Another study by Gierlinger group aimed to comprehensively analyze the cuticle and epidermis of spruce needles (Sasani et al., 2021). Using their mapping technique, the researchers assessed several phytochemicals for changes in concentration and orientation across the cuticle. The researchers

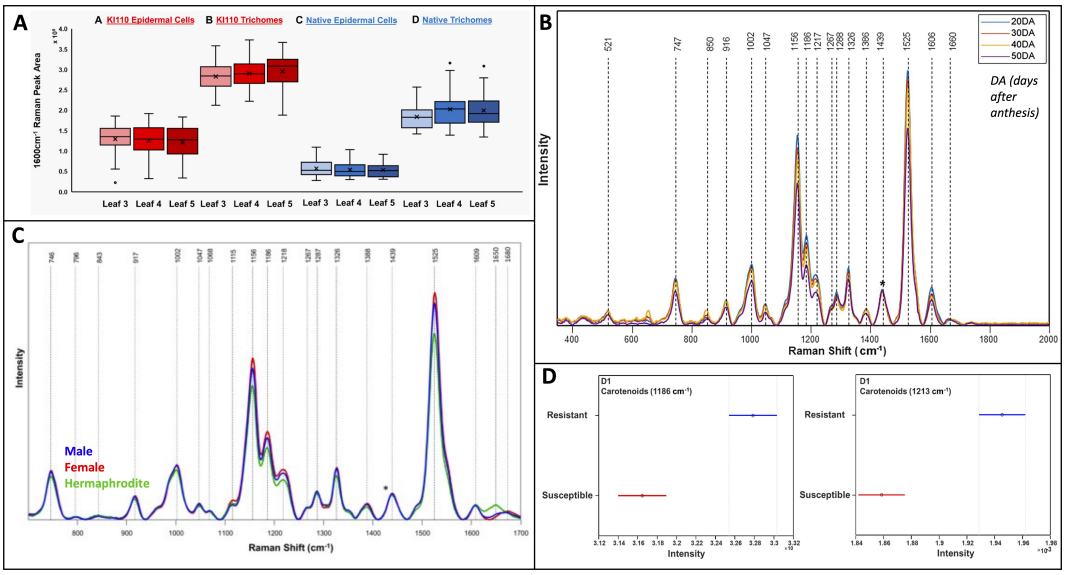


FIGURE 5
(A) Box and whiskers plot of rosmarinic acid abundance in spearmint comparing different leaf ages and structures for two cultivars; **(B)** Raman spectra of watermelon rind over maturation; **(C)** Raman spectra of mature cannabis plants comparing sex; **(D)** Raman spectra of phenotypically resistant and susceptible palmer amaranth one day after glyphosate exposure. Reproduced with permissions from (Li et al., 2021; Singh et al., 2021; Dhanani et al., 2022; Goff et al., 2022).

also used RS to track changes in tomato cuticle chemistry through maturity (González Moreno et al., 2023), Figure 6B. Specifically, the researchers identified variations in the concentrations of phenolics and flavonoids during ripening, enabling the construction of a novel model for cuticle development in tomatoes. This discovery revolutionized how we can study developmental plant biology. Separately, Zeng and co-workers used Raman mapping to visually quantify carotenoids and chlorophyll in tea leaves (Zeng et al., 2021). The researchers first used a confocal microscope to build a calibration model. This model was combined with map scanning spectra, allowing the researchers to predict analyte concentrations at any point on the Raman map. The study employed a 532 nm

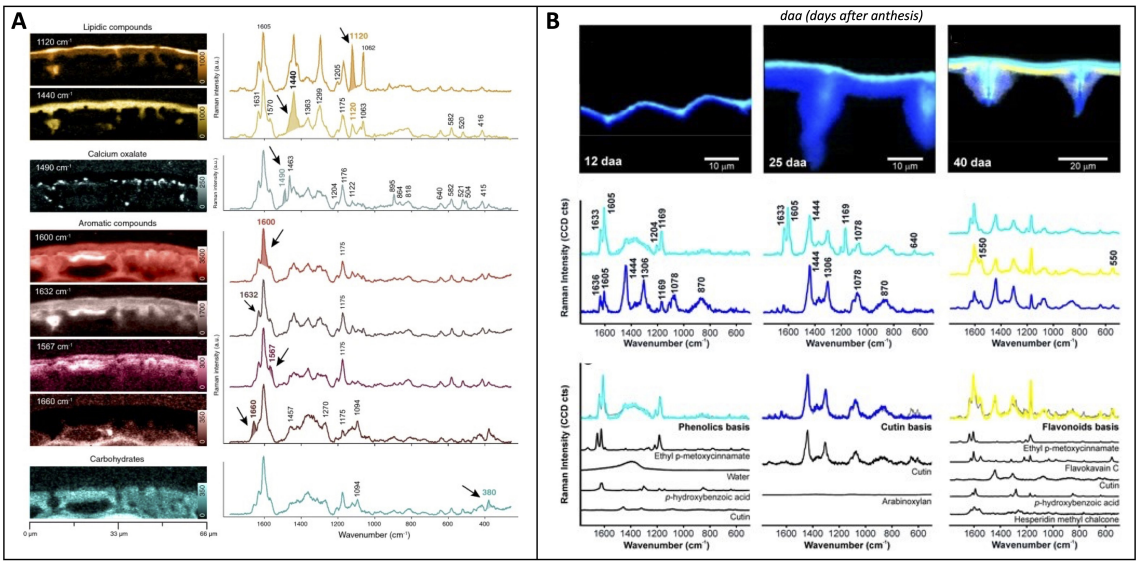


FIGURE 6
(A) Raman imaging of spruce needle visualized by the integration of various peaks corresponding to specific biomolecules; **(B)** Raman imaging visualized by unmixing analysis (NMF) of tomato cuticles from select stages of development, with the model fit spectra shown below the basis spectra. Reproduced with permissions from (Bock et al., 2021; González Moreno et al., 2023).

laser. These studies by the Gierlinger lab and Zeng et al. lay the groundwork for diverse applications of Raman mapping in plant science, paving the way for future endeavors.

The integration of RS into plant science has revolutionized phenotyping and genotyping, showcasing its versatility in agriculture. The recent applications emphasize the importance of carotenoids and aromatic molecules in Raman's ability to distinguish crop traits. While applications such as sexing and mapping have limited current studies, further expansion in the scientific application of these underlying principles is anticipated.

3.2 Applications of surface-enhanced Raman spectroscopy

3.2.1 Phytopathology and environmental stress

Traditional Raman in phytopathology has revolutionized proactive crop disease diagnosis by rapidly assessing plant metabolite concentrations. However, its diagnostic capabilities are limited by the magnitude of these changes. SERS offers solutions by detecting metabolites in minute quantities or detecting pathogens directly. Earlier SERS studies exemplify these strategies, albeit within a limited scope of research in plants (Kim et al., 2013; Lau et al., 2016; Zhou et al., 2020). Similar strategies could be used for detection of environmental stressors, as shown in a study tracking leaf-applied pesticide (Yang et al., 2017). Recent studies continue to expand on these SERS applications, exploring more phytopathogens and environmental factors.

Direct pathogen detection offers advantages over traditional spectroscopy, overcoming challenges posed by pathogens' evolved defenses that can destroy immune signaling molecules or evade detection by immune cells. This is crucial as these defenses often mask typical infection symptoms detected by Raman spectroscopy. Jiang and co-workers recently applied SERS to detect minute changes in carotenoid concentration in kiwifruit leaves caused by *Pseudomonas syringae*, the bacterium responsible for kiwifruit canker (Jiang et al., 2023b), Figure 7A. The study demonstrated SERS's effectiveness in greatly amplifying signals from carotenoids compared to normal RS, both in early and late stages of disease. The researchers utilized silver nanoparticles coated with iodide and calcium ions and a confocal Raman spectrophotometer equipped with a 532 nm laser. Given the nanoparticles' high specificity, one may question the need for individually designed nanoparticles for each specific analyte. A groundbreaking study by Son and co-workers showed that their nanoparticles could detect critical phytochemicals responsible for stress responses in watercress, wheat, and barley (Son et al., 2023). The researchers tracked changes in salicylic acid, ATP, phytoalexins, and glutathione, Figure 7B. These molecules exhibited unique vibrational bands that allowed the researchers to observe biotic factors (fusarium graminarium) and abiotic factors (wound stress, cold stress). The researchers also revealed that nanoparticles larger than 100 nm localized in the intercellular space without

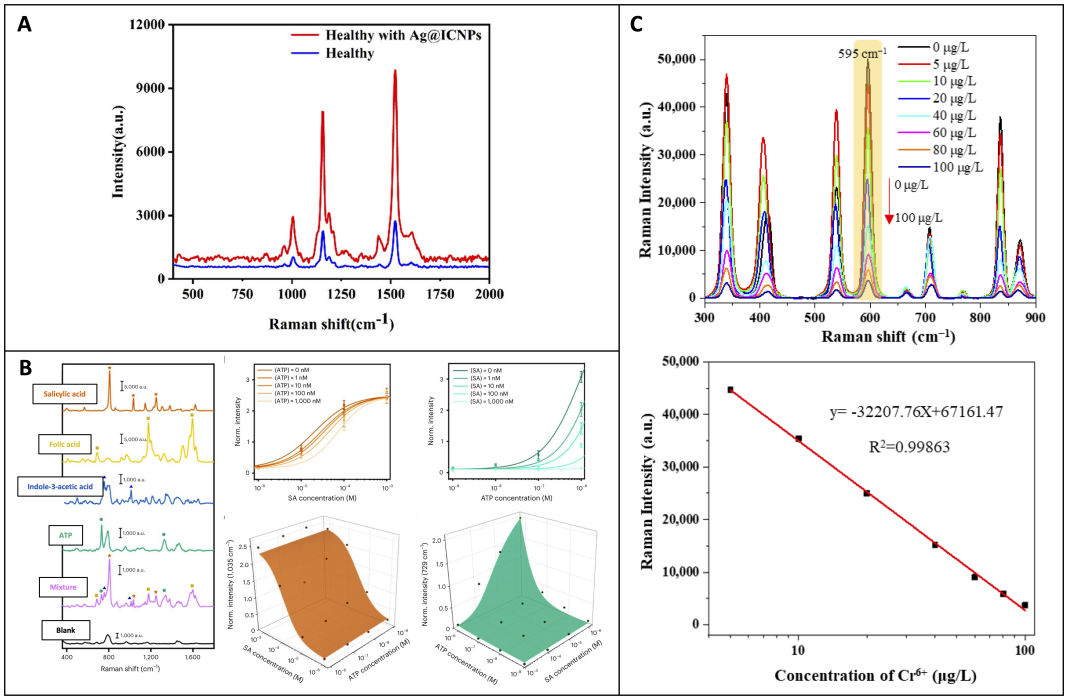


FIGURE 7
(A) Raman spectra of kiwifruit leaves with and without nanoparticles; **(B)** SERS spectra of various plant signaling molecules, along with concentration-dependence plots and three-dimensional plots of peak intensity at 1035 cm-1 and 729 cm^{-1} based on the combination ATP and salicylic acid concentrations. **(C)** SERS spectra of carbimazole at different concentrations of Cr6+, with a calibration curve below based on Raman intensity at the 595 cm-1 peak. Reproduced with permissions from (Jiang et al., 2023b; Son et al., 2023; Yin et al., 2023).

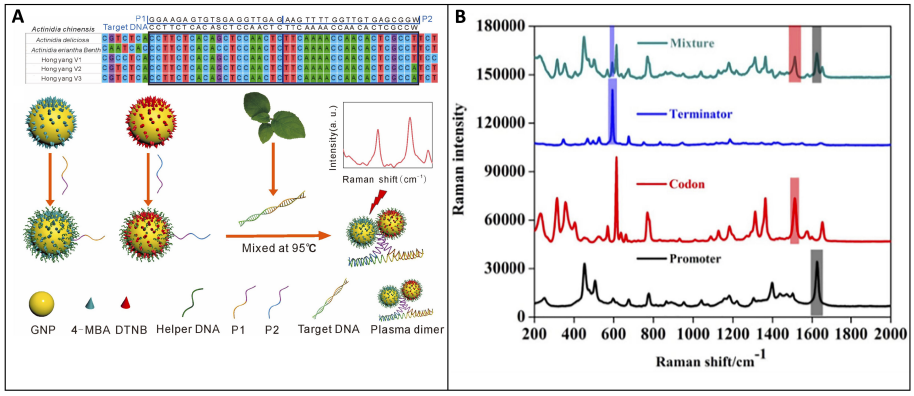


FIGURE 8
(A) The formation of GNPs combined with target DNA leading to plasmonic dimers with strong SERS enhancement of reporter molecules; **(B)** SERS spectra for three GMO components and differentiation within a mixture. Reproduced with permissions from (Yao et al., 2022; Jiang et al., 2023a).

causing adverse effects. The study used silver-capped silicon nanoparticles coated in poly(diallyl dimethylammonium chloride) and a confocal Raman spectrophotometer equipped with 660 nm and 785 nm lasers. This research emphasizes that diverse nanoparticle development may not be necessary, as a single nanoparticle can effectively detect a wide range of conditions.

High-performance liquid chromatography is commonly used to monitor pesticide bioaccumulation in crops, but real-time tracking is challenging due to its destructive nature and lack of structural uptake information (Sicbaldi et al., 1997). Yang and co-workers demonstrated SERS could be used to track the uptake of the pesticide thiabendazole in tomato plants (Yang et al., 2019). The researchers observed a concentration-dependent journey of the pesticide from the midrib to the leaf margin when taken up by the roots. The nanoparticles detected thiabendazole directly, with the SERS limit of detection at around 2 μg/g of leaf tissue. The study used citrate-capped gold nanoparticles and a dispersive Raman spectrophotometer equipped with a 780 nm laser. Like pesticides, heavy metals are a significant focus of bioaccumulation in crops. Like pesticides, heavy metals bioaccumulate in crops due to naturally high levels in many agricultural regions worldwide. Yin and co-workers applied SERS to measure hexavalent chromium bioaccumulation in tea leaves (Yin et al., 2023), Figure 7C. They employed nanoparticles coated with methimazole, a compound which exhibits a selective reaction with chromium. Measurement of chromium concentration was achieved by tracking linear intensity decreases at the 595 cm^{-1} peak, which corresponded to decreases in methimazole concentration. This system achieved a limit of detection of 0.945 mg/kg of leaf tissue. The study utilized silver-coated gold nanoparticles capped with carbimazole, which hydrolyzes to methimazole, and a confocal Raman spectrophotometer equipped with a 785 nm laser. It's worth noting that the study on thiabendazole was conducted *in vivo*, while the study on chromium used digested plant tissue. Both studies showcase diverse approaches to detecting minute toxic compounds in crops.

Whether by direct or indirect detection, there is major potential of SERS for highly specific detection of stress factors in crops in agriculture as researchers explore innovative detection methods. These SERS advancements hold promise for the future of crop disease diagnosis and environmental monitoring, offering several advantages over traditional Raman.

3.2.2 Phenotyping

Similar chemical principles underlie the application of Raman and Surface-Enhanced Raman Spectroscopy (SERS) in phenotyping. However, a key distinction lies in SERS's usage of custom nanoparticles for direct detection of specific proteins and DNA sequences (Almehmadi et al., 2019). This approach can resemble primer-based techniques such as qPCR, utilizing labels on nanoparticles; however, label-free detection of the nucleic acid itself is also possible.

SERS can overcome the challenges encountered by Raman and PCR in detecting nucleic acids, methods which are hindered by the low concentration of nucleic acids and the requirement for amplification. Dina and colleagues demonstrated that SERS could be used to amplify the intrinsic Raman signal of DNA without amplification, specifically focusing on potato and grapevine leaf tissue (Dina et al., 2022). The study utilized chloride-capped silver nanoparticles and a confocal Raman spectrophotometer with 532 and 785 nm lasers. This validates the feasibility of nucleic acid detection for plant tissue. Direct detection of specific DNA sequences has recently found success in sexing kiwi plants. Jiang and co-workers demonstrated that SERS could sex dioecious kiwi plants at an early age using gold nanoparticles coated in primer DNA and Raman reporters (Jiang et al., 2023a), Figure 8A. The researchers achieved a remarkably low limit of detection (100 fM) by employing two primers that bound to the male sex gene, vastly outperforming PCR. The resulting Raman signal could be used for sex assignment, relying on peak intensities at 1071 and 1326 cm^{-1} which corresponding to the two Raman reporters 4-MBA and DTNB. The study utilized a portable Raman spectrophotometer equipped with a 785 nm laser. Similarly, this technology finds application in detecting genetically modified organisms (GMO) components within crops. Yao and co-workers, for instance, utilized SERS in conjunction with a lateral flow strip to identify GMO components of soybeans (Yao et al., 2022), Figure 8B. Modified nanoparticles enabled the

detection of three GMO components, and a linear correlation between SERS signal and analyte concentration was observed. This study utilized silver-shelled gold nanoparticles conjugated with thiolated DNA and Raman reporters. It also utilized a confocal Raman spectrophotometer. All studies were conducted using extractions and were therefore destructive, raising the future possibility of improving the technology for *in vivo* detection.

While conventional Raman excels in phenotyping, the low concentration of nucleic acids creates an opportunity for SERS to play a pivotal role. These recent studies highlight the potential of SERS as an invaluable technique for swiftly phenotyping crops, particularly when equipped with pre-prepared genetic segments of DNA.

3.2.3 Other SERS-based studies

Recent SERS advancements in agriculture exhibit great potential yet pose questions about *in-vivo* nanoparticle use. Since engineered nanoparticles are meant to be absorbed by crops, they pose health risks for both plants and human consumers (Shrivastava et al., 2019). Variations in toxicity also complicate their applications; for instance, despite silver nanoparticles providing the best SERS enhancement, they also possess the greatest biological toxicity (Ferdous and Nemmar, 2020). Furthermore, nanoparticle translocation is not fully understood. Nanoparticles are primarily confined to intercellular spaces due to size limitations imposed by plant cell walls, so recent studies aim to understand nanoparticle move within plants and to improve their ability to infiltrate cells.

While studies have explored nanoparticle uptake, few have specifically tracked translocation with SERS. Yilmaz and team addressed this gap by exposing maize seedlings to silver nanoparticles and monitoring their translocation with RS (Yilmaz et al., 2021). The researchers observed accumulation in the root and the phloem of the stem, with smaller particle size linked to greater accumulation. Toxicity experiments revealed inhibition of root and leaf length, reduced chlorophyll content, increased protein content, and alterations in mineral composition. However, these effects will largely vary on the specifications of the nanoparticles. The study employed two sets of nanoparticles synthesized through chemical and green synthesis methods, utilizing a confocal Raman spectrophotometer equipped with a 785 nm laser. Size significantly influences the movement and effects of nanoparticles, and location-specific uptake is primarily restricted. For example, roots can take up nanoparticles smaller than 100 nm, leaves up to 50 nm, and cell walls generally limit cellular uptake to around 20 nm (Ballikaya et al., 2023; Wang et al., 2023). To solve this, Cupil-Garcia and co-workers designed 12 nm rod-shaped nanoparticles capable of penetrating past the tobacco leaf cell wall (Cupil-Garcia et al., 2023). Infiltration was verified with several techniques including, transmission electron microscopy, two-photon luminescence, photoacoustic imaging, and SERS. This multi-technique approach enables advancements in various spectroscopy fields. This study used silver coated gold nanorods and a confocal Raman spectrophotometer with a 785 nm. Importantly, successful nanoparticle infiltration into plant cells opens new possibilities for detecting intracellular proteins. There is still much to explore in the study of SERS nanoparticles, including their toxic effects and the development of new designs.

Advancing nanoparticle development for SERS applications in crops shows great promise. However, the utilization of nanoparticles in plants requires not only study of their photonic capabilities but also an assessment of the potential health risks they pose to both plants and humans. Still, as new nanoparticle designs emerge, the range of possible applications in agriculture will only continue to expand.

3.3 Other spectroscopic techniques

3.3.1 Infrared spectroscopy

The complementary nature of Raman and infrared spectroscopies reveals similarities and differences in their capabilities for plant detection. Each technique exhibits preferences for specific moieties. Therefore, IR complements RS by potentially filling the detection gaps for compounds that Raman may miss. Previous applications of IR have successfully detected pesticides in cucumbers and diseases like zebra chip in potatoes and sour rot infection in tomatoes (Jamshidi et al., 2015; Liang et al., 2018; Skolik et al., 2019). The largest applications of IR in plant science however are phenotypic applications, such as studying fruit composition and structural characterization (Cozzolino, 2014; Farber et al., 2019). Recent studies have further expanded IR's applications, investigating previously unstudied pesticides, other abiotic stresses, and continuing studies on plant composition.

In a study by Lu and co-workers, IR was applied to detect chlorpyrifos and carbendazim residues in cabbages (Lu et al., 2021). The researchers built chemometric models to quantify residue amounts in vegetables, allowing for the quick assessment of whether residues met food safety standards. The researchers noted lower decreased performance at trace pesticide levels, emphasizing the need for improved accuracy at lower detection limits. IR can also directly track metabolites, aiding in environmental stress detection. Zhang and co-workers used IR spectroscopy to identify drought stress in crops by measuring malondialdehyde content in pine needles (Zhang et al., 2021). The study extensively explored chemometrics, fine-tuning a PLS model for optimal prediction, ultimately achieving a model with an R^2 value of 0.66. Drought stress is challenging to model due to co-occurrence with effects like high temperature and low humidity, therefore one can expect other modeled stresses to have greater viability. These studies both emphasize the role data manipulation has in predictively using IR spectroscopy.

Determination of plant analytes is key for several sectors of the agriculture industry. Geskovski and co-workers used Mid-IR spectroscopy to quantify THC and CBD content in cannabis extract and flowers by constructing a multivariate model that achieved R^2 values above 0.95 (Geskovski et al., 2021), Figure 9. The researchers also considered spectral differences between THC/CBD and their acidic precursor forms, emphasizing the importance of accounting for different metabolite forms in concentration determination. On the other hand, Miao and co-workers demonstrated that near-IR could be

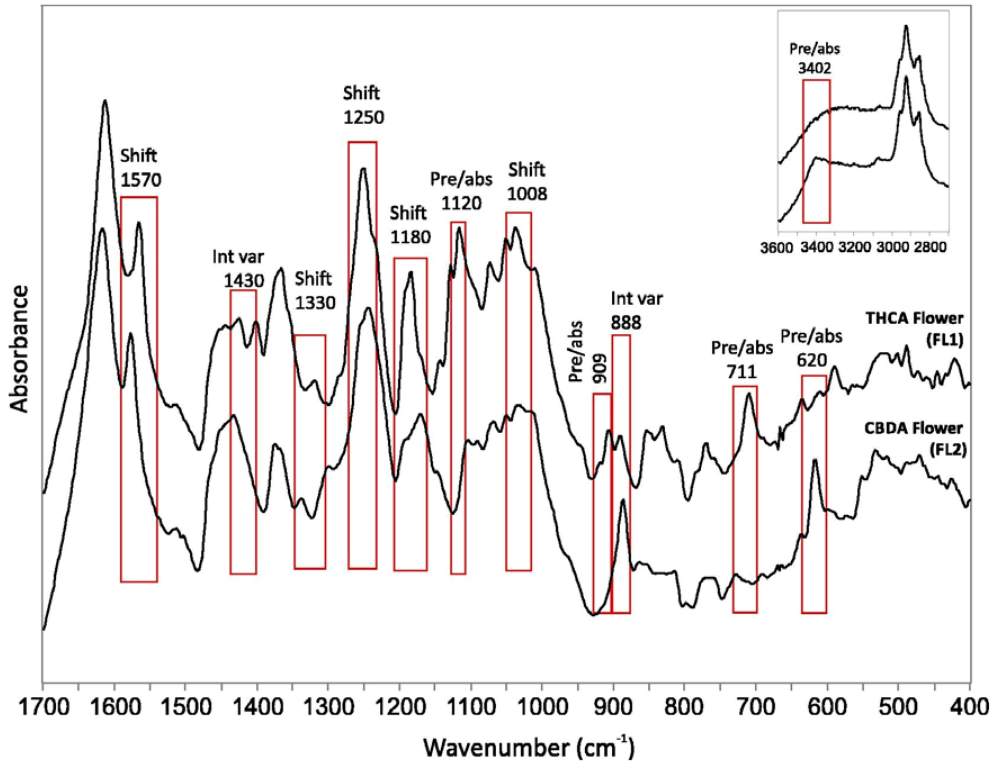


FIGURE 9
FTIR spectra of THCA and CBDA dominant flowers with major differences marked by rectangles. Reproduced with permissions from (Geskovski et al., 2021).

used for the accurate assessment of nitrogen content in rice, achieving an R^2 value greater than 0.95 with synergy interval PLS (Miao et al., 2023). This was primarily based on the peak at 4854 cm^{-1}, corresponding to the N-H stretch indicative of protein content. The study was conducted across various stages of crop growth, making the model relevant to all growth stages of rice. Although not applied to fresh plant tissue, the technique's ease of use for high-throughput phenotyping remains a significant advantage. Understanding the structural composition of plant cell walls is crucial for comprehending plant development and biomass utilization (Pettolino et al., 2012). Liu and co-workers employed ATR-FTIR to characterize 58 cell wall polysaccharides, learning that individual peaks could determine cellulose, hemicellulose, and pectin content (Liu et al., 2021). The cell walls could then be differentiated based on the abundance of these polysaccharides. Still, vibrational similarities in molecules like arabinan and galactan limit the ability to distinguish cell walls that vary in concentrations of these two polysaccharides. Nevertheless, the study's examination of numerous polysaccharides indicates strong potential for future applications in rapidly determining plant cell wall structure.

Many of these studies highlight the crucial role of chemometrics in distinguishing variations within the spectra. As models improve sensitivity, challenges like the strong signal from water molecules in IR should diminish. While Raman has seen more development in the plant field, there's a significant need to explore diverse diagnostic and quantitative applications of IR to better understand its limits.

4 Current limitations and future perspectives

RS has rapidly expanded in agriculture, solving issues posed by traditional techniques. It has thoroughly been utilized for studying various phytopathogens (bacterial, fungal, or viral) and environmental stresses critical in our changing climate. RS excels in these aspects, marking a significant paradigm shift towards digital farming in agriculture. Many recent studies have predominantly focused on proof-of-principle experiments under controlled growing conditions, isolating a single variable. While these studies contribute to constructing stress response models, it is crucial to compare these models with actual field growing conditions. RS detects plant stress by monitoring changes in metabolite content, typically involving carotenoids and phenylpropanoids in most cases. However, as demonstrated by studies using RS to determine OHD, these metabolites can vary in concentration due to maturation. In addition, field conditions seldom involve isolated stresses, often featuring the co-occurrence of multiple minor stresses. Consequently, conducting more longitudinal studies on stress in actual field crops would bridge the gap between these controlled experiments and real-world application.

Phenotypic applications of RS further showcase the technique's versatility, including determining OHD, crop sexing, and genotyping. The recent development of Raman mapping in agriculture is especially noteworthy. While offering limited advantages for real-time plant monitoring, it holds significant

promise for the detailed examination of plant structure. Raman mapping has transformed human histological studies of diseases, an application which could be translated to plants. This could offer insights into how various adverse growth conditions and diseases impact plant histology and structure. Finally, while each study highlights the usefulness of RS for a specific aspect of a particular crop, there is a lack of comprehensive efforts to fully integrate RS for all aspects within a single crop. This end goal is to develop a complete chemometric model capable of predicting multiple stresses for a crop and monitoring its overall growth, marking the final step toward the widespread implementation of RS in daily agricultural practices.

As was mentioned in the introduction, substantial costs of Raman spectrometers limit broad utilization of RS in farming and plant breeding. With the current instrumental price, RS can be implemented as a service rather than a technology that can be possessed by every farmer or breeding center. Nevertheless, miniaturization of Raman spectrometers and reduction of their cost could be an avenue that will transform the use of this innovative technique in farming. Furthermore, it remains unclear whether spectroscopic library acquired at one geographic location could be directly transferred to detect plant stresses at other geographic locations. The same question can be posed about variability of signals from different varieties of plants. From one perspective, such specificity is advantageous to identify resistant and susceptible cultivars. However, from another perspective, it remains unclear whether detection and identification of plant stresses would require additional variety-specific calibration. Finally, RS-based assessment of plant stresses would benefit from coupling of this highly sensitive technique with approaches, such as RGB or thermography drone or satellite-based imaging. These imaging techniques could be used to detect 'problem' areas in fields that can be later inspected by RS to identify the problem.

SERS development in agriculture, though slower than RS, presents equally robust applications. Its strengths include detecting minute concentrations and specified targets, with most stress detection studies emphasizing the former. Carotenoid-linked peaks dominate the normal Raman spectra of plants, so SERS could feasibly distinguish a stress detectable through trace metabolites like ATP and salicylic acid. An overlooked application of SERS in agriculture is the direct detection of pathogens via antibody or DNA coated nanoparticles. This would offer a faster alternative to traditional methods like PCR for confirming disease. Despite its versatility, concerns about the impact nanoparticles have on plant and human health limit widespread application. Few SERS studies conduct toxicity assays that would build confidence in food safety. Additionally, the use of nanoparticles is step beyond base Raman, so studies using them must compare their application to base Raman to justify the use of SERS. Nevertheless, ongoing nanoparticle advancements will continue to incite novel applications for SERS in agriculture.

IR spectroscopy has found several applications in agriculture, although not as extensively as Raman. Its strength lies in identifying compounds that are not Raman active, making it valuable for analyzing plant matter compositions. However, the significant water signal detected by IR often restricts studies to extracts or dried plant tissue. This limits its field application and renders it an often-destructive technique. Despite these drawbacks, IR analysis of plant tissue remains faster than methods like HPLC or ICP-MS. While IR may face challenges in transitioning to digital farming, its usefulness in existing applications should not be overlooked.

Author contributions

IJ: Conceptualization, Visualization, Writing – original draft. DK: Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This review was funded by the Institute for Advancing Health Through Agriculture.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Adar, F. (2017). Carotenoids - their resonance raman spectra and how they can be helpful in characterizing a number of biological systems. *Spectroscopy* 32, 12–20.

Agarwal, U. P. (2006). Raman imaging to investigate ultrastructure and composition of plant cell walls: distribution of lignin and cellulose in black spruce wood (Picea mariana). *Planta* 224, 1141–1153. doi: 10.1007/s00425-006-0295-z

Agarwal, U. P. (2014). 1064 nm FT-Raman spectroscopy for investigations of plant cell walls and other biomass materials. *Front. Plant Sci.* 5, 112. doi: 10.3389/fpls.2014.00490

Almehmadi, L. M., Curley, S. M., Tokranova, N. A., Tenenbaum, S. A., and Lednev, I. K. (2019). Surface enhanced Raman spectroscopy for single molecule protein detection. *Sci. Rep.* 9, 12356. doi: 10.1038/s41598-019-48650-y

Almeida, M. R., Alves, R. S., Nascimbem, L. B., Stephani, R., Poppi, R. J., and de Oliveira, L. F. (2010). Determination of amylose content in starch using Raman spectroscopy and multivariate calibration analysis. *Anal. Bioanal. Chem.* 397, 2693–2701. doi: 10.1007/s00216-010-3566-2

Altangerel, N., Huang, P.-C., Kolomiets, M. V., Scully, M. O., and Hemmer, P. R. (2021). Raman spectroscopy as a robust new tool for rapid and accurate evaluation of drought tolerance levels in both genetically diverse and near-isogenic maize lines. *Front. Plant Sci.* 12, 621711. doi: 10.3389/fpls.2021.621711

Ballikaya, P., Brunner, I., Cocozza, C., Grolimund, D., Kaegi, R., Murazzi, M. E., et al. (2023). First evidence of nanoparticle uptake through leaves and roots in beech (Fagus sylvatica L.) and pine (Pinus sylvestris L.). *Tree Physiol.* 43, 262–276. doi: 10.1093/treephys/tpac117

Bebber, D. P., Ramotowski, M. A., and Gurr, S. J. (2013). Crop pests and pathogens move polewards in a warming world. *Nat. Climate Change* 3, 985–988. doi: 10.1038/nclimate1990

Belisário, R., Robertson, A. E., and Vaillancourt, L. J. (2022). Maize anthracnose stalk rot in the genomic era. *Plant Dis.* 106, 2281–2298. doi: 10.1094/PDIS-10-21-2147-FE

Bock, P., Felhofer, M., Mayer, K., and Gierlinger, N. (2021). A guide to elucidate the hidden multicomponent layered structure of plant cuticles by Raman imaging. *Front. Plant Sci.* 12, 793330. doi: 10.3389/fpls.2021.793330

Cabrales, L., Abidi, N., and Manciu, F. (2014). Characterization of developing cotton fibers by confocal raman microscopy. *Fibers* 2, 285–294. doi: 10.3390/fib2040285

Cael, J. J., Koenig, J. L., and Blackwell, J. (1975). Infrared and raman spectroscopy of carbohydrates. 4. Normal coordinate analysis of V-amylose. *Biopolymers* 14, 1885–1903. doi: 10.1002/bip.1975.360140909

Cao, Y., Shen, D., Lu, Y., and Huang, J. (2006). A Raman-scattering study on the net orientation of biomacromolecules in the outer epidermal walls of mature wheat stems (Triticum aestivum). *Ann. Bot.* 97, 1091–1094. doi: 10.1093/aob/mcl059

Colthup, N. B., Daly, L. H., and Wiberley, S. E. (1990). *Introduction to infrared and raman spectroscopy*. (San Diego, CA: Academic Press).

Costa, C., Schurr, U., Loreto, F., Menesatti, P., and Carpentier, S. (2019). Plant phenotyping research trends, a science mapping approach. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01933

Cozzolino, D. (2014). Use of infrared spectroscopy for in-field measurement and phenotyping of plant properties: instrumentation, data analysis, and examples. *Appl. Spectrosc. Rev.* 49, 564–584. doi: 10.1080/05704928.2013.878720

Cupil-Garcia, V., Li, J. Q., Norton, S. J., Odion, R. A., Strobbia, P., Menozzi, L., et al. (2023). Plasmonic nanorod probes' journey inside plant cells for in *vivo* SERS sensing and multimodal imaging. *Nanoscale* 15, 6396–6407. doi: 10.1039/D2NR06235F

Devitt, G., Howard, K., Mudher, A., and Mahajan, S. (2018). Raman spectroscopy: an emerging tool in neurodegenerative disease research and diagnosis. *ACS Chem. Neurosci.* 9, 404–420. doi: 10.1021/acschemneuro.7b00413

Dhanani, T., Dou, T., Biradar, K., Jifon, J., Kurouski, D., and Patil, B. S. (2022). Raman spectroscopy detects changes in carotenoids on the surface of watermelon fruits during maturation. *Front. Plant Sci.* 13, 832522. doi: 10.3389/fpls.2022.832522

Dina, N. E., Muntean, C. M., Bratu, I., Tican, A., Halmagyi, A., Purcaru, M. A., et al. (2022). Structure and surface dynamics of genomic DNA as probed with surface-enhanced Raman spectroscopy: Trace level sensing of nucleic acids extracted from plants. *Spectrochimica Acta Part A: Mol. Biomolecular Spectrosc.* 279, 121477.

Dou, T., Sanchez, L., Irigoyen, S., Goff, N., Niraula, P., Mandadi, K., et al. (2021). Biochemical origin of raman-based diagnostics of huanglongbing in grapefruit trees. *Front. Plant Sci.* 12, 680991. doi: 10.3389/fpls.2021.680991

Edwards, H. G., Farwell, D. W., and Webster, D. (1997). FT Raman microscopy of untreated natural plant fibres. *Spectrochim. Acta A* 53, 2383–2392. doi: 10.1016/S1386-1425(97)00178-9

Egging, V., Nguyen, J., and Kurouski, D. (2018). Detection and identification of fungal infections in intact wheat and sorghum grain using a hand-held Raman spectrometer. *Analytical Chem.* 90, 8616–8621. doi: 10.1021/acs.analchem.8b01863

Engelsen, S. B., and Nørgaard, L. (1996). Comparative vibrational spectroscopy for determination of quality parameters in amidated pectins as evaluated by chemometrics. *Carbohydr. Polymers* 30, 9–24. doi: 10.1016/S0144-8617(96)00068-9

Fao, F. (2018). The impact of disasters and crises on agriculture and food security. *Report.* 32.

Farber, C., Bennett, J. S., Dou, T., Abugalyon, Y., Humpal, D., Sanchez, L., et al. (2021a). Raman-based diagnostics of stalk rot disease of maize caused by Colletotrichum graminicola. *Front. Plant Sci.* 12, 722898. doi: 10.3389/fpls.2021.722898

Farber, C., Islam, A. F., Septiningsih, E. M., Thomson, M. J., and Kurouski, D. (2021b). Non-invasive identification of nutrient components in grain. *Molecules* 26. doi: 10.3390/molecules26113124

Farber, C., Li, J., Hager, E., Chemelewski, R., Mullet, J., Rogachev, A. Y., et al. (2019). Complementarity of raman and infrared spectroscopy for structural characterization of plant epicuticular waxes. *ACS Omega* 4, 3700–3707. doi: 10.1021/acsomega.8b03675

Farber, C., Sanchez, L., Rizevsky, S., Ermolenkov, A., McCutchen, B., Cason, J., et al. (2020). Raman spectroscopy enables non-invasive identification of peanut genotypes and value-added traits. *Sci. Rep.* 10.

Farber, C., Shires, M., Ueckert, J., Ong, K., and Kurouski, D. (2023). Detection and differentiation of herbicide stresses in roses by Raman spectroscopy. *Front. Plant Sci.* 14, 1121012. doi: 10.3389/fpls.2023.1121012

Ferdous, Z., and Nemmar, A. (2020). Health impact of silver nanoparticles: a review of the biodistribution and toxicity following various routes of exposure. *Int. J. Mol. Sci.* 21. doi: 10.3390/ijms21072375

Geskovski, N., Stefkov, G., Gigopulu, O., Stefov, S., Huck, C. W., and Makreski, P. (2021). Mid-infrared spectroscopy as process analytical technology tool for estimation of THC and CBD content in Cannabis flowers and extracts. *Spectrochimica Acta Part A: Mol. Biomolecular Spectrosc.* 251, 119422.

Ghosh, D., Kokane, S., Savita, B. K., Kumar, P., Sharma, A. K., Ozcan, A., et al. (2022). Huanglongbing pandemic: current challenges and emerging management strategies. *Plants* 12, 160. doi: 10.3390/plants12010160

Goff, N. K., Guenther, J. F., Roberts, J. K. III, Adler, M., Molle, M. D., Mathews, G., et al. (2022). Non-invasive and confirmatory differentiation of hermaphrodite from both male and female cannabis plants using a hand-held raman spectrometer. *Molecules* 27. doi: 10.3390/molecules27154978

González Moreno, A., Domínguez, E., Mayer, K., Xiao, N., Bock, P., Heredia, A., et al. (2023). 3D (xyt) Raman imaging of tomato fruit cuticle: Microchemistry during development. *Plant Physiol.* 191, 219–232.

Gordon, K. C., and McGoverin, C. M. (2011). Raman mapping of pharmaceuticals. *Int. J. Pharmaceutics* 417, 151–162. doi: 10.1016/j.ijpharm.2010.12.030

Gupta, S., Huang, C. H., Singh, G. P., Park, B. S., Chua, N.-H., and Ram, R. J. (2020). Portable Raman leaf-clip sensor for rapid detection of plant stress. *Sci. Rep.* 10, 20206. doi: 10.1038/s41598-020-76485-5

Hara, R., Ishigaki, M., Ozaki, Y., Ahamed, T., Noguchi, R., Miyamoto, A., et al. (2021). Effect of Raman exposure time on the quantitative and discriminant analyses of carotenoid concentrations in intact tomatoes. *Food Chem.* 360, 129896. doi: 10.1016/j.foodchem.2021.129896

Hariharan, G., and Prasannath, K. (2021). Recent advances in molecular diagnostics of fungal plant pathogens: A mini review. *Front. Cell. Infection Microbiol.* 10, 600234. doi: 10.3389/fcimb.2020.600234

Higgins, S., Jessup, R., and Kurouski, D. (2022a). Raman spectroscopy enables highly accurate differentiation between young male and female hemp plants. *Planta* 255, 85. doi: 10.1007/s00425-022-03865-8

Higgins, S., Joshi, R., Juarez, I., Bennett, J. S., Holman, A. P., Kolomiets, M., et al. (2023). Non-invasive identification of combined salinity stress and stalk rot disease caused by Colletotrichum graminicola in maize using Raman spectroscopy. *Sci. Rep.* 13.

Higgins, S., Serada, V., Herron, B., Gadhave, K. R., and Kurouski, D. (2022b). Confirmatory detection and identification of biotic and abiotic stresses in wheat using Raman spectroscopy. *Front. Plant Sci.* 13, 1035522. doi: 10.3389/fpls.2022.1035522

Huang, C. H., Singh, G. P., Park, S. H., Chua, N.-H., Ram, R. J., and Park, B. S. (2020). Early diagnosis and management of nitrogen deficiency in plants utilizing Raman spectroscopy. *Front. Plant Sci.* 11, 663.doi: 10.3389/fpls.2020.00663

Jamieson, L. E., Li, A., Faulds, K., and Graham, D. (2018). Ratiometric analysis using Raman spectroscopy as a powerful predictor of structural properties of fatty acids. *R Soc. Open Sci.* 5, 181483. doi: 10.1098/rsos.181483

Jamshidi, B., Mohajerani, E., Jamshidi, J., Minaei, S., and Sharifi, A. (2015). Non-destructive detection of pesticide residues in cucumber using visible/near-infrared spectroscopy. *Food Additives Contaminants: Part A* 32, 857–863. doi: 10.1080/19440049.2015.1031192

Jiang, H., Zhu, H., Yu, T., Song, W., Zhou, B., Qu, C., et al. (2023a). Non-amplification on-spot identifying the sex of dioecious kiwi plants by a portable Raman device. *Talanta* 258, 124447. doi: 10.1016/j.talanta.2023.124447

Jiang, L., Zhang, Y., Kang, C., Zhao, Z., Chen, D., and Long, Y. (2023b). Nondestructive Determination of Carotenoids in Kiwifruit Leaves Infected with Pseudomonas syringae pv. actinidiae by Surface-enhanced Raman Spectroscopy Combined with Chemical Imaging. *Plant Pathology* 1022-1033. doi: 10.1111/ppa.13734

Käfer, J., Marais, G. A., and Pannell, J. R. (2017). On the rarity of dioecy in flowering plants. *Mol. Ecol.* 26, 1225–1241. doi: 10.1111/mec.14020

Kang, L., Wang, K., Li, X., and Zou, B. (2016). High pressure structural investigation of benzoic acid: raman spectroscopy and x-ray diffraction. *J. Phys. Chem. C.* 120, 14758–14766. doi: 10.1021/acs.jpcc.6b05001

Kim, S., Lee, S., Chi, H.-Y., Kim, M.-K., Kim, J.-S., Lee, S.-H., et al. (2013). Feasibility study for detection of turnip yellow mosaic virus (TYMV) Infection of Chinese Cabbage Plants Using Raman Spectroscopy. *Plant Pathol. J.* 29, 105. doi: 10.5423/PPJ.NT.09.2012.0147

Lau, H. Y., Wang, Y., Wee, E. J., Botella, J. R., and Trau, M. (2016). Field demonstration of a multiplexed point-of-care diagnostic platform for plant pathogens. *Analytical Chem.* 88, 8074–8081. doi: 10.1021/acs.analchem.6b01551

Li, Z., and Ai, Z. (2023). Mapping plant bioaccumulation potentials of pesticides from soil using satellite-based canopy transpiration rates. *Environ. Toxicol. Chem.* 42, 117–129. doi: 10.1002/etc.5511

Li, J., Wijesooriya, C. S., Burkhow, S. J., Brown, L. K., Collet, B. Y., Greaves, J. A., et al. (2021). Measuring plant metabolite abundance in spearmint (Mentha spicata L.) with Raman spectra to determine optimal harvest time. *ACS Food Sci. Technol.* 1, 1023–1029. doi: 10.1021/acsfoodscitech.1c00047

Liang, P.-S., Haff, R. P., Hua, S.-S. T., Munyaneza, J. E., Mustafa, T., and Sarreal, S. B. L. (2018). Nondestructive detection of zebra chip disease in potatoes using near-infrared spectroscopy. *Biosyst. Eng.* 166, 161–169. doi: 10.1016/j.biosystemseng.2017.11.019

Liu, X., Renard, C. M., Bureau, S., and Le Bourvellec, C. (2021). Revisiting the contribution of ATR-FTIR spectroscopy to characterize plant cell wall polysaccharides. *Carbohydr. Polymers* 262, 117935. doi: 10.1016/j.carbpol.2021.117935

Lu, Y., Li, X., Li, W., Shen, T., He, Z., Zhang, M., et al. (2021). Detection of chlorpyrifos and carbendazim residues in the cabbage using visible/near-infrared spectroscopy combined with chemometrics. *Spectrochimica Acta Part A: Mol. Biomolecular Spectrosc.* 257, 119759.

Mandrile, L., D'Errico, C., Nuzzo, F., Barzan, G., Matić, S., Giovannozzi, A. M., et al. (2022). Raman spectroscopy applications in grapevine: Metabolic analysis of plants infected by two different viruses. *Front. Plant Sci.* 13, 917226. doi: 10.3389/fpls.2022.917226

Miao, X., Miao, Y., Liu, Y., Tao, S., Zheng, H., Wang, J., et al. (2023). Measurement of nitrogen content in rice plant using near infrared spectroscopy combined with different PLS algorithms. *Spectrochimica Acta Part: Mol. Biomolecular Spectrosc.* 284, 121733.

Morey, R., Ermolenkov, A., Payne, W. Z., Scheuring, D. C., Koym, J. W., Vales, M. I., et al. (2020). Non-invasive identification of potato varieties and prediction of the origin of tuber cultivation using spatially offset Raman spectroscopy. *Analytical Bioanalytical Chem.* 412, 4585–4594. doi: 10.1007/s00216-020-02706-5

Mu, X., and Chen, Y. (2021). The physiological response of photosynthesis to nitrogen deficiency. *Plant Physiol. Biochem.* 158, 76–82. doi: 10.1016/j.plaphy.2020.11.019

Muik, B., Lendl, B., Molina-Díaz, A., Ortega-Calderón, D., and Ayora-Cañada, M. J. (2004). Discrimination of olives according to fruit quality using Fourier transform Raman spectroscopy and pattern recognition techniques. *J. Agric. Food Chem.* 52, 6055–6060. doi: 10.1021/jf049240e

Nikbakht, A. M., TAVAKKOLI, H. T., Malekfar, R., and Gobadian, B. (2011). Nondestructive determination of tomato fruit quality parameters using Raman spectroscopy. *J. Agr. Sci. Tech.* 13, 517-526.

Pan, T.-T., Pu, H., and Sun, D.-W. (2017). Insights into the changes in chemical compositions of the cell wall of pear fruit infected by Alternaria alternata with confocal Raman microspectroscopy. *Postharv. Biol. Technol.* 132, 119–129. doi: 10.1016/j.postharvbio.2017.05.012

Parlamas, S., Goetze, P. K., Humpal, D., Kurouski, D., and Jo, Y.-K. (2022). Raman spectroscopy enables confirmatory diagnostics of fusarium wilt in asymptomatic banana. *Front. Plant Sci.* 13, 922254. doi: 10.3389/fpls.2022.922254

Payne, W. Z., Dou, T., Cason, J. M., Simpson, C. E., McCutchen, B., Burow, M. D., et al. (2022). A proof-of-principle study of non-invasive identification of peanut genotypes and nematode resistance using raman spectroscopy. *Front. Plant Sci.* 12, 664243. doi: 10.3389/fpls.2021.664243

Pettolino, F. A., Walsh, C., Fincher, G. B., and Bacic, A. (2012). Determining the polysaccharide composition of plant cell walls. *Nat. Protoc.* 7, 1590–1607. doi: 10.1038/nprot.2012.081

Ploetz, R. C. (2021). Gone bananas? Current and future impact of fusarium wilt on production. *Plant Dis. Food Secur. 21st century. Springer)* 21-32.

Programme, U. N. E. (2022). "Plastics in agriculture – an environmental challenge," in *Foresight brief 029* (Nairobi).

Ristaino, J. B., Anderson, P. K., Bebber, D. P., Brauman, K. A., Cunniffe, N. J., Fedoroff, N. V., et al. (2021). The persistent threat of emerging plant disease pandemics to global food security. *Proc. Natl. Acad. Sci.* 118, e2022239118. doi: 10.1073/pnas.2022239118

Rubio, L., Galipienso, L., and Ferriol, I. (2020). Detection of plant viruses and disease management: Relevance of genetic diversity and evolution. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.01092

Rys, M., Juhasz, C., Surowka, E., Janeczko, A., Saja, D., Tobias, I., et al. (2014). Comparison of a compatible and an incompatible pepper-tobamovirus interaction by biochemical and non-invasive techniques: chlorophyll a fluorescence, isothermal calorimetry and FT-Raman spectroscopy. *Plant Physiol. Biochem.* 83, 267–278. doi: 10.1016/j.plaphy.2014.08.013

Sanaeifar, A., Li, X., He, Y., Huang, Z., and Zhan, Z. (2021). A data fusion approach on confocal Raman microspectroscopy and electronic nose for quantitative evaluation of pesticide residue in tea. *Biosyst. Eng.* 210, 206–222. doi: 10.1016/j.biosystemseng.2021.08.016

Sanchez, L., Baltensperger, D., and Kurouski, D. (2020a). Raman-based differentiation of hemp, cannabidiol-rich hemp, and cannabis. *Analytical Chem.* 92, 7733–7737. doi: 10.1021/acs.analchem.0c00828

Sanchez, L., Ermolenkov, A., Tang, X.-T., Tamborindeguy, C., and Kurouski, D. (2020b). Non-invasive diagnostics of Liberibacter disease on tomatoes using a hand-held Raman spectrometer. *Planta* 251, 1–6. doi: 10.1007/s00425-020-03359-5

Sanchez, L., Filter, C., Baltensperger, D., and Kurouski, D. (2020c). Confirmatory non-invasive and non-destructive differentiation between hemp and cannabis using A hand-held raman spectrometer. *RCS Adv.* 10, 3212–3216. doi: 10.1039/C9RA08225E

Sanchez, L., Pant, S., Irey, M., Mandadi, K., and Kurouski, D. (2019a). Detection and identification of canker and blight on orange trees using a hand-held Raman spectrometer. *J. Raman Spectrosc.* 50, 1875–1880. doi: 10.1002/jrs.5741

Sanchez, L., Pant, S., Mandadi, K., and Kurouski, D. (2020e). Raman spectroscopy vs quantitative polymerase chain reaction in early stage Huanglongbing diagnostics. *Sci. Rep.* 10, 10101. doi: 10.1038/s41598-020-67148-6

Sanchez, L., Pant, S., Xing, Z., Mandadi, K., and Kurouski, D. (2019b). Rapid and noninvasive diagnostics of Huanglongbing and nutrient deficits on citrus trees with a handheld Raman spectrometer. *Anal. Bioanal Chem.* 411, 3125–3133. doi: 10.1007/s00216-019-01776-4

Sasani, N., Bock, P., Felhofer, M., and Gierlinger, N. (2021). Raman imaging reveals *in-situ* microchemistry of cuticle and epidermis of spruce needles. *Plant Methods* 17, 1–15. doi: 10.1186/s13007-021-00717-6

Schulz, H., Baranska, M., and Baranski, R. (2005). Potential of NIR-FT-Raman spectroscopy in natural carotenoid analysis. *Biopolymers* 77, 212–221. doi: 10.1002/bip.20215

Shrivastava, M., Srivastav, A., Gandhi, S., Rao, S., Roychoudhury, A., Kumar, A., et al. (2019). Monitoring of engineered nanoparticles in soil-plant system: A review. *Environ. nanotechnology Monit. Manage.* 11, 100218. doi: 10.1016/j.enmm.2019.100218

Sicbaldi, F., Sacchi, G. A., Trevisan, M., and Del Re, A. A. (1997). Root uptake and xylem translocation of pesticides from different chemical classes. *Pesticide Sci.* 50, 111–119. doi: 10.1002/(SICI)1096-9063(199706)50:2<>1.0.CO;2-8

Singh, V., Dou, T., Krimmer, M., Singh, S., Humpal, D., Payne, W. Z., et al. (2021). Raman Spectroscopy Can Distinguish Glyphosate-Susceptible and-Resistant Palmer Amaranth (Amaranthus palmeri). *Front. Plant Sci.* 12, 657963. doi: 10.3389/fpls.2021.657963

Skolik, P., McAinsh, M. R., and Martin, F. L. (2019). ATR-FTIR spectroscopy non-destructively detects damage-induced sour rot infection in whole tomato fruit. *Planta* 249, 925–939. doi: 10.1007/s00425-018-3060-1

Son, W. K., Choi, Y. S., Han, Y. W., Shin, D. W., Min, K., Shin, J., et al. (2023). *In vivo* surface-enhanced Raman scattering nanosensor for the real-time monitoring of multiple stress signalling molecules in plants. *Nat. Nanotechnology* 18, 205–216. doi: 10.1038/s41565-022-01274-2

Synytsya, A., Čopíková, J., Matějka, P., and Machovič, V. (2003). Fourier transform Raman and infrared spectroscopy of pectins. *Carbohydr. Polym.* 54, 97–106. doi: 10.1016/S0144-8617(03)00158-9

Tympa, L.-E., Katsara, K., Moschou, P. N., Kenanakis, G., and Papadakis, V. M. (2021). Do microplastics enter our food chain via root vegetables? A raman based spectroscopic study on Raphanus sativus. *Materials* 14. doi: 10.3390/ma14092329

Wang, X., Xie, H., Wang, P., and Yin, H. (2023). Nanoparticles in plants: uptake, transport and physiological activity in leaf and root. *Materials* 16. doi: 10.3390/ma16083097

Wiercigroch, E., Szafraniec, E., Czamara, K., Pacia, M. Z., Majzner, K., Kochan, K., et al. (2017). Raman and infrared spectroscopy of carbohydrates: A review. *Spectrochim. Acta A* 185, 317–335. doi: 10.1016/j.saa.2017.05.045

Xu, J., Meng, J., and Quackenbush, L. J. (2019). Use of remote sensing to predict the optimal harvest date of corn. *Field Crops Res.* 236, 1–13. doi: 10.1016/j.fcr.2019.03.003

Yang, T., Doherty, J., Guo, H., Zhao, B., Clark, J. M., Xing, B., et al. (2019). Real-time monitoring of pesticide translocation in tomato plants by surface-enhanced Raman spectroscopy. *Analytical Chem.* 91, 2093–2099. doi: 10.1021/acs.analchem.8b04522

Yang, T., Zhao, B., Kinchla, A. J., Clark, J. M., and He, L. (2017). Investigation of pesticide penetration and persistence on harvested and live basil leaves using surface-enhanced Raman scattering mapping. *J. Agric. Food Chem.* 65, 3541–3550. doi: 10.1021/acs.jafc.7b00548

Yao, L., Xu, J., Cheng, J., Yao, B., Zheng, L., Liu, G., et al. (2022). Simultaneous and accurate screening of multiple genetically modified organism (GMO) components in food on the same test line of SERS-integrated lateral flow strip. *Food Chem.* 366, 130595. doi: 10.1016/j.foodchem.2021.130595

Yilmaz, M., Yilmaz, A., Karaman, A., Aysin, F., and Aksakal, O. (2021). Monitoring chemically and green-synthesized silver nanoparticles in maize seedlings via surface-enhanced Raman spectroscopy (SERS) and their phytotoxicity evaluation. *Talanta* 225, 121952. doi: 10.1016/j.talanta.2020.121952

Yin, L., Jayan, H., Cai, J., El-Seedi, H. R., Guo, Z., and Zou, X. (2023). Development of a sensitive SERS method for label-free detection of hexavalent chromium in tea using carbimazole redox reaction. *Foods* 12. doi: 10.3390/foods12142673

Yu, M. M., Schulze, H. G., Jetter, R., Blades, M. W., and Turner, R. F. (2007). Raman microspectroscopic analysis of triterpenoids found in plant cuticles. *Appl. Spectrosc.* 61, 32–37. doi: 10.1366/000370207779701352

Zandalinas, S. I., Fritschi, F. B., and Mittler, R. (2021). Global warming, climate change, and environmental pollution: recipe for a multifactorial stress combination disaster. *Trends Plant Sci.* 26, 588–599. doi: 10.1016/j.tplants.2021.02.011

Zeng, J., Ping, W., Sanaeifar, A., Xu, X., Luo, W., Sha, J., et al. (2021). Quantitative visualization of photosynthetic pigments in tea leaves based on Raman spectroscopy and calibration model transfer. *Plant Methods* 17, 1–13. doi: 10.1186/s13007-020-00704-3

Zhang, Y., Luan, Q., Jiang, J., and Li, Y. (2021). Prediction and utilization of malondialdehyde in exotic pine under drought stress using near-infrared spectroscopy. *Front. Plant Sci.* 12, 735275. doi: 10.3389/fpls.2021.735275

Zhou, S., Lu, C., Li, Y., Xue, L., Zhao, C., Tian, G., et al. (2020). Gold nanobones enhanced ultrasensitive surface-enhanced Raman scattering aptasensor for detecting Escherichia coli O157: H7. *ACS sensors* 5, 588–596. doi: 10.1021/acssensors.9b02600

frontiers | Frontiers in Plant Science

TYPE Methods
PUBLISHED 09 October 2024
DOI 10.3389/fpls.2024.1469675

OPEN ACCESS

EDITED BY
Miha Humar,
University of Ljubljana, Slovenia

REVIEWED BY
Xinsheng Liu,
Anhui Normal University, China
Ákos Malatinszky,
Hungarian University of Agricultural and Life Sciences, Hungary

*CORRESPONDENCE
Jianfeng Peng
jfpeng@vip.henu.edu.cn
Yafei Wei
weiyafe@henu.edu.cn

†These authors have contributed equally to this work and share first authorship

RECEIVED 24 July 2024
ACCEPTED 19 September 2024
PUBLISHED 09 October 2024

CITATION
Li J, Liu Y, Wei Y, Li J, Zhang K, Wei X and Peng J (2024) Dendroclimatological study of ancient trees integrating non-destructive techniques.
Front. Plant Sci. 15:1469675.
doi: 10.3389/fpls.2024.1469675

Dendroclimatological study of ancient trees integrating non-destructive techniques

Jinkuan Li[1†], Yameng Liu[1†], Yafei Wei[1*], Jiaxin Li[1], Keyu Zhang[1], Xiaoxu Wei[1] and Jianfeng Peng[1,2*]

[1]College of Geography and Environmental Science, Henan University, Kaifeng, China, [2]The Key Laboratory of Earth System Observation and Simulation of Henan Province, Kaifeng, China

Based on the need to protect previous ancient trees and the development of dendroclimatology, the use of non-destructive technologies in tree-ring research has gained increasing attention. This study focuses on the ancient *Pinus tabulaeformis* in Yu Xiang Forest Farm in Henan Province. Firstly, samples were collected using the traditional Increment borers and the Resistograph, a non-destructive method. Subsequently, the peak-valley analysis was used to filter the data obtained by the Resistograph to extract the tree ring width sequence, and the data's accuracy was verified by correlation analysis with tree ring width sequence by the Increment borers. Then, the optimal filtering method and an appropriate comprehensive threshold were determined, and tree ring width and density sequences were successfully extracted. Following that, the growth trend and residual resistance in the measurement process were corrected using linear fitting and Ensemble Empirical Mode Decomposition (EEMD) technology, thereby establishing the tree-ring width and density index series, which were further validated through correlation analysis and t-tests. Finally, analysis of the correlation with climatic factors, identified the main limiting factors for tree growth, and the accuracy of the tree-ring information extracted by the Resistograph was further verified. The results showed that spite of certain differences between the tree-ring width indices extracted by the Resistograph and the Increment borer, they were generally reliable. The radial growth of the ancient *P.tabulaeformis* in Yu Xiang Forest Farm is primarily influenced by temperature, with the maximum density of the tree rings responding more significantly to the mean maximum temperature, while the minimum density of the tree rings responded more significantly to the mean minimum temperature. These results not only provide a scientific and accurate age for the protection of ancient trees and verify the reliability of the data obtained by the Resistograph, but also facilitate the use of non-destructive technology for in-depth study of ancient trees, therefore enhancing our understanding of how climate change affects tree growth and provide valuable insights for the future protection and management of these ancient trees.

KEYWORDS

ancient trees, non-destructive technology, peak-valley analysis, tree ring index, climate response

1 Introduction

Global warming has become an undeniable fact (IPCC, 2021). With the warming of the global climate, the growth environment of trees faces huge challenges. In recent years, climate warming has led to increasing drought events and subsequent forest decline or death (Peng et al., 2011; Gauthier et al., 2015; Millar and Stephenson, 2015). As an important part of nature, tree growth not only is related to ecological balance, but also directly affects the living environment of human beings. Ancient trees play a crucial role in maintaining biodiversity, ecological balance, carbon storage, soil and water conservation, microclimate regulation, and providing aesthetic and cultural value (Seibold et al., 2018; Wu, 2019). They are indispensable components of ecosystems, providing habitats for many species and serving as important resources for scientific research and education (Takács and Malatinszky, 2021). Protecting ancient trees helps ensure the long-term preservation and utilization of these valuable resources (Gao, 2023; Ren et al., 2024).

Traditional tree ring research has some limitations in exploring environmental information of ancient trees because the protection and possible influence of ancient trees should be considered in sample collection (Xu et al., 2012). With the development of science and technology, a series of innovations have occurred in sampling technology, from the traditional Increment borers sampling to X-ray (Polge, 1966; Parker and Meleskie, 1970), stress wave (Bulleit and Falk, 1985; Schad et al., 1996), ultrasonic detection (Bauer et al., 1991; Yang and Wang, 2010), and Resistograph (Wang et al., 2013; Xia, 2023), and other non-destructive non-sampling detection equipment. These innovations, therefore, greatly improve the efficiency and convenience of data collection. However, most of the equipment is bulky and inconvenient to carry and use in the field. In contrast, the Resistograph is easy to carry, simple to use, low cost, and widely applicable, thus becoming an economical and efficient non-destructive tree ring measurement tool (Wang et al., 2013; Xia, 2023). Moreover it has gradually been applied in the analyses of tree rings (Rinn et al., 1996; Chantre and Rozenberg, 1997; Wang and Lin, 2001; Lima et al., 2007; Acuna et al., 2011; Liang, 2017; Wei, 2018; Pan, 2020; Xia, 2023; Zhang, 2024), tree decay detections (Rinn, 1994; Costello and Quarles, 1999; Ceraldi et al., 2001; Wu, 2011a; Wu et al., 2011b; Zhu et al., 2013; Nutto and Biechele, 2015), and the assessment of wood structure status (Winistorfer et al., 1995; Isik and Li, 2003; Zhang et al., 2007; Ukrainetz and O'Neill, 2010; Sun et al., 2011; Sun, 2012) and other fields. Using the Resistograph identification method, we can more comprehensively understand the history and growth of ancient trees, providing a scientific basis for their protection and inheritance. However, the issue of dating accuracy has not yet been resolved. The application of the Resistograph, both domestically and internationally, has not been widely popularized, especially in tree ring width and density research.

Henan Province, located in the central part of China, is crossed by the Qin-Huai line in its southern region and mostly features a warm temperate continental monsoon climate (Wang et al., 1990). The natural geographical environment is complex and diverse, and the historical and cultural heritage is profound, with a wealth of ancient tree resources. As a key area for ecological protection and high-quality development in the Yellow River basin, the ancient tree resources of Henan Province are of great significance to the construction of ecological civilization and the inheritance of historical culture (Ren et al., 2024). Therefore, it is critical/important to use ancient trees to study long-term climate change in this region for clarifying the evolution of civilization in the Central Plains.

The objectives of this study were to: 1) determine the optimal filtering method and the best comprehensive threshold for extracting tree ring information from ancient trees using the Resistograph; 2) establish chronologies of tree ring width and density (mean density, minimum density, and maximum density) indices; 3) identify and analyze the main climatic limiting factors that affect tree growth; 4) ascertain the reliability of the Resistograph sequence and the feasibility of ancient tree-ring research. The study also aimed to provide a new perspective for the extraction of tree ring information using the Resistograph and to offer a scientific basis for the protection and management of ancient trees.

2 Materials and methods

2.1 The study area

Yuxiang Forest Farm (34°28'-34°29'N, 114°56'-114°57'E) is located in Suixian County, Shangqiu City, Henan Province, covering an area of nearly 70 km^2.The area is situated in the eastern plains of Henan Province, within the Yellow River alluvial fan region, with flat terrain and an elevation ranging from 51 to 60 meters. Climatically, Yuxiang Forest Farm has a warm temperate semi-humid continental monsoon climate with mildly wet summers and cold dry winters (Zhang et al., 2023; Ji, 2022), as depicted in Figure 1. It has moderate and evenly distributed precipitation, providing favorable conditions for vegetation growth (Li and Jing, 1999). Despite the frequent human activity, low altitude, and relatively infertile soil conditions, the forest still stands tall with 96 ancient *Pinus tabuliformis*, each measuring 40 centimeters in diameter and nearing a century in age. These ancient trees have been officially recorded as part of the *P.tabuliformis* ancient tree group during the national survey of ancient tree clusters. They not only demonstrate remarkable survival capabilities and adaptability to their environment, but also have distinct annual rings, making them a valuable resource for dendroclimatic research.

2.2 Research methods

2.2.1 Introduction to the Resistograph

The Resistograph used in this study is the Germany-made Resistograph PD600 model (Figure 2A). The probe needle of this device is made of special steel with a diameter of only 1.5mm (Figure 2B), capable of measuring trees with a diameter of up to 200cm. The resistance curve diagram generated can provide key data such as drilling resistance (DR, the resistance encountered by the drill bit during rotation), feed resistance (FR, the resistance

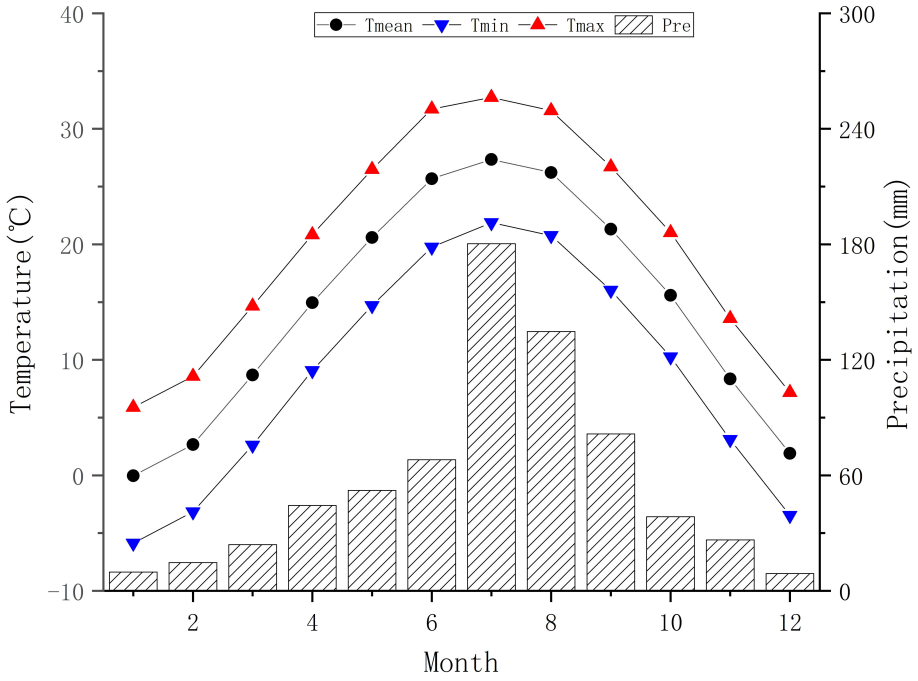


FIGURE 1
Gridded monthly meteorological data statistical chart for sampling in Yu Xiang Forest Farm.

encountered by the drill bit during advancement), and core information (Figure 2C). These curve diagrams can not only intuitively display the internal structure of the trees but also identify problems such as decay or cavities. Data are stored in the built-in microcomputer of the device and can be further processed and analyzed through the PD-Tools Pro software.

Detection Principle: The Resistograph is based on the resistance changes when the probe drills into the xylem, enabling it to infer wood characteristics such as age, tree ring width, and density. The advantages include rapid and accurate detection capabilities, simple operation, and non-destructive in nature, making it suitable for dendroclimatic studies of ancient trees (Yao et al., 2022a; b; 2023).

2.2.2 Samples collection

In October 2023, samples were collected from Yuxiang Forest Farm. The study adhered to the standards of the International Tree-Ring Data Bank (ITRDB), selecting thicker trees for sampling based on the sensitivity and replication principles of the trees. With an Increment borer, 2 cores were drilled from the base of each tree. To prevent sample damage, the samples were quickly placed into test tubes for preservation and numbered, for a total of 20 trees and 40 cores. Subsequently, in order to facilitate comparison with the Increment borer samples, the Resistograph was used to align the probe 3 cm above or 3 cm below the Increment borers sampling point. After starting the device, the probe was steadily inserted into the tree, recording the resistance changes and exporting the resistance curve from bark-pith-bark. It is necessary to keep the device stable during the measurement, and after the measurement, the data were then brought back to the laboratory for analysis.

2.2.3 Tree-ring information based on the increment borers

In the laboratory, all core samples were fixed, air-dried, sanded, measured for width, and cross-dated according to the methods of Stokes ma Smiley-TL (1968) and others in the ITRDB. The cross-dating results were corrected using the COFECHA program (Holmes, 1983), and any missing or false tree rings in the samples were corrected and quality controlled. Samples with low consistency with the main sequence were excluded. Finally, 20 trees with 40 core samples were retained. The age and ring width of each core sample were determined.

2.2.4 Extraction and verification of tree-ring information based on the Resistograph

In the laboratory, the data from the Resistograph were processed using Excel software to eliminate errors that might be caused by the drill bit entering and exiting the tree, and to exclude the potential impact of bark thickness on the measurement results, thus reducing the overall error range. The density differences between the earlywood (EW) and latewood (LW) of the tree cause resistance values to fluctuate (Yao et al., 2022c). The resistance line graph illustrates these variations with multiple peaks and valleys. However, not all peaks and valleys represent real tree ring changes; some minor fluctuations are indicative of false rings (Yao et al., 2022d; Pan, 2020). Therefore, setting an appropriate resistance threshold (Det) is necessary to distinguish between real tree rings and false rings.

The specific process for extracting the tree ring width sequence is as follows:

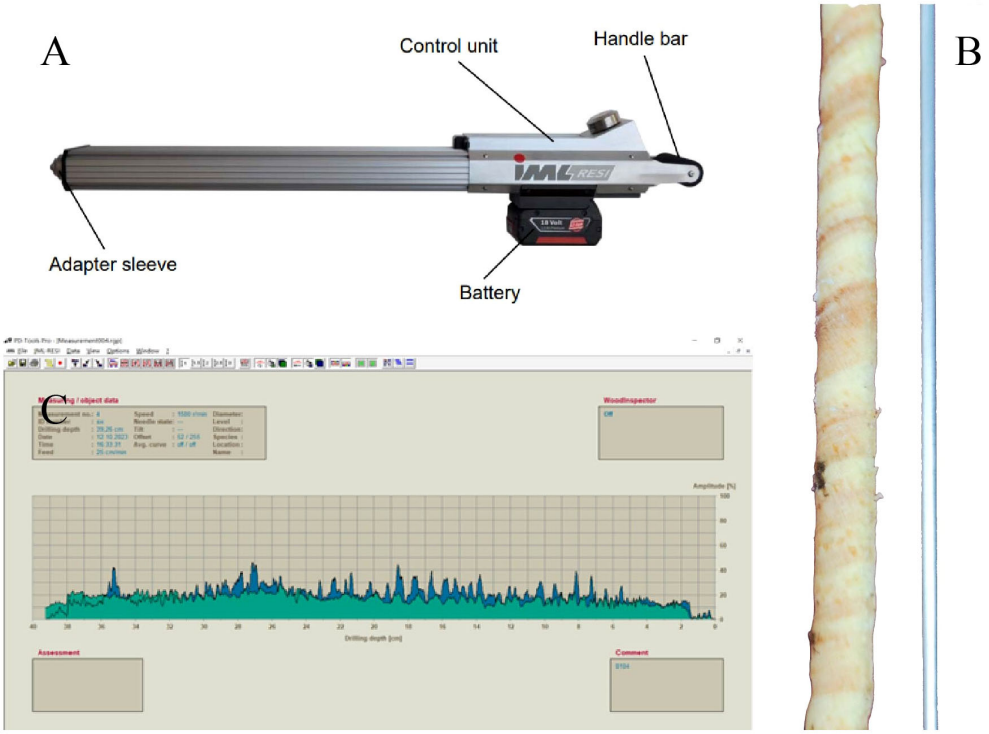


FIGURE 2
Resistograph and its application [**(A)**: Picture of the Resistograph; **(B)** Size comparison between the steel needle of the Resistograph (diameter 1.5mm) and the Increment borers sample embedded in the wooden slot (diameter 5.15mm); **(C)** Resistance curve exported by the Resistograph, green represents DR, blue represents FR, with a data resolution of 1/10 mm].

(1) Code the data from the Resistograph;

(2) Extract all extreme points;

(3) Filter the Resistograph data using several different methods of resistance threshold (Det) to remove minor peaks and valleys;

(4) Extract the tree ring width sequence using valid peaks and valleys (the encoding difference between extremes is directly related to the measurement of tree ring width, considering that the measurement accuracy of the Resistograph PD600 is 0.1mm).

2.2.4.1 Selection of the optimal filtering method

Based on past research and tree growth trends, eight threshold methods were selected to filter the resistance curve, making the age extracted by the Resistograph closest to the age extracted by the increment borers. The methods are as follows:

A. Fixed threshold;
B. Linear threshold starting at 0 (assign the value for the Resistograph near the bark as 2022, and the innermost value as the starting year determined by the Increment borers, with linear interpolation in between. Based on the growth trend of the tree ring width extracted by the Increment borers, a linear threshold is set. If the growth trend of the tree ring width extracted by the Increment borers continues to decline and the fitting curve is always greater than 0, then the threshold for the Resistograph to extract the tree ring width is set as a linear threshold that continues to decline to 0; if the growth trend of the tree ring width extracted by the Increment borers continues to decline and fits to 0 after a certain number of years, then a linear threshold that declines and the fitting curve is greater than 0 is set for the years where the growth trend declines and the fitting curve is greater than 0, and a threshold of 0 is set for the years where the fitting curve is less than 0);
C. Linear threshold starting at 0.01;
D. Linear threshold starting at 0.02;
E. Linear threshold starting at 0.03;
F. Linear threshold starting at 1/10 of the maximum threshold value in method B;
G. Linear threshold starting at 2/10 of the maximum threshold value in method B;
H. Linear threshold starting at 3/10 of the maximum threshold value in method B.

The tree ring width sequences extracted by the Resistograph using the eight methods were analyzed for correlation with the tree ring width sequences extracted by the Increment borers to determine the optimal filtering method.

2.2.4.2 Selection of the optimal comprehensive threshold

The Resistograph assigns the value near the bark 2022 year, and the innermost value as the comprehensive starting year of the tree ring data extracted by the Increment borers (1944, Figure 3), with linear interpolation in between. The overall growth trend of the trees in this area continues to decline, fitting to 0 in the year 2018 (Figure 3), and the maximum threshold is determined based on the correlation between tree ring width sequence extracted by the Increment borers and the corresponding tree ring width sequence extracted by the Resistograph, with linear interpolation in between.

2.2.5 Tree-ring data standardization

The variation in tree ring width is influenced by genetic and environmental factors. Genetics mainly affect the ring width through tree age; as the tree ages, the ring width changes, a phenomenon known as "growth trend". When measuring with a Resistograph, the probe penetrates the tree rings, increasing the contact area and friction, which leads to measured values that include a combination of density and friction, with the friction part being residual resistance.

To improve the accuracy of ring width and density measurements, linear fitting and Ensemble Empirical Mode Decomposition (EEMD) are used to remove signals caused by the tree's own characteristics and non-synchronous microhabitat factors (Fritts, 1976), and residual resistance during density measurement (Xia, 2023). This helps accurately analyze and interpret tree ring data, and better understand the relationship between tree growth and climate change.

The principle of standardization: according to the fitted curves of tree width and density, the expected growth value for the tree species is calculated, reflecting the degree to which the ring is influenced by meteorological changes after excluding other influences. The formula is:

$$It = Wt/Yt$$

where: It is the ring index of ring width (density) in year t; Wt is the measured value of ring width (density) in year t; Yt is the expected value of ring width (density).

2.2.6 Acquisition of meteorological data

To address the differences in altitude and distance between the sampling site and the meteorological station, this study used the 1km resolution monthly meteorological dataset provided by the Qinghai-Tibet Plateau Data Center (Peng, 2020a; b; c; d) to extract monthly mean temperature (Tmean), mean minimum temperature (Tmin), mean maximum temperature (Tmax), and precipitation (Pre) near Yuxiang Forest Farm (34°28'47"-34°29'17"N, 114°56'31"-114°57'1"E) from 1950 to 2021 (Figure 1). Obviously, this area belongs to a continental monsoon climate, characterized by precipitation concentrated in July-August and concurrent precipitation and heat during the monsoon season. The mean annual Tmean is 14.44°C and mean annual precipitation is 683.57mm.

2.2.7 Correlation analyses

The DendroClim2002 (Biondi and Waikul, 2004) was used to analyze the correlation between the tree ring indices and climatic factors. The aim was to explore the response of tree growth to climate change and determine the main climatic factors limiting tree growth. Tree ring indices included ring width indices measured by Increment borers, as well as, ring width and density indices (mean, minimum, maximum) extracted from Resistograph curves. Considering the possible "lag effect" of climatic factors on tree growth, this study selected 21 months of climatic data from the previous March to the current November for analysis.

3 Results and analysis

3.1 Results of different filtering methods

The optimal filtering method was selected by comparing eight methods (Figure 3) through correlation analysis between the tree ring width sequences extracted by the Resistograph and the sequences measured by the Increment borers.

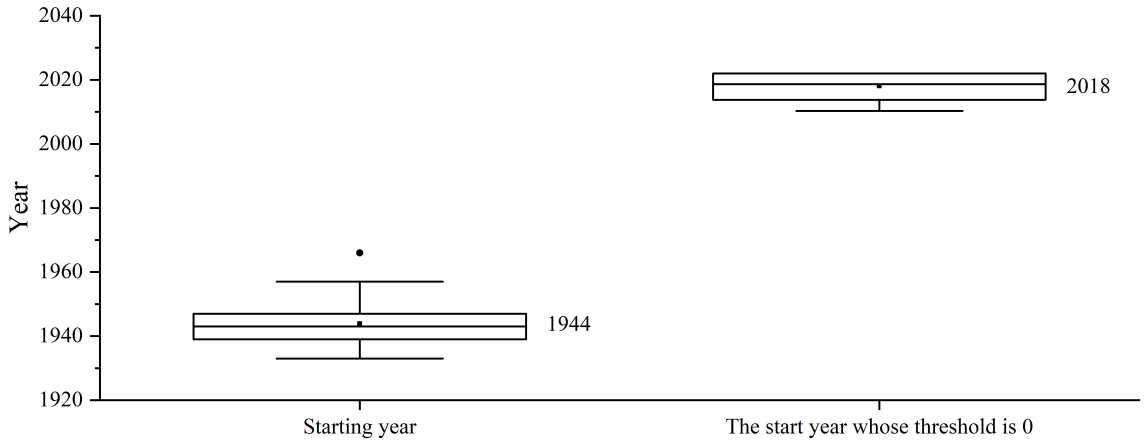


FIGURE 3
The mean start year of regional tree growth compared to the start year with an initial threshold of 0.

The correlation analysis (Table 1) revealed that there are a higher number of significant correlations between tree ring width sequences extracted by the Resistograph (A linear threshold whose initial threshold is 0, Table 1 B column) and the Increment borers, with the best correlation observed between the comprehensive sequences (Figure 4), indicating that the tree ring width sequence extracted by the Resistograph using an initial threshold value of 0 for the linear threshold is credible for both high-frequency (Figure 4A) and low-frequency signals (Figure 4B).

3.2 Optimal comprehensive threshold

Using an initial threshold value of 0(2018-2022, Figure 2), the maximum threshold values as shown Figure 5(i: The mean of all maximum threshold values; ii: iii: iv: v: The average values of the maximum thresholds for significant correlations in sequences I, II, III, and IV after removing outliers), for the linear threshold values in between, linear interpolation is applied.

Using the correlation analysis of the tree ring width sequences extracted by the Resistograph and the Increment borers, the optimal comprehensive threshold is selected from among five candidate thresholds.

Through correlation analysis, it was found that the tree ring width sequences, extracted using the Resistograph with five different methods and the Increment borers, have a higher number of significant correlations (Table 2), with the best correlation observed between the comprehensive sequences (Figure 6). However, method c is the best. This indicates that the tree ring width sequences extracted by the Resistograph using the comprehensive threshold value are credible for both high-frequency (Figure 6A) and low-frequency (Figure 6B) signals.

TABLE 1 Number of significant correlations in each method (The tree ring width sequences, including both the individual sequences extracted using the Resistograph and the Increment borers (I), their 5-year moving averages (II), the individual sequence extracted using the Resistograph and the comprehensive tree ring width sequence derived from the mean of all sequences using the Increment borers (III), their 5-year moving averages (IV)), the same below.

		A	B	C	D	E	F	G	H
I	Drilling Resistance	18	44	41	30	27	33	23	19
	Feed Resistance	3	43	38	37	37	32	13	10
II	Drilling Resistance	36	52	52	51	49	52	42	40
	Feed Resistance	19	52	52	51	51	49	31	17
III	Drilling Resistance	13	30	30	25	23	25	18	17
	Feed Resistance	7	30	29	26	26	24	13	8
IV	Drilling Resistance	22	31	31	31	29	31	27	24
	Feed Resistance	13	31	31	30	30	30	28	15

3.3 Establishment of tree ring indices

3.3.1 Tree ring width index

The growth trend of tree width is removed using linear fitting. Considering the characteristics of the tree ring width samples, at least six samples are required to establish a reliable chronology (as shown in Figure 7, the darker lines represent reliable chronologies, same below). The Increment borers extracted tree ring width index (IW), drilling resistance extracted tree ring width index (DW), and feed resistance extracted tree ring width index (FW) were ultimately obtained.

By comparing and analyzing the tree ring width indices and their first-order differences (Figures 7A, B), there were differences in specific values between the indices extracted by the Increment borers and Resistograph. Nevertheless, the tree ring width index extracted by the Resistograph is generally reliable.

3.3.2 Tree ring density index

Using EEMD to remove the residual resistance of the Resistograph and based on the specific characteristics of the sample sequence, at least six samples are required when establishing a balanced chronology (Figure 8). The Drilling resistance extracted tree ring mean density (DDmean), maximum density (DDmax), minimum density (DDmin) index, and Feed resistance extracted tree ring mean density (FDmean), maximum density (FDmax), minimum density (FDmin) index for ancient *P.tabulaeformis* at Yuxiang Forest Farm were ultimately obtained.

By comparing the tree ring density indices and their first-order differences extracted by the drilling resistance and feed resistance of the Resistograph, it is found that despite some differences, the tree ring density indices extracted by the Resistograph demonstrate high consistency between drilling and feed resistance methods. The tree ring density index curves extracted by drilling resistance and feed resistance showed higher reliability in both high-frequency and low-frequency signals.

3.4 Relationship between tree ring index and climatic factors

The results of correlation analyses between the IW, DW, FW, DDmean, DDmax, DDmin, FDmean, FDmax, FDmin and the Tmean, Tmax, Tmin, and Pre are shown in Figures 9, 10:

IW and DW both have a high negative correlated with Tmean, Tmax and Tmin from the previous year to the current year; FW is higher negatively correlated with Tmean of C7-8, Tmax of P7-9, P11, C4, C7-8, Tmin of C8; and higher positively correlated with Tmean, Tmin of P3, C3. IW, DW and FW show a weak response to Pre and exhibit significant differences. DW and IW show a high consistency in their response to climate, while there are certain differences in the response of FW to climate compared to IW.

DDmean, DDmax, DDmin, FDmean, FDmax and FDmin exhibit a high consistency in their response to climate, they all show a positive correlation with temperature, they have a weak response to Pre. However, there are certain differences; DDmean, DDmax, and DDmin are more sensitive to climate responses than FDmean, FDmax, and FDmin. Moreover, DDmax and FDmax are

10.3389/fpls.2024.1469675

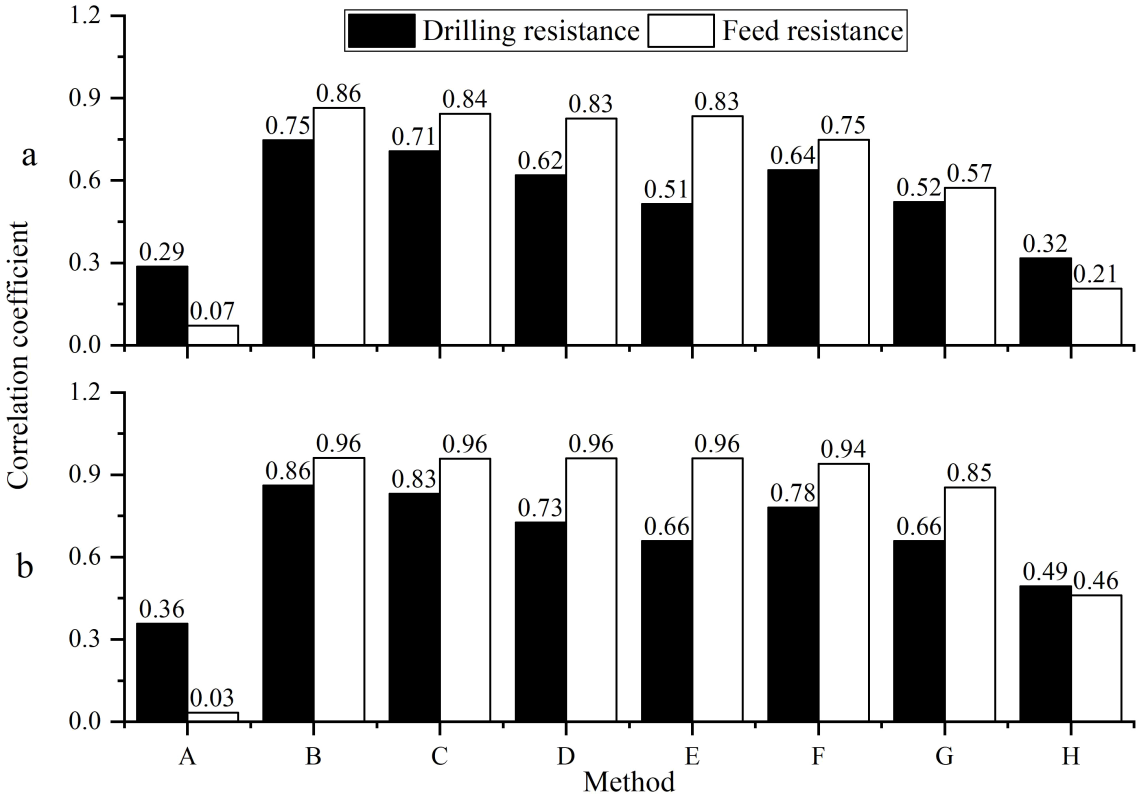


FIGURE 4
The correlation of the comprehensive tree ring width sequence **(A)** was extracted using both the Resistograph and the Increment borers, their 5-year moving averages **(B)**.

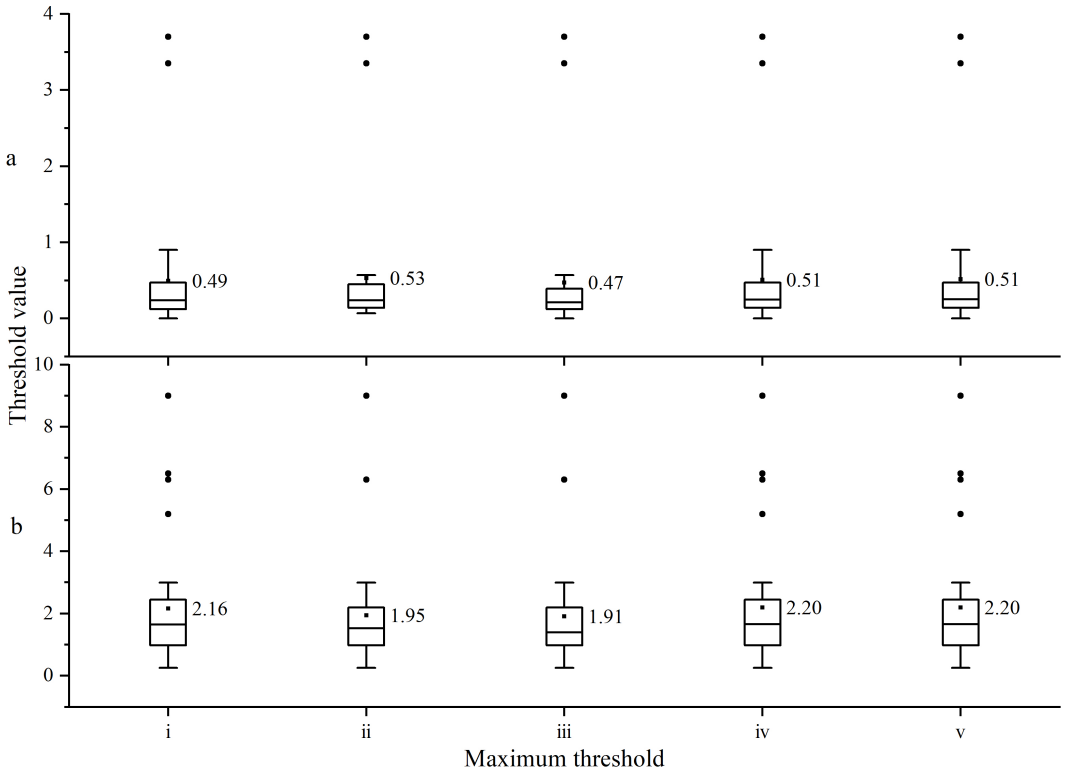


FIGURE 5
Maximum threshold values of **(A)** drilling resistance and **(B)** feed resistance.

10.3389/fpls.2024.1469675

TABLE 2 Number of significant correlations in each comprehensive threshold.

		i	ii	iii	iv	v
III	drilling Resistance	39	40	39	39	39
	Feed Resistance	38	38	38	38	38
IV	drilling Resistance	41	41	41	41	41
	Feed Resistance	41	41	41	41	41

more sensitive to responses to Tmax, while DDmin and FDmin are more sensitive to responses to Tmin.

4 Discussion

4.1 The feasibility of extracting tree-ring information using the Resistograph

In this study, we have innovatively improved upon the traditional automatic tree-ring information extraction methods used by the Resistograph, introducing a comprehensive linear threshold for the extraction of tree-ring information. After validation through correlation analysis, the new method has achieved significant enhancements in extraction accuracy and data reliability, demonstrating higher precision and reliability compared to the original technique.

Through correlation analysis and T-tests between IW, DW, and FW, it was found that DW and FW are significantly correlated with IW (0.41, 0.19, $p< 0.01$), and the differences are not significant (ρ = 0.55, 0.99). Additionally, the IW, DW, and FW indices exhibit a high level of consistency in their response to climate, mainly influenced by the temperatures of the previous and current years. This means that the tree ring index extracted by the Resistograph using the optimal comprehensive threshold is credible and can be used for studying the tree growth history. However, there are certain differences between DW, FW, and IW, which may be due to the different measurement methods used. When measuring tree ring width, the Increment borers typically identifies the edge of latewood as the ring boundary, whereas the Resistograph identifies the location of maximum resistance rather than the late wood boundary. Furthermore, the correlation coefficient between DW and IW was higher than that between FW and IW, indicating that tree ring width index extracted by the Resistograph using drilling resistance was more accurate than that extracted using feed resistance. This may be due to the lower temporal resolution of feed resistance measurement, which cannot provide enough details

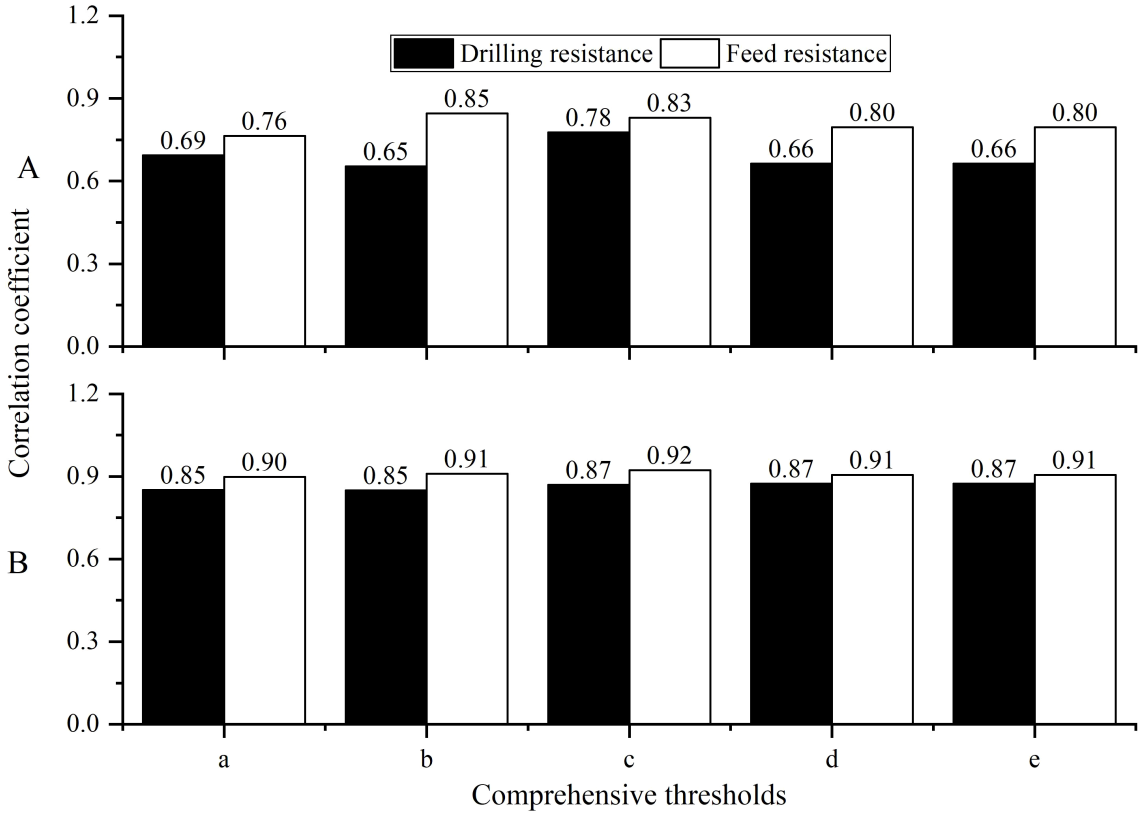


FIGURE 6
The correlation of the comprehensive tree ring width sequence **(A)** was extracted using both the Resistograph and the Increment borers, their 5-year moving averages **(B)**.

10.3389/fpls.2024.1469675

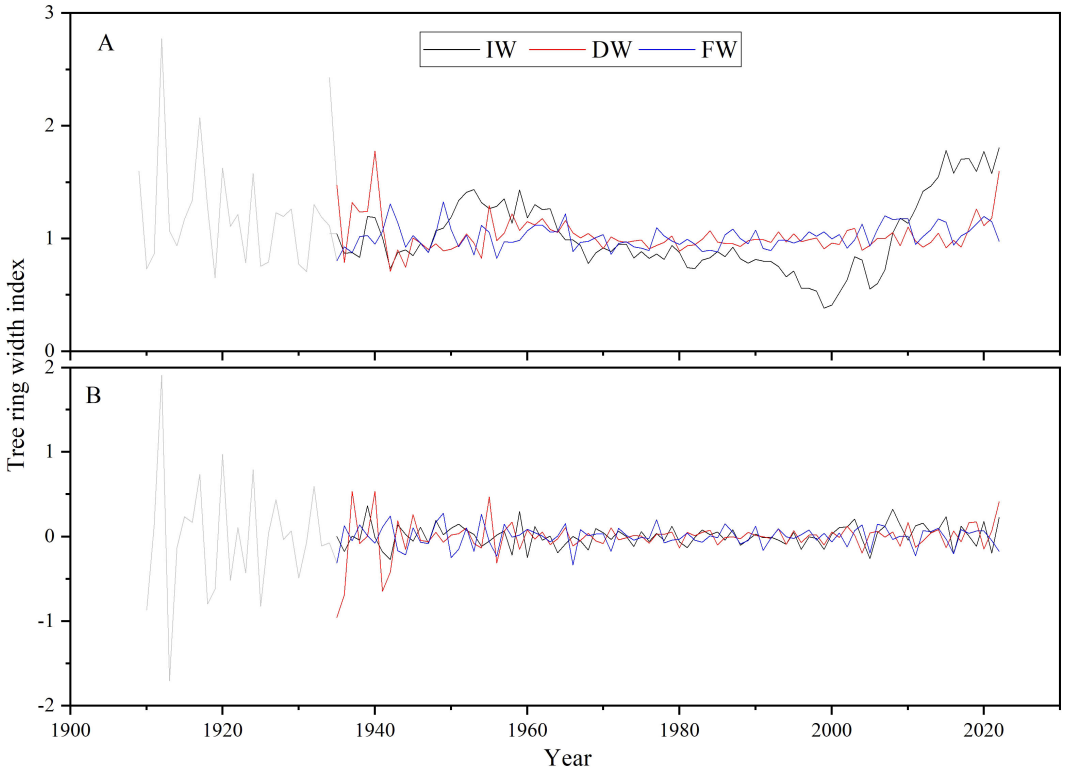


FIGURE 7
Tree ring width indices were developed (**A**: Tree ring width index; **B**: First-order difference of tree ring width index).

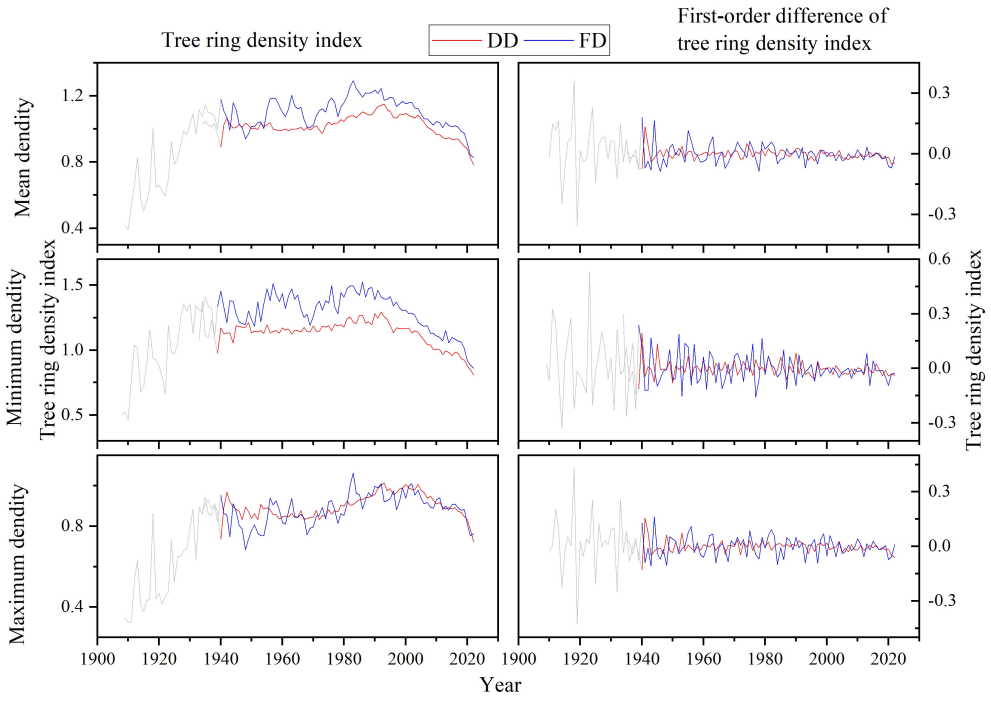


FIGURE 8
Tree ring density indices of ancient *P.tabulaeformis* at Yuxiang Forest Farm (DD: Tree ring density index extracted by drilling resistance; FD: Tree ring density index extracted by feed resistance).

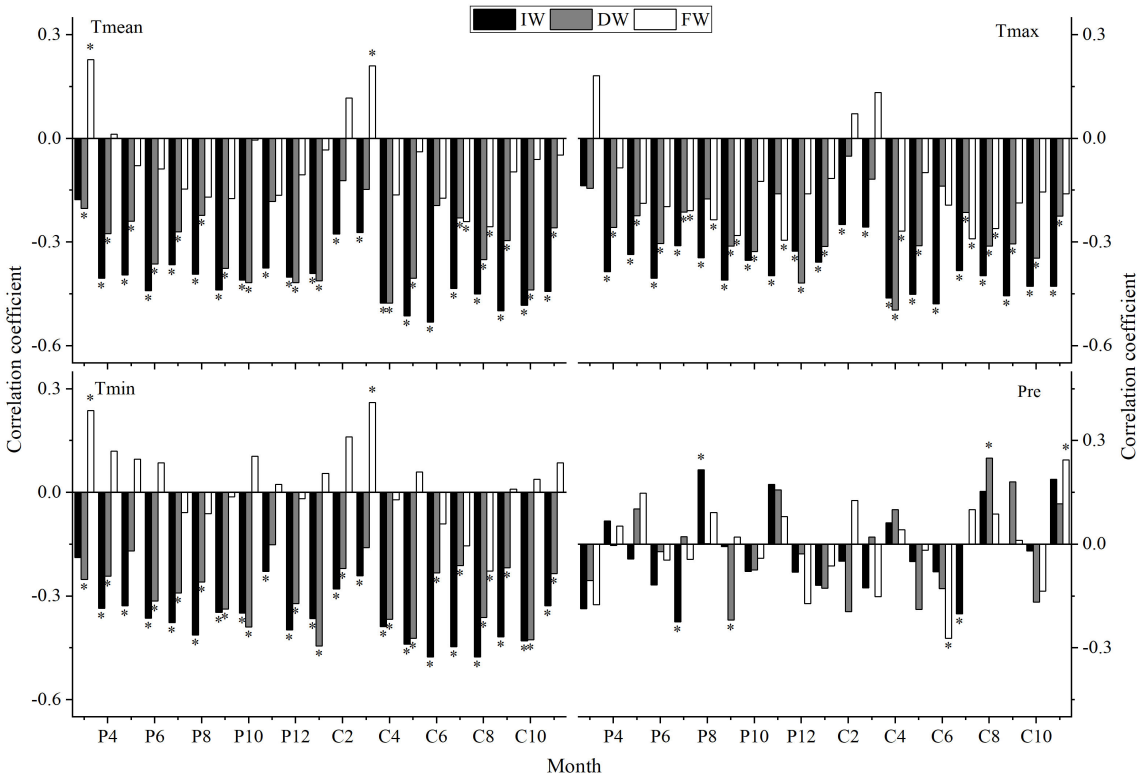


FIGURE 9
Correlation analyses between different width indices and climatic factors (P: represents the previous year, C: represents the current year; * indicates significant correlation at the 95% confidence level, same below).

to accurately distinguish the growth period of each year, and thus may not be as effective in extracting ring information as drilling resistance. Therefore, drilling resistance is usually a more reliable choice for precise analysis of tree growth history and environmental response (Li et al., 2016). Despite these differences, comparative analysis indicates that the tree-ring width data extracted by the Resistograph remains reliable, particularly the tree-ring information extracted using drilling resistance. This finding holds significant implications for future research on the Resistograph.

4.2 Relationship between tree ring and climatic factors

Yuxiang Forest Farm, situated in the low-altitude plain area at an elevation of only 62 meters, is located at the southern edge of ancient *P.tabulaeformis*'s distribution area, where higher temperatures are detrimental to its growth. Higher temperatures may trigger drought stress, which not only affects the photosynthesis and respiration of trees, but may also slows down the rate of cell division. These changes ultimately affect the number of wood cells and the content of cell wall substances, thereby significantly impacting the width and density of tree rings (Xu et al., 2011; Hansen et al., 1997; Zhang et al., 2021). This phenomenon has been widely verified in studies of tree growth (Ruxianguli et al., 2023). The temperature of the previous year has an obvious "lag effect" on the growth of ancient *P.tabulaeformis* (Fritts, 1976; Wu, 1990; Wei, 2018), and excessively high temperatures have an adverse effect on the growth of the current year. During the growing season, high temperatures cause trees to close their stomata and experience "carbon starvation", reducing photosynthesis and increasing the consumption of respiration, slowing cell division, accelerating lignification, thereby inhibiting radial growth (Li et al., 2022). This aligns with findings from other regions in Henan Province (Peng et al., 2014; 2018; 2019; Zhao et al., 2019) and increasing the density of tree rings (Ruxianguli et al., 2023), this effect has also been reflected in the *Picea crassifolia* Kom of Qilian Mountain (Xu et al., 2011) and *Picea schrenkiana Fischet* Mey in Kongnaes region (Zhang et al., 2011). By the end of the growing season, trees have largely completed cell growth and enter a phase of photosynthetic accumulation. At this point, high temperatures may lead to a reduction in the water supply available to trees or a decrease in water utilization efficiency, which in turn triggers the closure of leaf stomata, reducing cell growth and the demand for photosynthesis (Yang et al., 2021) and thus limit tree growth (Rahman et al., 2016; Zhang et al., 2020). As a result, trees accumulate more photosynthetic products, mainly used to strengthen the thickness of the cell walls of the latewood cells of the tree rings (Petit and Crivellaro, 2014). At this stage, the latewood cells of the trees have a smaller diameter and thicker cell walls (Li et al., 2015), mainly playing a supporting role (Hervé et al., 2004; Sperry et al., 2006), contributing less to the transport of nutrients

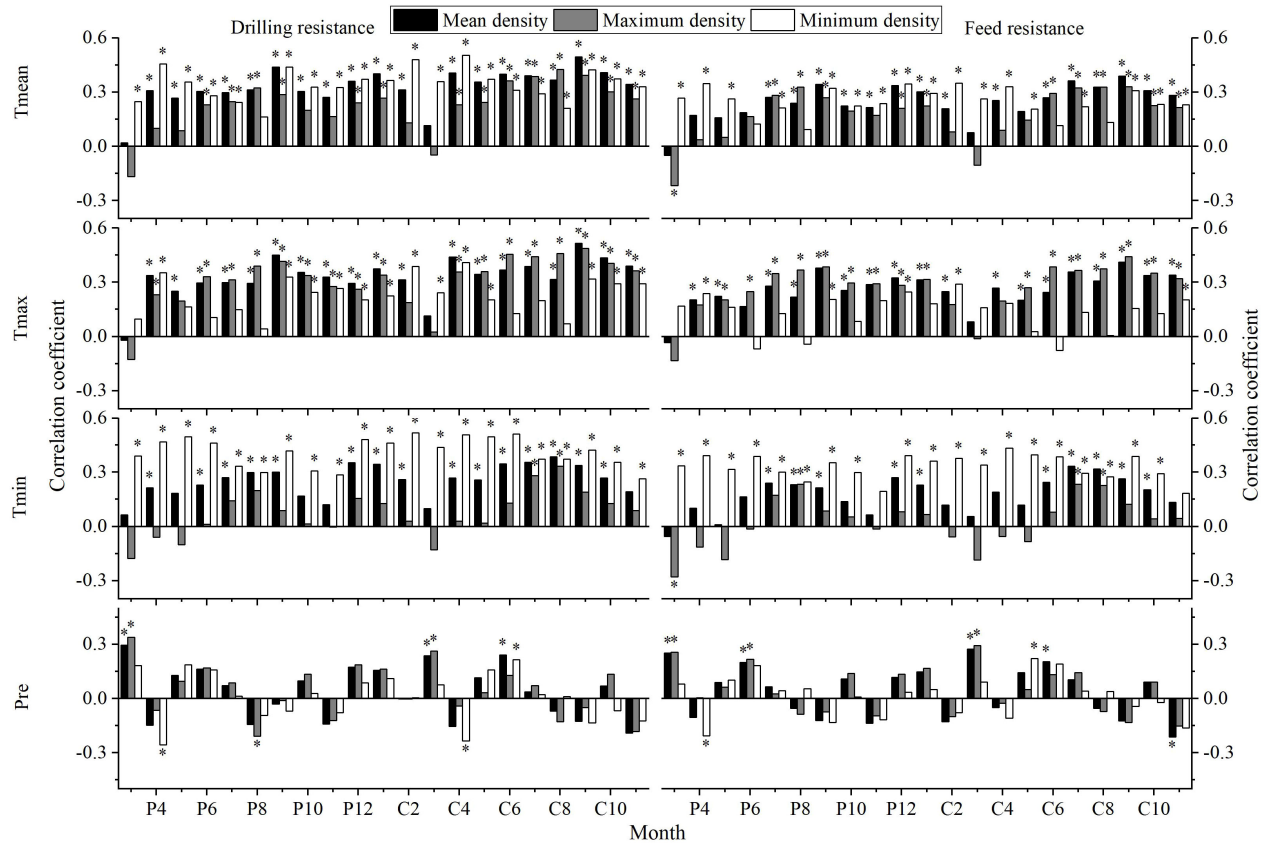


FIGURE 10
Correlation analyses between different tree ring density indices and climatic factors (P: represents the previous year, C: represents the current year; * indicates significant correlation at the 95% confidence level).

and water by the tree, with a very low degree of cell hollowing, thus making the density of latewood relatively large. This phenomenon has also been recorded in the *Pinus koraiensis* and *Abies fabri (Mast.) Craib* studies of Changbai Mountain (Ruxianguli et al., 2023).

However, there are certain differences in climatic response of different density indices; the maximum density is more significantly affected by Tmax, while it shows a significant correlation with Tmin only within two to three months prior. This is because Tmax usually occurs during the day, which is the period when the tree's metabolic activities are most active. When the temperature is too high, it will intensify the transpiration of the tree, leading to drought stress, inhibiting tree growth, causing the annual ring density to begin the growth phase earlier, thus affecting the maximum density of the annual ring (Chen et al., 2010). On the other hand, the annual minimum ring density is more influenced by Tmin, which usually occurs at night and plays a key role in the accumulation of cell wall substances and the lignification process, thereby affecting the value of the annual ring density. Therefore, the intensity of the tree's photosynthesis and respiration is crucial for the variation in annual ring density (Antonova and Stasova, 1993; Gou et al., 2007; Yoshihiro et al., 2002; Ruxianguli et al., 2023). Nighttime warming will promote leaf respiration and root respiration, stimulating the consumption of carbohydrates and other nutrients within the cells, leading to a reduction in the accumulation of nutrients (Zhu and Zheng, 2022). At this time, the earlywood cells expand more slowly, the cell walls are thicker, the cells have a higher saturation level, and it is easier to form a higher annual ring density.

5 Conclusions

This study focuses on *P. tabulaeformis* the Yuxiang Forest Farm in Henan Province, utilizing non-destructive tree ring data collection methods such as the Resistograph and Increment borers to explore the feasibility of extracting tree-ring information using the Resistograph and to discuss the impact of meteorological factors on tree-ring characteristics. The main conclusions drawn are:

(1) The Resistograph's use of a linear threshold for extracting tree-ring information is more accurate and credible compared to the use of a fixed threshold.

(2) In the Yuxiang Forest Farm, the initial comprehensive threshold for extracting tree-ring information from ancient *P. tabulaeformis* using drilling and feed resistance is 0, with the maximum thresholds being 0.47 and 1.91, respectively. Linear threshold is identified as the optimal comprehensive threshold, which provides the best filtering effect. The tree-ring information extracted is reliable. Moreover, the drilling

resistance is more accurate than the feed resistance in extracting tree ring information.

(3) The variations in both width and density of the ancient *P. tabulaeformis* at Yuxiang Forest Farm are primarily influenced by temperature. The maximum density is more significantly affected by the average maximum temperature, whereas the minimum density is more noticeably influenced by the average minimum temperature.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding authors.

Author contributions

JKL: Conceptualization, Data curation, Investigation, Methodology, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. YW: Supervision, Writing – review & editing. YL: Supervision, Visualization, Writing – review & editing. JXL: Validation, Visualization, Writing – review & editing. KZ: Supervision, Validation, Visualization, Writing – review & editing. XW: Supervision, Visualization, Writing – review & editing. JP: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. Project funding: This study was supported by the National Natural Science Foundation of China (No. 42077417; 41671042).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Acuna, L., Basterra, L. A., Casado, M. M., Lopez, G., Ramon-Cueto, G., Relea, E., et al. (2011). Application of resistograph to obtain the density and to differentiate wood species. *Mater. Constucc.* 61, 451–464. doi: 10.3989/mc.2010.57610

Antonova, G. ,. F., and Stasova, V. V. (1993). Effects of environmental factors on wood formation in Scots pine stems. *Trees* 7, 214–219. doi: 10.1007/BF00202076

Bauer, C., Kilbertus, G., and Bucur, V. (1991). Technique Ultrasonore de Caractérisation du Degré d'Altération des Bois de Hêtre et de Pin Soumis à l'Attaque de Différents Champignons. *Holzforschung* 45, 41–46. doi: 10.1515/hfsg.1991.45.1.41

Biondi, F., and Waikul, K. (2004). Dendroclim2002: A C++ program for statistical calibration of climate signals in tree-ring chronologies. *Comput. Geosci.* 30, 303–311. doi: 10.1016/j.cageo.2003.11.004

Bulleit, W. M., and Falk, R. H. (1985). Modeling stress wave passage times in wood utility poles. *Wood Sci. Technol.* 19, 183–191. doi: 10.1007/BF00353080

Ceraldi, C., Mormone, V., and Russo Ermolli, E. (2001). Resistographic inspection of ancient timber structures for the evaluation of mechanical characteristics. *Mat. Struct.* 34, 59–64. doi: 10.1007/BF02482201

Chantre, G., and Rozenberg, P. (1997). *Can drill resistance profiles (Resistograph) lead to within-profile and within-ring density parameters in Douglas fir wood*? (Quebec City, Canada: Forintek Canada Corp), 41–47.

Chen, F., Yuan, Y. J., Wei, W. S., Yu, S. L., Shang, H. M., Zhang, R. B., et al. (2010). Dendroclimatic reconstruction of mean maximum may-august temperature from the maximum density of the Larix sibirica in Hoboksar, China. *Acta Ecol. Sin.* 30, 4652–4658. doi: 10.0000/j.1000-0933.2010301246524658

Costello, L. R., and Quarles, S. L. (1999). Detection of wood decay in blue gum and elm: an evaluation of the Resistograph and the portable drill. *J. Arboric.* 25, 311–318. doi: 10.48044/jauf.1999.041

Fritts, H. C. (1976). *Tree Rings and Climate.* (Academic Press).

Gao, X. (2023). Research progress on the identification methods of the age of ancient trees. *J. Green Sci. Technol.* 25, 163–168 + 172. doi: 10.16663/j.cnki.lskj.2023.03.054

Gauthier, S., Bernier, P., Kuuluvainen, T., Shvidenko, A. Z., and Schepaschenko, D. G. (2015). Boreal forest health and global change. *Science* 349, 819–822. doi: 10.1126/science.aaa9092

Gou, X. H., Chen, F. H., Yang, M. X., Jacoby, G., Fang, K. Y., Tian, Q. H., et al. (2007). Asymmetric changes in maximum and minimum temperatures revealed by tree rings from the northeastern tibetan plateau. *Sci. China Earth Sci.* 37, 1480–1492. doi: 10.3969/j.issn.1674-7240.2007.11.007

Hansen, J. R., Türk,, Vogg, G., Heim, R., and Beck, E. (1997). *Conifer carbohydrate physiology: Updating classical views* (Leiden, Netherlands: Backhuys Publisher), 97–108.

Hervé, C., Fabienne, F., Stefan, M., and Catherine, C. (2004). Xylem wall collapse in water-stressed pine needles. *Plant Physiol.* 134, 401–408. doi: 10.1104/pp.103.028357

Holmes, R. L. (1983). Computer-assisted quality control in tree-ring dating and measurement. *Tree-Ring Bull.* 43, 69–78. doi: 10.1006/biol.1999.0214

IPCC (2021). *Climate change 2021: the physical science basis* (Cambridge: Cambridge University Press).

Isik, F., and Li, B. (2003). Rapid assessment of wood density of live trees using the Resistograph for selection in tree improvement programs. *Can. J. For. Res.* 33, 2426–2435. doi: 10.1139/x03-176

Ji, S. H. (2022). Climate change in henan province from 1961 to 2020 and its impact on climate productivity. *J. Shaanxi. Meteor.* 06), 42–47.

Li, M. Y., Wang, L. L., Fan, Z. X., and Shen, C. C. (2015). Tree-ring density inferred late summer temperature variability over the past three centuries in the Gaoligong Mountains, southeastern Tibetan Plateau. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 422, 57–64. doi: 10.1016/j.palaeo.2015.01.003

Li, Q., and Jing, D. (1999). Division of hydrological regions in henan province. *Yellow. River.* 8, 33–34.

Li, X., Dai, J., Qian, W., and Chang, L. H. (2016). Effect rule of different drill speeds on the wooden micro-drill resistance. *J. Beijing Univ. Technol.* 42, 1066–1070. doi: 10.11936/bjutxb2015060080

Li, X., Peng, J. F., Li, J. R., Yang, L., Cui, J. Y., Peng, M., et al. (2022). Climate-growth response of *Pinus tsbulaeformis* in the south slope of Longchiman, Mt. Funiu, central China. *Acta Ecol. Sin.* 42, 2865–2877. doi: 10.5846/stxb202101070076

Liang, X. E. (2017). *Response of the tree ring to the environmental change based on the Resistograph data* (Nanjing Forestry University).

Lima, J. T., Sartório, R., Trugilho, P. F., Cruz, C. R., and Vieira, R. D. S. (2007). Use of the resistograph for Eucalyptus wood basic density and perforation resistance estimative. *Sci. For.* 35, 85–93. doi: 10.1590/S0100-68062007000300007

Millar, I. C., and Stephenson, L. N. (2015). Temperate forest health in an era of emerging megadisturbance. *Science* 349, 823–826. doi: 10.1126/science.aaa9933

Nutto, L., and Biechele, T. (2015). "Drilling resistance measurement and the effect of shaft friction – using feed force information for improving decay identification on hard tropical wood," in *19th International Nondestructive Testing and Evaluation of Wood Symposium*. (Geneva University Press), 156–160.

Pan, H. (2020). Studies on age estimation algorithms of living trees based on the micro-damage measurement of resistograph. *Chinese. Chin. Acad. Forestry Sci.* doi: 10.13275/j.cnki.lykxyj.2021.01.003

Parker, M. L., and Meleskie, K. R. (1970). Preparation of X-ray negatives of tree-ring specimens for dendrochronological analysis. *Tree-Ring Bull.* 30, 11–22.

Peng, S. Z. (2020a). *-km monthly mean temperature dataset for China, (1901-2022)* (National Tibetan Plateau/Third Pole Environment Data Center).

Peng, S. Z. (2020b). *-km monthly mean minimum temperature dataset for China, (1901-2022)* (National Tibetan Plateau/Third Pole Environment Data Center).

Peng, S. Z. (2020c). *-km monthly mean maximum temperature dataset for China, (1901-2022)* (National Tibetan Plateau/Third Pole Environment Data Center).

Peng, S. Z. (2020d). *-km monthly precipitation dataset for China, (1901-2022)* (National Tibetan Plateau/Third Pole Environment Data Center).

Peng, J. F., Li, J. B., Wang, T., Huo, J. X., and Yang, L. (2019). Effect of altitude on climate–growth relationships of Chinese white pine (*Pinus armandii*) in the northern Funiu Mountain, central China. *Clim. Change.* 154, 273–288. doi: 10.1007/s10584-019-02416-7

Peng, J. F., Liu, Y. Z., and Wang, T. (2014). A tree-ring record of 1920's—1940's droughts and mechanism analyses in Henan Province. *Acta Ecol. Sin.* 34, 3509–3518. doi: 10.1016/j.chnaes.2014.05.012

Peng, C. H., Ma, Z. H., Lei, X. D., Zhu, Q. A., Chen, H., Wang, W. F., et al. (2011). A drought-induced pervasive increase in tree mortality across Canada's boreal forests. *Nat. Clim. Change* 1, 467–471. doi: 10.1038/nclimate1293

Peng, J. F., Peng, K. Y., and Li, J. B. (2018). Climate-growth response of Chinese white pine (*Pinus armandii*) at different age groups in the Baiyunshan National Nature Reserve, central China. *Dendrochronologia* 49, 102–109. doi: 10.1016/j.dendro.2018.01.001

Petit, G., and Crivellaro, A. (2014). Comparative axial widening of phloem and xylem conduits in small woody plants. *Trees* 28, 915–921. doi: 10.1007/s00468-014-1006-1

Polge, H. (1966). [amp]]Eacute;tablissement des courbes de variation de la densité du bois par exploration densitométrique de radiographies d'échantillons prélevés à la tarière sur des arbres vivants: applications dans les domaines Technologique et Physiologique. *Ann. For. Sci.* 23, 1–206. doi: 10.1051/FOREST/19660101

Rahman, M. H., Begum, S., Nakaba, S., Yamagishi, Y., Kudo, K., Nabeshima, E., et al. (2016). Relationship between the earlywood-to-latewood transition and changes in levels of stored starch around the cambium in locally heated stems of the evergreen conifer *Chamaecyparis pisifera*. *Trees* 30, 1619–1631. doi: 10.1007/s00468-016-1395-4

Ren, W. Q., Hou, H. P., Bi, H. T., and Sun, Y. M. (2024). Characteristics of old tree resources in different regions of Henan Province and protection strategies. *J. Henan. Agric. Univ.* 58 (3), 444–455. doi: 10.16445/j.cnki.1000-2340.20240023.002

Rinn, F. (1994). "Catalogue of relative density profiles of trees, poles and timber derived from Resistograph microdrillings," in *9th International Symposium on Non-destructive Testing*. (University of Wisconsin-Madison), 61.

Rinn, F., Schweingruber, F. H., and Schär, E. (1996). RESISTOGRAPH and X-ray density charts of wood. Comparative evaluation of drill resistance profiles and X-ray density charts of different wood species. *Holzforschung* 50, 303–311. doi: 10.1515/hfsg.1996.50.4.303

Ruxianguli, A. B. D. R. H. M., Zhang, T. W., Yu, S. L., Yuan, Y. J., Zhang, R. B., Wang, Z. P., et al. (2023). Variation in characteristics and climate response of tree ring wood density between *Pinus koraiensis* and *Abies nephrolepis* in Changbai Mountain. *J. Earth Environ.* 14, 573–587. doi: 10.7515/JEE222036

Schad, K. C., Schmoldt, D. L., and Ross, R. J. (1996). Nondestructive methods for detecting defects in softwood logs. *Res. Paper. Fpl.* 546, 13. doi: 10.1007/978-3-642-33191-6_21

Seibold, S., Hagge, J., Müller, J., Gruppe, A., Brandl, R., Bässler, C., et al. (2018). Experiments with dead wood reveal the importance of dead branches in the canopy for saproxylic beetle conservation. *For. Ecol. Manage.* 409, 564–570. doi: 10.1016/j.foreco.2017.11.052

Sperry, J. S., Hacke, G. U., and Jarmila, P. (2006). Size and function in conifer tracheids and angiosperm vessels. *Am. J. Bot.* 93, 1490–1500. doi: 10.3732/ajb.93.10.1490

Stokes ma Smiley-TL (1968). *Tree-ring Dating* (Chicago, IL, USA: The University of Chicago Press).

Sun, Y. L. (2012). *Determining Wood Density and Mechanical Properties of Ancient Architectural Timbers with Micro-Drilling Resistance* (BeiJing Forestry University).

Sun, Y. L., Zhang, H. J., Zhu, L., and Yan, H. C. (2011). Application research of micro-drill resistance meter in detecting wood density. *Hunan. Agric. Sci.* 10), 43–44. doi: 10.16498/j.cnki.hnnykx.2011.10.002

Takács, M., and Malatinszky, Á. (2021). Half of the ancient trees in Hungary stand in human-altered environments. *Sustainability* 13, 12803. doi: 10.3390/su132212803

Ukrainetz, U. K. N., and O'Neill, O. A. G. (2010). An analysis of sensitivities contributing measurement error to Resistograph values. *Can. J. For. Res.* 40, 806–811. doi: 10.1139/X10-019

Wang, Y. X., Gao, L. S., and Zhao, X. H. (2013). Tree-ring width chronology of populus tremula and its relationship with the weather in changbai mountain of northeastern China. *J. Northeast For. Univ.* 41, 10–13. doi: 10.13759/j.cnki.dlxb.2013.01.025

Wang, S. Y., and Lin, C. J. (2001). Application of the drill resistance method for density boundary evaluation of early wood and late wood of Taiwania (*Taiwania cryptomerioides* Hay.) plantation. *Taiw. J. For. Sci.* 16, 197–200. doi: 10.1007/s100860300018

Wang, W. K., Mao, J. Z., and Chen, D. G. (1990). Henan geographical records. *Henan. People's. Publ. House.* 8, 1–8.

Wei, H. C. (2018). *Response of Tree Rings to Different Planting Densities Based on the Data of Resistograph: A Case Study of Poplar Plantation in Dongtai Forest Farm of Jiangsu province* (Nanjing Forestry University).

Winistorfer, P., Xu, W., and Wimmer, R. (1995). Application of a drill resistance technique for density profile measurement in wood composite panels. *For. Prod. J.* 45, 90–93. doi: 10.1007/BF00816385

Wu, X. D. (1990). *Tree Rings and Climate Change* (China Meteorological Press).

Wu, F. S. (2011a). *Preliminary Study on Non-destructive Method and Safety Assessment of Tree* (Anhui Agricultural University). doi: 10.7666/d.d185609

Wu, J. (2019). Developing general equations for urban tree biomass estimation with high-resolution satellite imagery. *Sustainability* 11, 4347. doi: 10.3390/su11164347

Wu, F. S., Wu, Y. J., and Shao, Z. P. (2011b). Determination of cedar sample interior characteristics based on stress wave and resistograph. *J. Anhui. Agric. Univ.* 38, 127–130.

Xia, Q. Z. (2023). *Response of tree Growth to Climate Change Based on Tree Ring Characteristic from Nondestructive Measurement* (Nanjing Forestry University).

Xu, H. M., Huang, H. P., Zhang, J. L., Yuan, C. W., and Liu, X. (2012). Research summary of ancient and famous woody plants. *Hubei. For. Sci. Technol.* 174, 34–37 + 57. doi: 10.3969/j.issn.1004-3020.2012.02.011

Xu, J. M., Lv, J. X., Bao, F. C., Huang, R. F., Liu, X. D., Robert, E., et al. (2011). Response of wood density of *Picea crassifolia* to climate change in Qilian Mountains of northwestern China. *J. Beijing For. Univ.* 33, 115–121. doi: 10.1007/s11676-011-0141-4

Yang, L., Qin, C., and Li, G. (2021). Climatic signals recorded by Qinghai spruce tree-ring density in the western part of Qilian Mountains, China. *Chin. J. Appl. Ecol.* 32, 3636–3642. doi: 10.13287/j.1001-9332.202110.026

Yang, H. M., and Wang, L. H. (2010). Research and development on 2D lmaging technology of the dicay in trees and logs. *Sci. Silvae Sin.* 46, 170–175. doi: 10.3788/HPLPB20102208.1751

Yao, J. F., Fu, L. Y., Song, X. Y., Wang, X. F., Zhao, Y. D., Zheng, Y. L., et al. (2022c). Feasibility study on measuring density of earlywood and latewood by micro drill resistance method. *J. For. Eng.* 7, 66–73. doi: 10.13360/j.issn.2096-1359.202203028

Yao, J. F., Lu, J., and Ding, X. L. (2022a). Investigating the resistance expression method of wood resistance drill instruments. *For. Prod. J.* 73, 231–238. doi: 10.13073/FPJ-D-23-00005

Yao, J. F., Lu, J., and Fu, L. Y. (2022b). Micro drill resistance instrument measurements at different feed speeds: novel conversion algorithm for enhanced accuracy. *J. Nondestruct. Eval.* 42, 56. doi: 10.1007/s10921-023-00968-4

Yao, J. F., Wu, Z. Y., Zheng, Y. L., Rao, B. Q., Li, Z. F., Hu, Y. C., et al. (2023). Design of a tree micro drill instrument to improve the accuracy of wood density estimation. *Forests* 14, 2071. doi: 10.3390/f14102071

Yao, J. F., Zhao, Y. D., Fu, L. Y., Song, X. Y., Lu, J., and Li, S. S. (2022d). Tree-rings measurement method based on micro drill resistance. *Trans. Chin. Soc Agric. Mach.* 53, 52–59. doi: 10.6041/i.issn.1000-1298.2022.04.005

Yoshihiro, H., Masato, Y., Takanori, I., and Takashi, O. (2002). Diurnal difference in the amount of immunogold-labeled glucomannans detected with field emission scanning electron microscopy at the innermost surface of developing secondary walls of differentiating conifer tracheids. *Planta* 215, 1006–1012. doi: 10.1007/s00425-002-0824-3

Zhang, K. Y. (2024). Methods and research on tree age determination for ancient and famous trees. *J. Northeast For. Univ.*

Zhang, F., Ding, J. F., Yang, M., and Peng, C. (2023). Snowfall identification and response of snowfall to temperature and precipitation under climate change in Henan province. *Adv. Clim. Change Res.* 19, 714–722. doi: 10.12006/j.issn.1673-1719.2023.105

Zhang, X. F., Li, H., Liu, X. Y., Zhang, S. B., and Huang, R. F. (2007). Application of resistograph testing technique for wood. *China Wood Ind.* 21, 41–43. doi: 10.3969/j.issn.1001-8654.2007.02.014

Zhang, H., Yan, Y., Hu, Y. N., Huang, Z. J., Wu, P. F., and Ma, X. Q. (2020). Analysis on the relationship between radial growth of Cunninghamia lanceolataand climatic factors based on tree-ring climate correlation. *J. Fujian. Agric. For. Univ. (Nat. Sci. Ed.)* 49, 59–66. doi: 10.13323/j.cnki.j.fafu(nat.sci.).2020.01.011

Zhang, T. W., Yuan, Y. J., Yu, S. L., Wei, W. S., Shang, H. M., Zhang, R. B., et al. (2011). Contrastive analysis and climatic response of tree-ring gray values and tree-ring densities. *Acta Ecol. Sin.* 31, 6743–6752. doi: 10.1097/RLU.0b013e3181f49ac7

Zhang, H., Zhang, Y., Hu, Y. N., Yan, Y., Wu, P. F., Zeng, A. C., et al. (2021). Response of tree ring density to climatic factors of Cunninghamia lanceolata under climate warming. *Acta Ecol. Sin.* 41, 1551–1563. doi: 10.5846/stxb201908261764

Zhao, Y. S., Shi, J. F., Shi, S. Y., Ma, X. Q., Zhang, W. J., Wang, B. W., et al. (2019). Early summer hydroclimatic signals are captured well by tree-ring earlywood width in the eastern Qinling Mountains, central China. *Clim. Past.* 15, 1113–1131. doi: 10.5194/cp-15-1113-2019

Zhu, L., Zhang, H. J., Sun, Y. L., Wang, X. P., and Yan, H. C. (2013). Mechanical properties non-destructive testing of wooden components of Korean pine based on stress wave and micro-drilling resistance. *J. Nanjing. For. Univ. Nat. Sci. Ed.* 37, 156–158. doi: 10.3969/j.issn.1000-2006.2013.02.028

Zhu, J. T., and Zheng, J. H. (2022). Effects of diurnal asymmetric warming on terrestrial ecosystems. *Chin. J. Ecol.* 41, 777–783. doi: 10.13292/j.1000-4890.202203.001

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 22 November 2024
DOI 10.3389/fpls.2024.1470520

OPEN ACCESS

EDITED BY
Daniel Cozzolino,
University of Queensland, Australia

REVIEWED BY
Daisuke Miki,
Chinese Academy of Sciences (CAS), China
Ahmed M. Saad,
Zagazig University, Egypt

*CORRESPONDENCE
Pilar Hernandez
phernandez@ias.csic.es

RECEIVED 25 July 2024
ACCEPTED 15 October 2024
PUBLISHED 22 November 2024

CITATION
Mérida-García R, Gálvez S, Solís I, Martínez-Moreno F, Camino C, Soriano JM, Sansaloni C, Ammar K, Bentley AR, Gonzalez-Dugo V, Zarco-Tejada PJ and Hernandez P (2024) High-throughput phenotyping using hyperspectral indicators supports the genetic dissection of yield in durum wheat grown under heat and drought stress.
Front. Plant Sci. 15:1470520.
doi: 10.3389/fpls.2024.1470520

High-throughput phenotyping using hyperspectral indicators supports the genetic dissection of yield in durum wheat grown under heat and drought stress

Rosa Mérida-García[1], Sergio Gálvez[2], Ignacio Solís[3], Fernando Martínez-Moreno[3], Carlos Camino[4], Jose Miguel Soriano[5], Carolina Sansaloni[6], Karim Ammar[6], Alison R. Bentley[7], Victoria Gonzalez-Dugo[1], Pablo J. Zarco-Tejada[1,8] and Pilar Hernandez[1]*

[1]Institute for Sustainable Agriculture (IAS), Consejo Superior de Investigaciones Científicas (CSIC), Córdoba, Spain, [2]Department of Languages and Computer Science, ETSI Informática, Universidad de Málaga, Andalucía Tech, Málaga, Spain, [3]Department of Agronomy, ETSIA (University of Seville), Seville, Spain, [4]European Commission (EC), Joint Research Centre (JRC), Ispra, Italy, [5]Department of Agricultural and Forest Sciences and Engineering, University of Lleida - AGROTECNIO, Lleida, Spain, [6]International Maize and Wheat Improvement Center (CIMMYT), Texcoco, México, Mexico, [7]Research School of Biology, Australian National University, Canberra, ACT, Australia, [8]School of Agriculture, Food and Ecosystem Sciences (SAFES), Faculty of Science (FoS), and Faculty of Engineering, and Information Technology (IE-FEIT), University of Melbourne, Melbourne, VIC, Australia

High-throughput phenotyping (HTP) provides new opportunities for efficiently dissecting the genetic basis of drought-adaptive traits, which is essential in current wheat breeding programs. The combined use of HTP and genome-wide association (GWAS) approaches has been useful in the assessment of complex traits such as yield, under field stress conditions including heat and drought. The aim of this study was to identify molecular markers associated with yield (YLD) in elite durum wheat that could be explained using hyperspectral indices (HSIs) under drought field conditions in Mediterranean environments in Southern Spain. The HSIs were obtained from hyperspectral imagery collected during the pre-anthesis and anthesis crop stages using an airborne platform. A panel of 536 durum wheat lines were genotyped by sequencing (GBS, DArTseq) to determine population structure, revealing a lack of genetic structure in the breeding germplasm. The material was phenotyped for YLD and 19 HSIs for six growing seasons under drought field conditions at two locations in Andalusia, in southern Spain. GWAS analysis identified 740 significant marker-trait associations (MTAs) across all the durum wheat chromosomes, several of which were common for YLD and the HSIs, and can potentially be integrated into breeding programs. Candidate gene (CG) analysis uncovered genes related to important plant processes such as photosynthesis, regulatory biological processes, and

plant abiotic stress tolerance. These results are novel in that they combine high-resolution hyperspectral imaging at the field scale with GWAS analysis in wheat. They also support the use of HSIs as useful tools for identifying chromosomal regions related to the heat and drought stress response in wheat, and pave the way for the integration of field HTP in wheat breeding programs.

KEYWORDS

durum wheat, heat, drought, stress, HTP, yield, hyperspectral, GWAS

Introduction

Wheat is one of the foremost crops around the world, providing around 20% of the global human intake of calories and 20% of protein (FAOSTAT, 2023). It is the most important cereal in Mediterranean agriculture thanks to its adaptation to semi-arid environments, where it is mainly cultivated under rainfed conditions (Arriagada et al., 2020). Moreover, wheat is not only a highly significant crop for its pivotal role in primary production, but also because of the associated food industry chains (Arriagada et al., 2020). These are some of the reasons why there is a rising demand for increased wheat production, linked to the predictions of increasing global wheat requirements (Leegood et al., 2010) and the current geopolitical context (Bentley et al., 2022). However, given the limited availability of land for agricultural use, this increased demand tends to rely mainly on breeding programs focused on breeding crops with higher yield potential and stability under changing environmental conditions (Rufo et al., 2021c). The main constraint on wheat yield mainly originates from stress conditions such as water deficit and high temperature conditions during the grain filling stages, both of which are common in Mediterranean environments (Araus et al., 2002; Barakat et al., 2016). These environments have therefore been identified as a major sensitive region for yield reductions as a result of climate change (Rufo et al., 2021c). Climate change models (IPPC report, 2023) predict a decrease of about 20% in annual precipitations and an increase of approximately 4°C in temperature during the 21st century. Depending on their time and intensity, drought and heat stresses, along with other environmental pressures, can reduce wheat yields from 10% to 90% (Reynolds et al., 2004). For this reason, wheat breeding programs are becoming more focused on the adaptability and stability of productivity in dry areas (Bhatta et al., 2018). The genetic dissection of the complex mechanisms behind the heat and drought response in wheat relies on the availability of suitable phenotyping methods.

Phenotyping using traditional manual methods is currently considered as a bottleneck which prevents faster selection for increased yield and related traits in breeding programs (Araus and Cairns, 2014). This limits our ability to dissect the genetics of critical traits determining yield (Blum, 2011; Cabrera-Bosquet et al., 2012). For this reason, plant breeders need to improve the capacity to phenotype large number of lines rapidly in order to identify superior genotypes accurately (Araus and Cairns, 2014). Breeding populations can include thousands of lines, and accurately assessing and characterizing them simultaneously is a daunting task (McMullen et al., 2009). This is where high-throughput phenotyping (HTP) approaches offer powerful tools to assess phenotypes in large-scale field experiments, using a range of sensors and efficient image-processing systems (Jin et al., 2021; Hussain et al., 2022). HTP integrates equipment for data acquisition, a control terminal and a platform for data analysis, and possesses advantages such as facilitating the non-destructive, high-throughput detection of seen and unseen traits (Berger et al., 2010; Xiao et al., 2022). As a consequence, many plant breeding programs are exploring the use of HTP (Morisse et al., 2022), for example, through the use of vegetation spectral indices, which represent a breeding tool which could improve genetic gains for several plant traits (Babar et al., 2006) or serve as tools for extracting spectral characteristics related to drought-adaptive processes (Condorelli et al., 2018).

Spectral reflectance indices (SRIs) are calculated using reflectance data captured by hyperspectral or multispectral cameras, encompassing the visible (380-740nm) and the invisible near- and short wave-infrared (740-2500nm) regions, and, depending on the spectral domain, these SRIs provide information related to a plant's photosynthesis and water status (Barakat et al., 2016). Different studies have demonstrated the efficient use of vegetation spectral indices to measure several physiological traits related to crop canopies, such as total dry matter, leaf area index or photosynthetic capacity (Babar et al., 2006; Hassan et al., 2019; Wei et al., 2019), to detect and assess crops under different stress conditions (Aparicio et al., 2002; Condorelli et al., 2018; Camino et al., 2021), or the use of vegetation indices as predictors of crop yield (Sultana et al., 2014; Hassan et al., 2019; Vannoppen et al., 2020) or abiotic stresses (Lowe et al., 2017; Liu et al., 2019) in breeding programs. There is increasing interest in the potential applications of HTP for the genetic dissection of complex traits including yield or drought stress tolerance, through analyses such as QTL mapping or genome-wide association analysis (GWAS). GWAS is a powerful, high-efficiency and high-resolution tool that provides significative associations between

molecular markers and traits of interest using empirical models (Xiao et al., 2022).

The combined used of GWAS and HTP at different levels (proximal or remote, in greenhouses or in the field) has great potential for improving our understanding of plant growth and crop breeding (Xiao et al., 2022). Several studies have reported the use of proximal SRIs obtained using handheld devices for GWAS analysis, concluding that they are useful tools to understand the genetic basis of agronomic physiological or quality traits in wheat under yield potential and heat stress conditions both for bread wheat (Gizaw et al., 2016, 2018; Liu et al., 2019; Lozada et al., 2020; Barakat et al., 2021; Krishnappa et al., 2023) and durum wheat (Nigro et al., 2019). HTP based on hyperspectral imaging in greenhouse experiments and GWAS analysis has been recently integrated for dissecting drought traits in bread wheat (Zhang et al., 2024). The use of semi-automated devices in the field increases phenotyping throughput for GWAS, and facilitates the genetic dissection of N deficiency response in bread wheat using sensors with Red-Green Blue (RGB) spectral bands and Near-infrared (NIR) mounted on a tractor (Jiang et al., 2019), and for canopy height and stem elongation rates in winter wheat by using LiDAR (Light Detection and Ranging) on the FIP platform (Field Phenotypig Platform, Kronenberg et al., 2021; Roth et al., 2024). The first report of GWAS analysis using unmanned aerial vehicles (UAVs, UAS, RPAS) was carried out for durum wheat, when the NDVI index was mapped using multispectral imaging (Condorelli et al., 2018). This was followed by the analysis of lodging traits in spring wheat using RGB and multispectral imaging (Singh et al., 2019) and the identification of QTL hotspots for VIs in rainfed wheat (Rufo et al., 2021b).

This study carried out a GWAS analysis using SNP markers (from DArTseq) to identify significant associations for YLD and vegetation spectral indices in elite durum wheat lines grown in Mediterranean environments under drought field conditions. The availability of genome sequences for durum wheat (Maccaferri et al., 2019) and bread wheat (IWGSC, 2018) enabled candidate gene analysis to identify genes involved in key crop processes including photosynthesis, plant stress responses, and hormonal regulation. In this study, we combine, for the first time in wheat, the use of an aerial HTP platform equipped with hyperspectral imaging for field phenotyping, with GWAs analysis of spectral traits, to dissect the genetic basis of yield formation under drought conditions. This approach combines the highest level of spectral resolution (hyperspectral imaging) to derive crop stress indicators with high-throughput capabilities in an aerial platform in the field.

Materials and methods

Plant materials and field trials

Field experiments were conducted using a panel of 536 durum wheat genotypes, comprising 494 elite lines from the International Maize and Wheat Improvement Center (CIMMYT) and 42 commercial varieties (Supplementary Table S1). The commercial varieties were included as a representative group of wheats adapted to the environmental conditions of the Mediterranean locations assessed in this study. The experiments were grown under rainfed conditions in two locations: Location 1 (37° 32' 17" N, 5° 06' 57" W) (Seville, Spain) in 2014, 2015, 2016, 2017, 2018 and 2021, and Location 2 (37° 27' 28"N, 6° 21' 52"O) (Huelva, Spain) in 2021. The average maximum and minimum temperatures, together with accumulated rainfall, were obtained from daily data recorded by the agroclimatic stations of the local government, Junta de Andalucía (RIA, 2023), located in the proximities of both locations. The experimental design at each location and for each experiment consisted of an augmented design with two replicated checks for 100 of the elite durum wheat lines, and a three-replicated, randomized, complete block for the 42 durum wheat varieties. For the trials, six individual row plots of 7.2 m^2 each were used, with a sowing density of 360 seeds/m^2. The wheat plots were sown between 20th November and 15th December each year and were managed following the standard agricultural practices in both locations.

DNA isolation and genotyping

The durum wheat lines were sampled at the 4th leaf stage [DC 14 on the Zadoks scale (Zadoks et al., 1974)] for genetic analyses. The plant material was collected at field trials and immediately frozen using dry ice. All the samples were preserved at -80 °C until DNA isolation. About 100mg of the frozen leaf tissue per line was used for DNA extraction with a DNeasy Plant Mini Kit (catalogue number 69104 and 69106) from (Qiagen, Hilden, Germany), following the manufacturer's protocol. The quality and concentration of each sample was assessed by electrophoresis on a 0.8% agarose gel. In addition, the restriction enzyme Tru1I (MseI, catalogue number ER0982) (ThermoFisher, Waltham, MA, USA) was used to confirm absence of nucleases in DNA prior to genotyping. Approximately 81% of the samples were genotyped by Diversity Arrays Technology Pty Ltd. (University of Canberra, Bruce, Australia) (DArT), and the remaining 19% at the Genetic Analysis Service for Agriculture (SAGA, Mexico). Sequence data for samples genotyped were first aligned against the bread wheat IWGSC RefSeq v2.0 (https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies) and Svevo durum wheat (https://www.interomics.eu/durum-wheat-genome), using end-to-end alignment.

A panel of 46,935 biallelic SNP markers was obtained (Figure 1). After thinning the marker's panel by retaining markers with a minor allele frequency (MAF) ≥ 0.05 using Tassel 5 software (Bradbury et al., 2007), the final dataset contained 10,641 biallelic SNP markers (Figure 1).

Phenotypic trait measurement and image acquisition

Yield (YLD; kg/ha) and 19 vegetation spectral indices (Table 1) were evaluated across multiple years and environments. We derived the vegetation spectral indices from high-resolution RGB, hyperspectral and thermal remote sensing imagery collected

FIGURE 1
SNP markers density raw dataset (left) and thinned dataset (right) plot chromosome wise representing the number of SNP markers. Horizontal axis shows chromosome length (Mb); color legend depicts number of SNP markers.

during several airborne campaigns. Hyperspectral imagery was spatially and atmospherically processed to obtain the vegetation indices presented in Table 1, following the methods outlined by Zarco-Tejada et al. (2016) and Camino et al. (2019). An aircraft managed by the Laboratory for Research Methods in Quantitative Remote Sensing [QuantaLab, IAS-CSIC, Spain), equipped with a micro-hyperspectral imager (Micro-Hyperspec VNIR model, Headwall Photonics, Fichburg, MA, USA), was used for acquiring the images. The flights were conducted at similar times in the crop cycle (Supplementary Table S2) to coincide with the pre-anthesis and anthesis stages of wheat [stages 49 to 69 on the Zadoks scale (Zadoks et al., 1974)]. The spectral vegetation indices used in this analysis (Table 1) were classified based on their ability to assess various physiological and structural traits in crop canopies, as follows: 1) Chlorophyll fluorescence indices, which utilize blue (e.g., BF1) and red-edge (e.g., SIF2) spectral regions to monitor photosynthetic capacity; 2) Chlorophyll indices, related to chlorophyll content and essential for assessing photosynthesis (e.g., MCARI), combining visible and near-infrared regions (NIR); 3) Carotenoid indices, reflecting the presence of carotenoids that protect against oxidative stress using green and red-edge spectral regions; 4) Xanthophyll indices, related to light management and photoprotection, such as the Photochemical Reflectance Index (PRI), which primarily utilizes the green spectral region around 550 nm; 5) Plant disease indices, assessing physiological responses to pathogens; and 6) Structural indices, which are related to biomass, leaf area, and overall structural characteristics, focusing on the red and NIR.

Population structure and linkage disequilibrium assessment

The thinned molecular markers dataset was used to assess the population structure by principal components analysis (PCA) in Tassel 5.0 (Bradbury et al., 2007). The results were then plotted in R using the '*plot*' function (R Core Team, 2020).

Linkage disequilibrium (LD) between pairs of genetic locations across the two wheat sub-genomes (A, B) were evaluated using Tassel 5 (Bradbury et al., 2007). Pairwise LD (square allele frequency, r^2) for SNP markers pairs was calculated following Weir (1997). The intersection of the fitted curve with the cut-off threshold was the mean r^2 value for each chromosome (Breseghello and Sorrells, 2006; Liu et al., 2017). LD decay was then plotted in R following Remington et al. (2001) using the mean r^2 value of each chromosome and the genetic distance in base pairs (bp).

Statistical analysis and marker-trait associations

Phenotypic correlations between assessed traits were analyzed by the 'cor' function in R (Kendall, 1938, 1945; Becker et al., 1988) across years and environments, and also plotted in R using the '*ggfortify*' package (Horikoshi and Tang, 2016).

GWAS was conducted across years and environments using best linear unbiased estimates (BLUEs) for YLD and 19 spectral indices, and 10,641 SNP markers to identify marker-trait associations using the Tassel 5.0 software (Bradbury et al., 2007). A weighted mixed linear model (W-MLM) (Casstevens and Wang, 2015) was applied using the PCA matrix, with the first and second principal components as fixed effects and the kinship matrix (K-mat) (Supplementary Table S3) as a random effect, at the optimum compression level and following the model equation:

$$y = X\beta + Z\mu + \varepsilon$$

where y is a vector of observed phenotypes; X and Z are matrices related to β and μ, respectively; β is a vector of fixed effects; μ is a vector of genetic effects (with covariance proportional to a kindship or relationship matrix); and ε is a vector of residuals. R was used to extract significant MTAs between markers and assessed traits, according to a Bonferroni-corrected threshold of $-\log_{10}(0.05/n) = 5.33$, where n is the total number of SNPs (10,641), and $\alpha = 0.05$. Manhattan and quantile-quantile (QQ) plots were visualized using

TABLE 1 Spectral indices assessed for durum wheat panels grouped by index type (in bold).

Spectral Index	Acronym	Reference
Photosynthetic Activity and Chlorophyll Fluorescence emissions		
Blue fluorescence index	BF1	Zarco-Tejada et al. (2018)
Blue fluorescence index	BF2	Zarco-Tejada et al. (2018)
Solar-induced chlorophyll fluorescence	SIF2	Plascyk and Gabriel (1975); Moya et al. (2004)
Reflectance curvature index	CUR	Zarco-Tejada et al. (2000)
Chlorophyll pigments		
Blue/green index	BGI1	Zarco-Tejada et al. (2005)
Blue/green index	BGI2	Zarco-Tejada et al. (2005)
Carotenoid xanthophyll pigment index	DCabxc	Datt (1998)
Transformed chlorophyll absorption in reflectance index/Optimized soil-adjusted vegetation index (TCARI/OSAVI)	TCARI/ OSAVI	Haboudane et al. (2002)
Normalized phaeophytinization index	NPQI	Barnes et al. (1992)
Modified chlorophyll absorption in reflectance	MCARI	Haboudane et al. (2004)
Transformed chlorophyll absorption in reflectance index	TCARI	Haboudane et al. (2002)
Carotenoid pigments		
Simple ratio carotenoids – CARter index	CAR	Hernández-Clemente et al. (2012)
Carotenoid concentration index	CRI700	Gitelson et al. (2003, 2006)
Carotenoid concentration index	CRI700m	Gitelson et al. (2003, 2006)
Carotenoid concentration index	CRI550	Gitelson et al. (2003, 2006)
Carotenoid concentration index	CRI550m	Gitelson et al. (2003, 2006)
Xanthophyll indices		
Photochemical reflectance index	PRI	Gamon et al. (1992)
Carotenoid and Xanthophyll pigments		
Carotenoid xanthophyll pigment index	DCabxc	Datt (1998)
Assessing Plant Health and Disease Stress		
Health index	HI	Mahlein et al. (2013)
Structural and biomass changes		
Normalized difference vegetation index	NDVI	Rouse et al. (1973)

the R package '*Cmplot*' (Yin et al., 2021) (script can be found at https://github.com/YinLiLin/CMplot).

Candidate gene analysis

As described in Mérida-García et al. (2020), the sequences of associated SNP markers were blasted against the bread wheat reference assembly RefSeq v2.0 (https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies) and the Svevo durum wheat reference assembly (https://www.interomics.eu/durum-wheat-genome), with no indels or mismatches allowed, using an *ad hoc* Java program to confirm the physical mapping location on each genome. To estimate the position of the MTAs, measured in centimorgans (cM), a map of correspondences between the positions in bp and cM was created for every Svevo chromosome. This map uses the data provided in Supplementary Table 2 of Maccaferri et al. (2014), which provides a large set of markers, including their nucleotide sequences, and their estimated cM positions on the correct chromosome. To calculate their positions in bp, a BLAST search into the Durum Interomics pseudomolecules (https://doi.org/10.1038/s41588-019-0381-3) was performed (parameter -ungapped). From the resulting map, the only markers retained were those with a public sequence or available for research purposes, and with a single best hit (maximum bitscore) in the correct chromosome. Finally, the map was sorted by chromosome and cM, and checked to remove those markers whose positions in bp were unsorted. Using the resulting map, and knowing the

10.3389/fpls.2024.1470520

TABLE 2 Physical position (cM) for marker-trait associations based on Maccaferri et al. (2014).

Marker	Chr	Pos (cM)	Traits
SNP229	1A	29.0	CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP76228	1A	61.0	CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP1275	1A	71.0	BF1, BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP1276	1A	71.0	BF1, BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP1481	1A	85.5	BF1, BF2, BGI1, CAR, CUR, DCabxc, MCARI, NDVI, PRI, SIF2, TCARI_OSAVI, TCARI
SNP77275	1A	111.6	CRI700m
SNP2019	1A	123.2	CAR, CUR, MCARI, NDVI, PRI
SNP2648	1B	12.0	CRI550, CRI550m
SNP2830	1B	33.6	MCARI, NDVI, PRI
SNP26551	1B	37.6	CAR, YLD
SNP3098	1B	46	CUR, MCARI, PRI
SNP3549	1B	47.8	BF1, BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP981	1B	48.8	CAR
SNP3877	1B	65.0	YLD
SNP3937	1B	67.5	CUR, MCARI, TCARI
SNP23059	1B	93.6	CAR, MCARI, TCARI_OSAVI, TCARI
SNP5762	1B	136.2	YLD
SNP32258	1B	156.2	YLD
SNP45417	2A	38.2	BF1, BF2, BGI1, CAR, CUR, DCabxc, MCARI, NDVI, PRI, SIF2, TCARI_OSAVI, TCARI
SNP46534	2A	46.6	BF1, BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, SIF2, TCARI_OSAVI, TCARI
SNP6223	2A	52.1	MCARI
SNP33554	2A	91.0	CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI
SNP6626	2A	96.9	BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP6675	2A	99.9	CUR, PRI
SNP7069	2A	109.6	BF1, BF2, BGI1, CAR, CUR, DCabxc, MCARI, NDVI, PRI, SIF2, TCARI_OSAVI, TCARI
SNP7835	2A	132.2	MCARI, PRI, TCARI
SNP11390	2A	136.2	CAR, CUR, MCARI, PRI, TCARI
SNP8165	2A	151.2	MCARI
SNP8198	2A	154.6	BF1, BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP8232	2A	154.6	CAR, CUR, MCARI, PRI
SNP13427	2A	163.3	BF1, BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP46141	2A	210.8	CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI
SNP46142	2A	210.8	BF1, BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP9483	2B	24.7	BF1, BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP9484	2B	24.7	BF1, BF2, BGI1, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP70996	2B	45.3	MCARI, PRI
SNP9901	2B	55.4	BF1, BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP9976	2B	57.7	CAR, YLD
SNP10568	2B	91.3	HI

(Continued)

10.3389/fpls.2024.1470520

TABLE 2 Continued

Marker	Chr	Pos (cM)	Traits
SNP10840	2B	95.3	BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI
SNP10841	2B	95.3	CAR, CUR, MCARI, NDVI, PRI
SNP46997	2B	101.6	CAR, CRI550m
SNP11217	2B	115.1	CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP13388	2B	137.9	BF1, BF2, BGI1, CAR, CUR, DCabxc, MCARI, NDVI, PRI, SIF2, TCARI_OSAVI, TCARI, YLD
SNP43735	2B	166.6	CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI
SNP12651	2B	181.6	BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP46683	3A	7.9	CAR, CUR, MCARI, PRI, TCARI
SNP77245	3A	33.6	BF1, BF2, BGI1, CAR, CUR, DCabxc, MCARI, NDVI, PRI, SIF2, TCARI_OSAVI, TCARI
SNP14668	3A	66.8	MCARI
SNP14760	3A	67	CAR, PRI
SNP15000	3A	80.1	CAR, HI, YLD
SNP38516	3A	81.4	CAR, YLD
SNP15180	3A	90.3	BF1, BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP15291	3A	97.4	BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP15292	3A	97.4	CAR, NDVI, PRI
SNP15681	3A	123.1	BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP15835	3A	136.4	YLD
SNP76391	3B	16.9	BF1, BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP16842	3B	25.4	CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI
SNP17449	3B	68.3	CAR
SNP17455	3B	69.1	CUR, DCabxc, MCARI, NDVI, PRI, TCARI
SNP17862	3B	81.2	BF1, BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP18017	3B	88.0	CAR, MCARI, PRI
SNP20210	3B	136.9	BF1, BF2, BGI1, CAR, CUR, DCabxc, MCARI, NDVI, PRI, SIF2, TCARI_OSAVI, TCARI, YLD
SNP76785	3B	136.9	BF1, BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, SIF2, TCARI_OSAVI, TCARI
SNP76832	3B	136.9	BF21, BF2, BGI1, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP20617	4A	15.5	YLD
SNP21331	4A	57.3	CAR, MCARI, PRI
SNP21648	4A	65.1	CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP21687	4A	69.4	BF1, BF2, BGI1, CAR, CUR, DCabxc, MCARI, NDVI, PRI, SIF2, TCARI_OSAVI, TCARI
SNP21759	4A	79.3	BF1, BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP22313	4A	133.9	BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI
SNP23640	4B	41.6	CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP24067	4B	52.9	CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP25678	5A	20.6	BF1, BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP25731	5A	26.9	BF1, BF2, BGI1, CAR, CUR, DCabxc, MCARI, NDVI, PRI, SIF2, TCARI_OSAVI, TCARI, YLD
SNP16002	5A	27.4	CAR, CUR, MCARI, NDVI, PRI
SNP26048	5A	48.3	CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI

(Continued)

10.3389/fpls.2024.1470520

TABLE 2 Continued

Marker	Chr	Pos (cM)	Traits
SNP47527	5A	48.6	CAR
SNP47528	5A	48.6	CAR
SNP47529	5A	48.6	CAR, PRI
SNP47530	5A	48.6	CAR
SNP47531	5A	48.6	CAR
SNP47532	5A	48.6	CAR
SNP47533	5A	48.6	CAR
SNP47536	5A	48.6	CAR, CRI550m
SNP47537	5A	48.6	CAR
SNP26845	5A	90.3	BF1, BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP28567	5B	6.5	BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI
SNP29706	5B	54.4	CAR
SNP29849	5B	68.5	BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP31932	5B	75.9	PRI
SNP32147	5B	146.1	CAR, NPQI, YLD
SNP27817	5B	148.4	BF1, BF2, BGI1, CAR, CUR, DCabxc, MCARI, NDVI, PRI, SIF2, TCARI_OSAVI, TCARI
SNP30955	5B	150.9	BF2, CAR, CUR, MCARI, NDVI, PRI
SNP32334	6A	0.9	CAR, CUR, MCARI, NDVI, PRI
SNP32837	6A	44.3	BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP32939	6A	49.7	BF1, BF2, BGI1
SNP33144	6A	53.2	MCARI
SNP33615	6A	63.5	CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP33665	6A	67.3	BF1, BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP34751	6B	21.6	CAR, NPQI, YLD
SNP34891	6B	27.1	BF1, BF2, BGI1, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP34892	6B	27.1	CAR, CUR, MCARI, PRI, TCARI
SNP47538	6B	31.2	CAR, PRI
SNP47539	6B	31.2	CAR
SNP47540	6B	31.2	CAR
SNP47541	6B	31.2	CAR, PRI
SNP35255	6B	45.7	CAR, MCARI, PRI
SNP13219	6B	52.5	BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP37933	6B	76.3	BF1, BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP37996	6B	86.0	BF1, BF2, BGI1, CAR, CRI550m, CUR, DCabxc, MCARI, NDVI, PRI, SIF2, TCARI_OSAVI, TCARI, YLD
SNP36835	6B	86.2	HI
SNP37315	6B	96.7	BGI2
SNP76145	6B	137.2	CAR, CUR, MCARI, NDVI, PRI, TCARI
SNP73562	7A	8.4	YLD

(Continued)

TABLE 2 Continued

Marker	Chr	Pos (cM)	Traits
SNP38478	7A	14.3	YLD
SNP38846	7A	53.4	CAR, CUR, MCARI, NDVI, PRI, TCARI
SNP38848	7A	53.4	BF1, BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP40233	7A	113.6	CAR
SNP40908	7A	147.5	CAR, CRI550m, PRI, YLD
SNP76958	7A	170.7	BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP46389	7A	170.8	BF1, BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP41357	7A	172.9	BF2, CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI
SNP45972	7B	54.0	HI, YLD
SNP43797	7B	96.1	BF1, BF2, BGI1, CAR, CUR, DCabxc, MCARI, NDVI, PRI, SIF2, TCARI_OSAVI, TCARI
SNP44041	7B	109.3	CAR
SNP21473	7B	195.9	CAR
SNP45528	7B	208.3	CAR, CUR, DCabxc, MCARI, NDVI, PRI, TCARI_OSAVI, TCARI

positions in bp of our markers, their positions in cM were interpolated. To compare with the meta-QTL (MQTL) analysis reported by Soriano et al. (2017), the physical position of the MQTLs was inferred based on the closest DArTseq or SNP marker to the MQTL. A confidence interval of 5kbp to the left and right of the marker was established.

Candidate genes were identified and manually chosen based on their annotations within a window of ±50kbp. Gene expression analyses were performed using the publicly available transcriptomics analyses under different heat and drought stress conditions previously published for bread wheat (Liu et al., 2015; Ma et al., 2017; Gálvez et al., 2019). These results were drawn as a heatmap using the data retrieved by Wheat Expression (www.wheat-expression.com/) and the R package 'NMF 0.21.0' (Gaujoux and Seoighe, 2010). The samples analyzed were: (1) seedling samples grown under controlled conditions included in NCBI SRA ID SRP045409 (control, IS; heat and drought (PEG induced drought) stress for 1 and 6 hours, PEG1 and PEG6, respectively) (Liu et al., 2015); (2) samples grown in a shelter and corresponding to NCBI SRA ID SRP102636 (anther stage irrigated leaf phenotype, AD_C; anther stage drought-stressed leaf phenotype, AD_S; tetrad stage irrigated developing spike phenotype, T_C; and tetrad stage drought-stressed developing spike phenotype, T_S) (Ma et al., 2017); and (3) flag leaf samples from field experiments corresponding to NCBI SRA ID SRP119300 (irrigated, IF; mild stress, MS; and severe stress, SS, flag leaves samples) (Gálvez et al., 2019).

Results

Agroclimatic conditions

Locations 1 and 2 are both in Mediterranean climate-zones, characterized by hot and dry summers, and short and mild winters with irregular precipitation. Figure 2 shows the patterns for maximum and minimum average temperatures (°C) and monthly accumulated precipitation (mm) during the crop cycle (from November to June) for each growing season in the two locations. For Location 1, the 2018 season was the wettest, with 488 mm of precipitation, whereas the driest was 2015, with 243 mm (Figure 3).

Figure 2 reveals increasing temperatures from March until the end of the crop season for all the years assessed, together with irregular precipitation throughout the crop cycle in both testing locations. Yearly variations in precipitation and temperatures were reflected in the differences found in the final YLD (Figure 3). However, this relationship was not always clearly evident, with contrasting patterns sometimes being found, as for season 2015 in Location 1 and season 2021 in Location 2 (Figure 3), which could be attributed to high soil fertility, as suggested by Royo et al. (2021).

Phenotypic analyses

The yearly means of crop final yield (Kg/ha) are shown in Figure 4. Variations in precipitation and temperatures were reflected in the differences found in the final yield (Figure 4). For instance, Location 1 had the highest values of YLD in 2018 (5,250 kg/ha), likely due to the highest level of accumulated precipitation during the crop cycle (Figures 3, 4). However, this relationship was not obvious in some cases, with different patterns found, as in Location 2 in 2021, with a high average yield (5,605 kg/ha), although the accumulated precipitation (352 mm) was not significantly different from the average yearly rainfall.

Phenotypic correlations were found between yield and HSIs related to plant photosynthesis processes, and canopy structure and density (Figure 5), which directly or indirectly affected final crop production. Positive correlations (r = 0.30) between YLD and the structural index NDVI (normalized difference vegetation) were also found, as previously reported by Basnyat et al. (2004); Chandel et al.

10.3389/fpls.2024.1470520

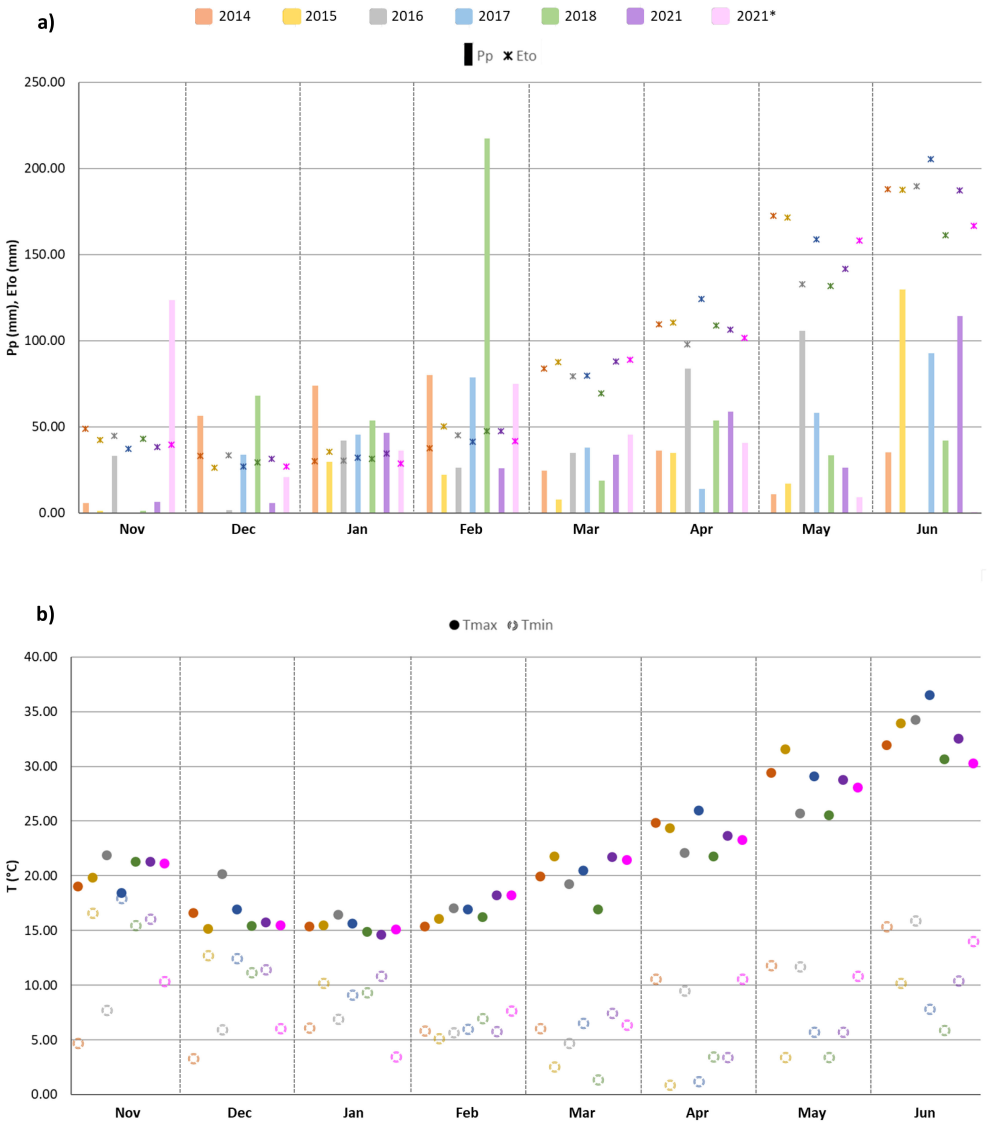


FIGURE 2
(A) Monthly precipitation (Pp, mm) and evapotranspiration (Eto, mm); **(B)** average maximum (Tmax, °C) and minimum (Tmin, °C) temperatures (below) during the crop cycle (from November to June) for each growing season (2014, 2015, 2016, 2017, 2018 and 2021 in Location 1; 2021* in Location 2).

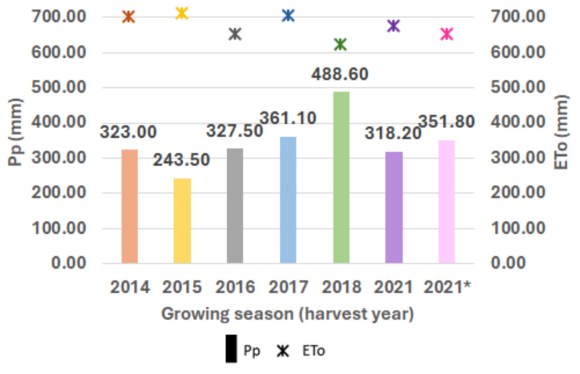


FIGURE 3
Crop cycle rainfall variation (y-axis) for each growing season (x-axis). Accumulated precipitation (mm) is indicated above each boxplot. Evapotranspiration (mm) at Location 1 for each growing season (2014, 2015, 2016, 2017, 2018 and 2021 in Location 1; 2021* in Location 2).

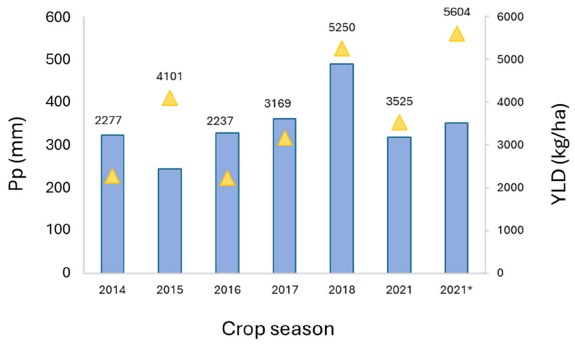


FIGURE 4
Mean values of yield and crop cycle accumulated precipitation (mm) at Location 1 for each year of assessment. Pp: yearly accumulated precipitation (mm); YLD: yield (kg/ha); 2021*: means corresponded to Location 2 in 2021.

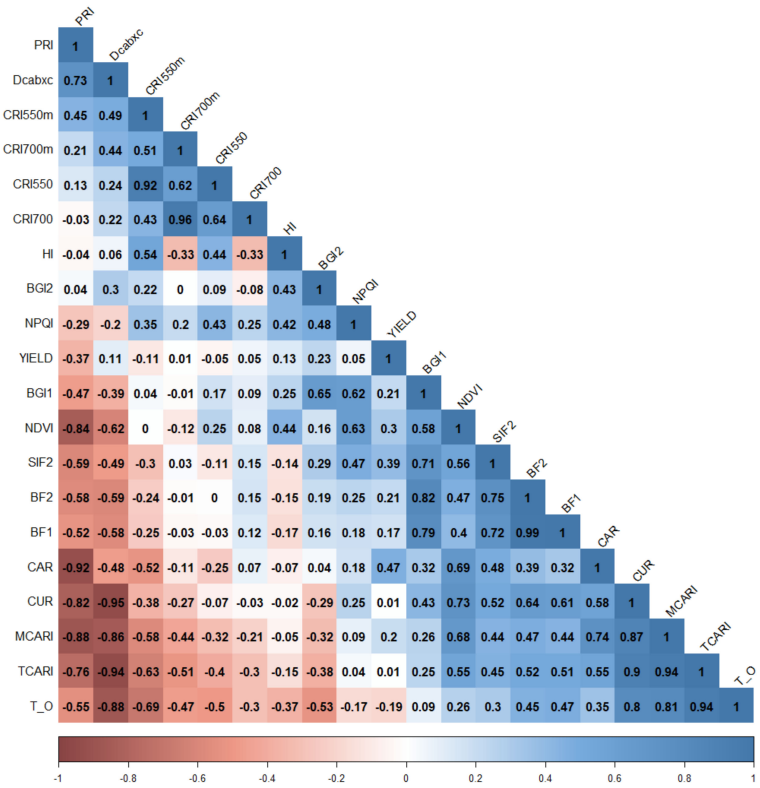


FIGURE 5
Phenotypic correlations between assessed traits across years and environments. Spectral index abbreviations are shown in Table 1. Color intensities show degrees of positive and negative significance ($p < 0.05$).

(2019) and Rufo et al. (2021a), using different wheat populations. Moreover, phenotypic correlations were found between YLD and HSIs related to plant photosynthesis processes (indices of simple carotenoid ratio (CAR, r = 0.47), solar-induced chlorophyll fluorescence (SIF2, r = 0.39) and photochemical reflectance (PRI, r = -0.37)) (Figure 5). This highlights the important impact of the assimilation processes during grain filling on final yield. As expected, correlations were also found between spectral indices which were classified in the same group (Figure 5; Table 1).

Significant correlations between HSIs were also observed, several of which were observed between spectral indices belonging to the same categories as described above (Table 1). These correlations suggest that these indices capture similar spectral regions that are sensitive to plant traits, such as the chlorophyll region or Solar-Induced Fluorescence (SIF) emission related to photosynthetic capacity. Furthermore, correlations were found between different groups of indices, exemplified by connections between indices of a group of chlorophyll pigments (MCARI, TCARI, and TCARI/OSAVI) with those of a group including photosynthetic activity and chlorophyll fluorescence emission (CUR) (Figure 5). Indices from this group were also found to be closely correlated to those of a group of carotenoid and xanthophyll pigments (DCabxc) (Figure 5).

Population structure and linkage disequilibrium

In the PCA analysis, the first and second principal components (PC) accounted for 3.3% and 2.6% of the genetic variation, respectively (Figure 6). No genetic sub-structure was identified in the panel. LD decay was estimated around 3.98 kbp for all the chromosomes (Figure 7).

Marker-trait associations and candidate genes

A total of 740 significative marker-trait associations were identified for the 20 analyzed traits (Supplementary Table S4). A summary of the results for all the traits across years and environments is reported in Figure 8. The physical position of the associated markers is shown in Supplementary Figure S1. Manhattan and QQ-plots can be found in Supplementary Figure S2. 721 SNP markers were linked to spectral indices (Supplementary Table S4) and 19 to YLD. Twelve of the latter were also associated with one or more spectral indices. The carotenoid index (CAR), which was correlated with YLD

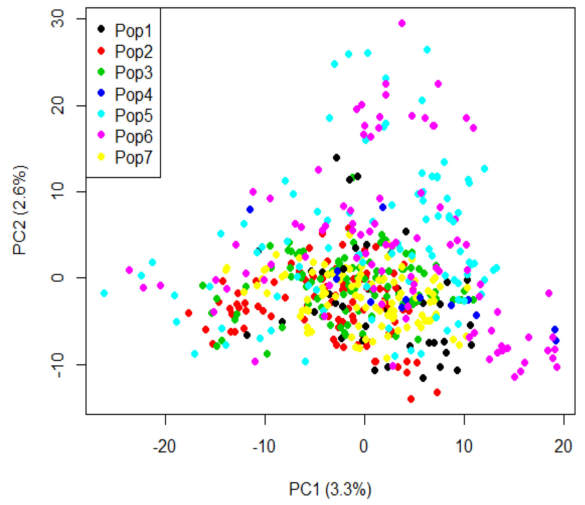


FIGURE 6
Principal component analysis of genotypic data using 10,641 SNP markers. Pop1: durum wheat varieties; Pop2: elite durum wheat lines collected in 2014; Pop3: elite durum wheat lines collected in 2015; Pop4: elite durum wheat lines collected in 2016; Pop5: elite durum wheat lines collected in 2017; Pop6: elite durum wheat lines collected in 2018; and Pop7: elite durum wheat lines collected in 2021.

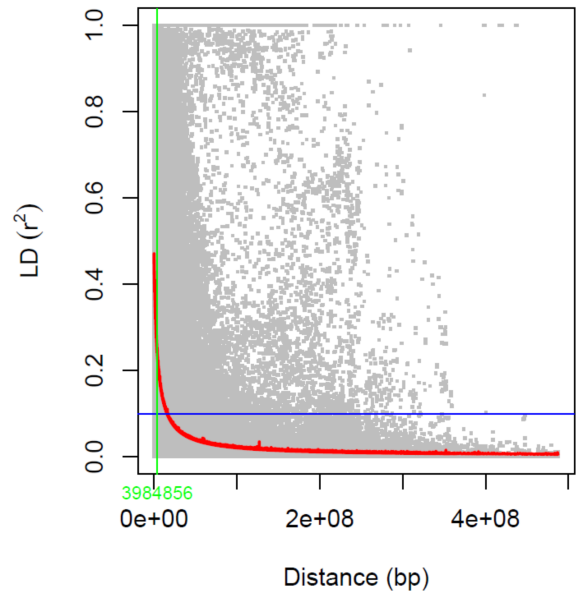


FIGURE 7
Linkage disequilibrium (LD) decay analysis using SNP markers. Estimated r^2 against linkage distance in bp is shown. LD decay was measured at 3.98 kbp.

(Figure 5), showed the highest number of significative associations, with 14% of the total number of MTAs, followed by the photochemical reflectance index PRI (11.62%), as well as a moderate correlation with YLD (Figure 5), and the chlorophyll index MCARI (11.35%), which is sensitive to chlorophyll a+b content. The indices of CUR (10.00%), sensitive to fluorescence emission, NDVI (9.32%), related to structure, TCARI (9.19%), sensitive to chlorophyll, and DCabxc (8.11%), sensitive to chlorophyll and carotenoids (Figure 8), also showed a medium-high number of significative associations (Figure 8). Fifteen of the 19 indices analyzed showed co-localization with YLD (Supplementary Table S7). Among these, the CAR, PRI, MCARI and CUR indices were those with the highest number of co-localized MTAs, with 32, 25, 23 and 22, respectively. To our knowledge, this is the first report of co-localization of fluorescence (BF1, BF2, CUR, SIF2), chlorophyll a+b (BGI1, DCabxc, TCARI/OSAVI, NPQI, MCARI, TCARI), carotenoid (CAR, CRI550m), plant disease (HI) and xanthophyll (PRI) indices with YLD.

The number of MTAs per chromosome across years and environments ranged from 11 on wheat chromosome 4B to 100 on chromosome 2B (Figure 8). Genome A accounted for 54.05% of the total marker-trait associations (320 MTAs), and the remaining 45.95% (272 MTAs) corresponded to genome B.

The physical position for MTAs (Table 2) and MQTLs, previously described in Soriano et al. (2017) and Arriagada et al. (2022) for each chromosome, are shown in Figure 9.

The QTLs found for yield, some of which were shared with one or more vegetation spectral indices, were placed in several chromosomes (Supplementary Table S4), which agree with the different QTLs and MQTLs described in previous studies (Soriano et al., 2017; Anuarbek et al., 2020; Arif et al., 2020; Farouk et al., 2021; Mangini et al., 2021; Arriagada et al., 2022; Mulugeta et al., 2023; Valladares García et al., 2023) for durum wheat yield or yield-related traits. Eleven markers (*SNP26551, SNP9976, SNP13388, SNP15000, SNP38516, SNP20210, SNP25731, SNP32147, SNP34751, SNP37996* and *SNP40908*) were found to be significatively associated with yield and the simple carotenoid ratio index (some of them were also associated with other spectral indices, see Supplementary Table S4), which were well correlated (r = 0.47) (Figure 5), in agreement with the importance and influence of carotenoids on yield as precursors of vitamin-A and plant hormones involved in plant growth and its responses to adverse growth conditions (Mi et al., 2022). Marker *SNP9976*, mapped on durum wheat chromosome 2B, was found within the MQTL15 (Soriano et al., 2017), described for grain weight (GW), and within a grain yield MQTL found under irrigated conditions, as described in Arriagada et al. (2022) (Figure 9). *SNP32147* and *SNP34751*, were found to be significantly associated with yield and the normalized phaeophytinization index (Supplementary Table S4), which belongs to chlorophyll pigments group (Table 1), and is thus related to the process of photosynthesis in the plant. It has been described in previous studies as an efficient SRI for indirect selection of grain yield (Liu et al., 2019). In addition, *SNP32147*, mapped on wheat chromosome 5B, was found in proximity to the MQTL49 (Soriano et al., 2017), related to GY and GW (Figure 9). *SNP15000* and *SNP45972* were both significatively associated with yield and health index (Supplementary Table S4), which has been previously described as relevant for yield estimation in spring wheat and used to determine patterns of drought distribution in agricultural areas (Zuhro et al., 2020). *SNP15000* and *SNP38516* (also associated with YLD), both mapped on wheat chromosome 3A, were found in

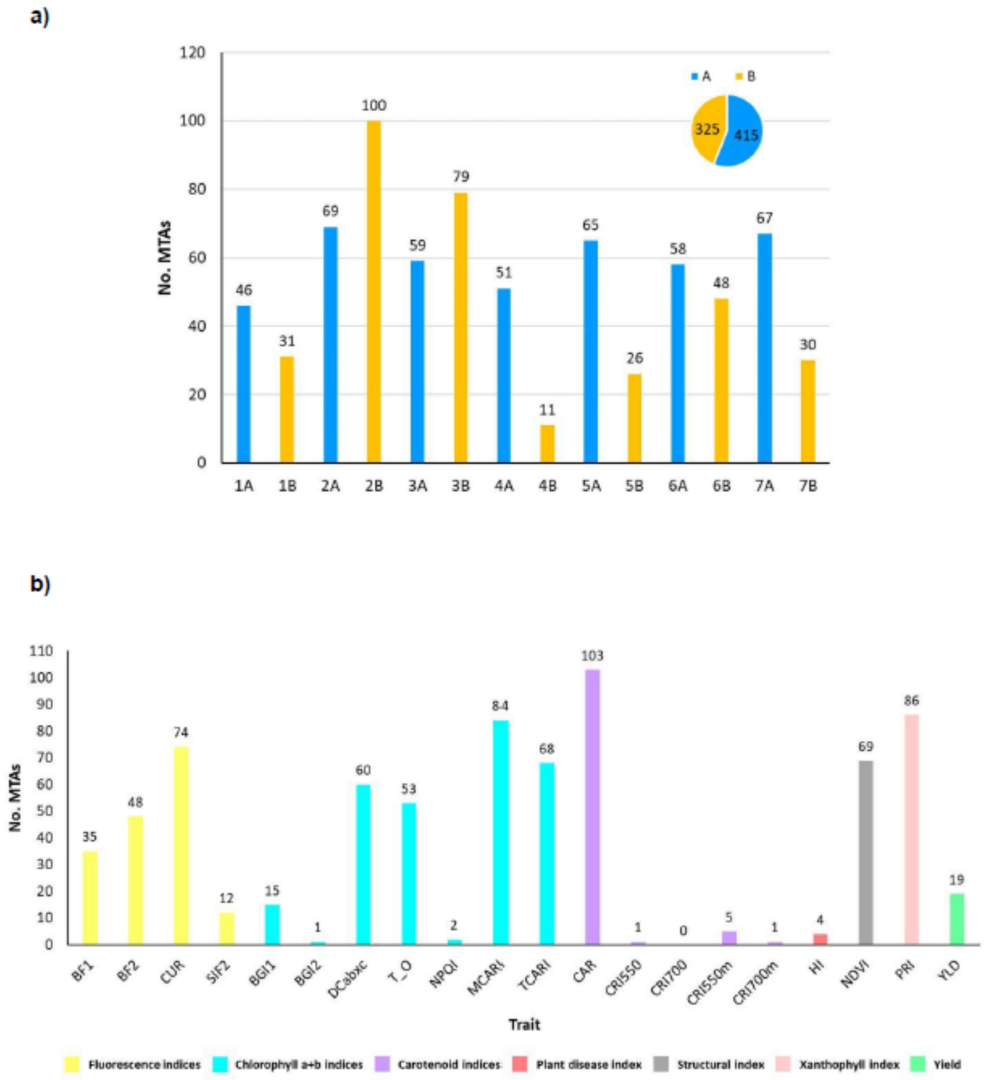


FIGURE 8
(A) Number of MTAs found for each chromosome. Bars for genome A chromosomes are indicated in blue, genome B in yellow; **(B)** Number of MTAs for each trait assessed. Chr: chromosome; No. MTAs: number of significant marker-trait associations; A: durum wheat genome A; B: durum wheat genome B; spectral indices abbreviations are given in Table 1; YLD, yield (Kg/ha).

the proximity of a YLD MQTL (irrigated conditions) and a yield and yield-related traits MQTL (rainfed conditions), respectively, both described in Arriagada et al. (2022) (Figure 9). Finally, *SNP73562* and *SNP38478*, both mapped on wheat chromosome 7A and associated with YLD, were found within the MQTL59 (Soriano et al., 2017) described for GW.

The search for candidate genes aimed to identify corresponding gene models, in durum and bread wheat. We also analyzed the corresponding gene expression under different drought levels, and stress conditions were performed for the significantly associated markers. Gene annotation from the durum wheat genome (https://www.interomics.eu/durum-wheat-genome) allowed the identification of 695 candidate genes (Supplementary Table S5). Among these, there were 244 HC genes related to different plant processes including stress responses, but also photosynthesis, and structural and regulatory plant biological processes. Of these, we can highlight the HC genes which encode photosystem I and II assembly proteins, NAD(P)H-quinone oxidoreductases, cytochrome subunits, F-box family proteins, disease resistance proteins, kinase family proteins, aspartic proteinases, or glycosyltransferases, among others (Supplementary Table S5). Most orthologs of these genes were also found in gene annotation from the bread wheat reference assembly RefSeq v2.0 (https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies) (Supplementary Table S6). The results for the gene expression analyses under different stress conditions (Liu et al., 2015; Ma et al., 2017; Gálvez et al., 2019) are shown as a heatmap in Supplementary Figure S3.

Discussion

This study focused on the phenotypic and yield response of elite durum wheat in field experiments conducted under Mediterranean

10.3389/fpls.2024.1470520

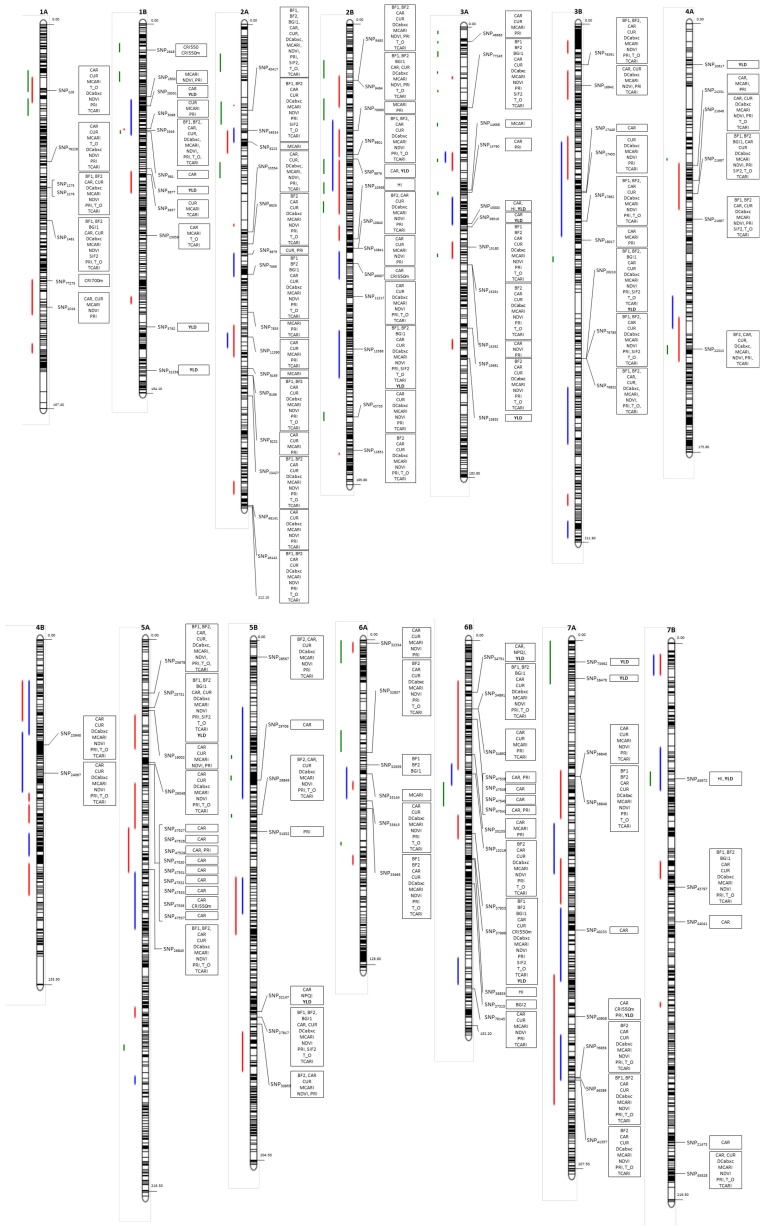


FIGURE 9
Genetic map for the significant marker-trait associations [position in cM based on Maccaferri et al. (2014)] based on the previously-described yield or yield-related traits (Soriano et al., 2017; Arriagada et al., 2022) for each chromosome. Green: MQTL from Soriano et al., 2017, blue: MQTLs detected under irrigated conditions (Arriagada et al., 2022); red: MQTLs detected under rainfed conditions (Arriagada et al., 2022).

field conditions. These growing environments are characterized by irregular precipitation during the crop growth cycle, and high temperatures during anthesis and grain filling (Araus et al., 2002; Barakat et al., 2016), exhibiting varying environmental conditions influenced by climate change, specifically characterized by heat and drought.

Marker trait associations

The dissection of the genetic basis of complex traits is a key objective in breeding programs (Rufo et al., 2021a). In this context, the identification of marker-trait associations as well as QTL related to traits of interest, such as final yield in wheat, are major goals in plant breeding (Arriagada et al., 2020), and can be encompassed within the objectives of these breeding programs. Hyperspectral indices have recently been proposed to assist the genetic dissection of bread (Rasheed et al., 2014; Crain et al., 2018; Chandel et al., 2019; Jiang et al., 2019; Liu et al., 2019; Singh et al., 2019; Rufo et al., 2021a; Govta et al., 2022; Yu et al., 2024; Zhang et al., 2024) and durum wheat traits (Condorelli et al., 2018b; Zendonadi dos Santos et al., 2021; Safdar et al., 2023). The present GWAS analysis resulted in 740 significant MTAs for yield and HSIs, distributed across all durum wheat chromosomes (Supplementary Table S4). The pseudomolecule position distribution

reveals several QTLs where the HSIs co-locate with yield (Supplementary Table S7), highlighting the potential of HSIs as spectral plant traits in yield genomic dissection.

The NDVI (Normalized Difference Vegetation Index), which is an indicator of the plant structure and response to drought (Ji and Peters, 2023) and belongs to the structural and biomass changes indices group, showed 69 associations across almost all durum wheat chromosomes (Supplementary Table S4), in agreement with the results presented by Condorelli et al. (2018). Nineteen of them (27.54%) are co-localized with yield associations on several chromosomes (2B, 3A, 3B, 5A, 5B, 6B and 7A). Importantly, *SNP43735*, mapped on wheat chromosome 2B and linked to several HSIs including NDVI, was found within the MQTL19 (Soriano et al., 2017) described for index NDVI. Previous studies, including Xiao et al. (2013) and Li et al. (2014), reported QTL for NDVI which co-locate with YLD on chromosome 3AS. Our results are in agreement with the influence of vegetative growth on final wheat production (Chandel et al., 2019), and thus the relation that can be found between this vegetation spectral index and final crop productivity (Labus et al., 2010). As Li et al. (2020) recently highlighted, water and nitrogen availability can be considered as highly limiting factors in crop production. In fact, nitrogen is the most important element for plant growth and development, affecting the biochemical and physiological functions of the plant and also increasing final yield (Leghari et al., 2016). There are different studies on this topic which used NDVI to predict or estimate final yield in winter and durum wheat, including Chandel et al. (2019), who concluded that the NDVI-YLD relation was stronger in the heading stage (96% accurate estimation of grain and biomass yields in irrigated wheat), and Panek and Gozdowski (2020), whose results agree with Chandel et al., paving the way for forecasting cereal grain yield. *SNP32837* (mapped on chromosome 6A), found in association with several spectral indices including NDVI, was located within the MQTL54 (Soriano et al., 2017) described for GW. Moreover, NDVI has been associated with drought-adaptive traits, as well as grain yield, under stress conditions in wheat crops (Bort et al., 2005; Reynolds et al., 2007; Bowman et al., 2015; Tattaris et al., 2016; Yousfi et al., 2016).

Numerous marker-trait associations were found for HSIs, including indices of group of xanthophyll pigments (86 MTAs), chlorophyll pigments (such as MCARI and TCARI, with 84 and 68 MTAs, respectively) or photosynthetic activity and chlorophyll fluorescence emission (e.g., CUR, with 74 MTAs), among other photosynthesis-related spectral traits (Supplementary Table S4). These associations resulted of great interest and value thanks to their relation to photosynthesis processes in plants, as well as to their direct or indirect relation to yield, since, as Paul (2021) highlighted, the role of photosynthesis is pivotal in driving the biological processes involved in final crop yields.

The identification of molecular markers linked to yield and/or different hyperspectral indices facilitates their subsequent use in marker-assisted selection (MAS) or other applications in wheat breeding programs. Moreover, the common associations found for yield and some HSIs can be applied to select vegetation indices as possible estimators of yield, and used for monitoring the development of the crop more efficiently across different growth stages.

The main innovation of this study is the use of high-resolution hyperspectral cameras, opening up new possibilities for exploring a broader spectrum of spectral indices. The versatility of hyperspectral imaging provides researchers with a more comprehensive dataset for characterizing plant traits associated with essential photosynthetic processes. This strategic use of hyperspectral imaging not only advances our understanding of plant physiology but also contributes significantly to the remote sensing community, showcasing its potential to uncover a deeper layer of information for enhanced crop monitoring and phenotypic analysis.

Candidate gene analysis

This study used hyperspectral indices assessed during the pre- and anthesis stages and final yield phenotyped in durum wheat lines grown under hot, dry, Mediterranean field conditions to perform a combined GWAS analysis. This combined approach unveiled specific genomic regions associated with crop adaptation and yield in response to the challenging climatic conditions of heat and drought.

SNP1275 (1A chromosome) and *SNP1276* (1B), both significantly linked to various HSIs (Supplementary Table S4), were found in the proximity (-8bp) of the durum wheat HC genes *TRITD1Av1G177430.1* and *TRITD1Bv1G163490.1*, respectively. Both genes encode a membrane-associated kinase regulator G, an enzyme which belongs to the protein kinase family, which are involved in plant stress response as regulatory components and in controlled cellular activities (Wang et al., 2020). This agrees with the decreased expression of both genes under increasing drought stress conditions in the field (Supplementary Figure S3). Markers *SNP46997* (1B), *SNP47527, SNP47529, SNP47530, SNP47532* and *SNP47537* (all mapped on chromosome 3B) are all associated with the carotenoid index (CAR), related to the pigment pool involved during the photosynthesis process. These markers were found in the vicinity (within the window of ±50kbp) of several HC genes (Supplementary Table S6) which encoded photosystem I P700 chlorophyll a apoproteins and photosystem II CP47 reaction center proteins. The photosystems I and II play important roles in photosynthesis processes (Gao et al., 2018), and carotenoids are part of their co-factors (Fromme et al., 2001; Gao et al., 2018), with some of these genes decreasing their expression under stress conditions (Supplementary Figure S3). *SNP3549* (1B), associated with several fluorescence, chlorophyll, structural and xanthophyll indices (Supplementary Table S4), was found in the surroundings (within a window of ±15kbp) of gene *TRITD1Bv1G134220.2*, which encodes an ABC family transporter (Supplementary Table S6). ABC transporters have been described in plants as proteins which play key roles in plant growth, nutrition, development, response to abiotic stresses, and interaction with its environment (Kang et al., 2011). In keeping with the ABC protein function, this gene decreases its expression under field drought stress conditions (Supplementary Figure S3).

The *SNP46534* (2A), associated with different HSIs (Supplementary Table S4), was found in the proximity (-3262bp) of the gene *TRITD2Av1G018540.1*, which encodes a cytochrome P450 family protein. It is an enzymatic protein with a key role in plant development and stress defense (Ohkawa et al., 1998; Li et al.,

2012; Jun et al., 2015). This gene is also involved in plant development, showing decreased expression under field drought conditions (Supplementary Figure S3).

Similarly, *SNP77245* (3A), associated with several HSIs (Supplementary Table S4), was found in the surroundings (within the window of ±50kbp) of genes *TRITD3Av1G017220.1, TRITD3Av1G017230.1, TRITD3Av1G017250.1* and *TRITD3Av1G017260.1*, which also encode cytochrome P450 proteins. These genes also showed decreased expression under increased field drought conditions and PEG stress treatment (Supplementary Figure S3). A similar reaction in gene expression was found for *TRITD3Bv1G076220.1*, which encodes a cytochrome b559 subunit alpha (Supplementary Figure S3), one of the main components of the photosystem II reaction center (Chu and Chiu, 2016). This gene was found in the proximity (-42250bp) of marker *SNP47527* (3B), which is associated with the carotenoid index CAR. In the vicinity (-21,235 bp) of *SNP47529* (3B), associated with CAR and PRI, we found gene *TRITD3Bv1G076150.1*, which encodes a cytochrome b6 (Cytb6), a protein specific to chloroplasts which participates in the electron transport chain in photosynthesis (Cramer, 2020). In the expression heatmap shown in Supplementary Figure S3, this gene increased its expression under heat and drought stress field conditions (including IF, control) and under PEG stress treatment, but decreased it under AD_C, T_C and T_S conditions (anther stage irrigated leaf phenotype, tetrad stage irrigated developing spike phenotype and tetrad stage drought-stressed developing spike phenotype, respectively). *SNP46997* (1B), was linked to carotenoid indices CAR and CRI550m. This marker was found in proximity (-17bp) to gene *TRITD1Bv1G206480.4* (Supplementary Table S6), encoding a NAD(P)H-quinone oxidoreductase subunit 2, which plays crucial roles in several biological plant processes including photosynthesis (Hu et al., 2018). Moreover, the candidate gene analysis showed another 7 HC genes in the same durum wheat chromosome 1B, albeit more distanced (within a window of -15 and -20kbp) of the *SNP46997* (see Supplementary Table S6), which form a cluster (*TRITD1Bv1G206400.1, TRITD1Bv1G206410.1, TRITD1Bv1G206420.1, TRITD1Bv1G206420.2, TRITD1Bv1G206420.3, TRITD1Bv1G206420.4* and *TRITD1Bv1G206430.1*), all of which encode a NAD(P)H-quinone oxidoreductase subunit 1. *SNP5762* (1B), associated with yield, was found in the proximity (-169bp) of gene *TRITD1Bv1G215590.1* (Supplementary Table S6), which encodes an aspartic proteinase. This enzyme has been described as part of a group of enzymes related to gliadins in the wheat endosperm (Belozersky et al., 1989). As described in Tenikecier and Genctan (2020), increased or decreased seed size, influenced by endosperm size, affects the final yield in wheat. Moreover, the chromosome where this gene was mapped agrees with previous wheat studies which described major genomic regions for gluten strength and genes related to endosperm proteins as gliadins (Kaan et al., 1993; Ruiz and Carrillo, 1993; IWGSC, 2018; Mérida-García et al., 2019). *SNP9483* and *SNP9484*, both located on wheat chromosome 2B, and *SNP13427* (2A) associated with several HSIs (Supplementary Table S4), were found in the proximity (-235 and -772bp, respectively) of HC genes *TRITD2Bv1G013040.1* and *TRITD2Bv1G231900.1* (Supplementary Table S6), which encode NBS-LRR (leucine-rich repeats) disease resistance protein, LRRs and immunoglobulin-like domains protein 3 G, respectively. The LRRs are involved as cellular controllers in different plant processes such as cell division or differentiation (Chakraborty et al., 2019), as well as in stress (Torii, 2004; Dufayard et al., 2017) and defense (Lee and Yeom, 2015) responses. A group of 9 markers composed of *SNP33554, SNP8198* and *SNP33554* (2A), *SNP20210, SNP76785* and *SNP76832* (3B), *SNP13219* (4A), *SNP44041* and *SNP73562* (7B) were significantly linked to different HSIs, and some of them also with yield (Supplementary Table S4). All were related in terms of greater or lesser proximity (within the window of ±50kbp) to genes encoding F-box proteins (Supplementary Table S6), which is one of the largest protein families in plants. F-box proteins can participate as positive regulators in plant responses to stress, such as drought conditions, and also influence plant immunity and hormone signaling (Abd-Hamid et al., 2020).

The marker *SNP15681* (3A), linked to several HSIs (Supplementary Table S4), was found in a proximal region (-2662bp) to the HC gene *TRITD3Av1G246000*, which encodes a disease resistance protein responsible for plant immune responses (Belkhadir et al., 2004). *SNP13388* (2B) was associated with different HSIs and yield (Supplementary Table S4), and was found in the proximity (-6,765bp) of the HC gene *TRITD2Bv1G222900.1*, which encodes the enzyme glycosyltransferase G. This enzyme is important in plants due to its involvement in photosynthetic processes during the transformation of photosynthesis products into disaccharides, oligosaccharides and polysaccharides (Keegstra and Raikhel, 2001). Moreover, some glycosyltransferases have been described as being involved in the cell wall polysaccharide synthesis of grain wheat endosperm (Suliman et al., 2013). *SNP17449*, (3B), associated with the carotenoid index CAR, was also found in the surroundings (-3470bp) of gene *TRITD7Av1G013810.1*, which also encodes a glycosyltransferase enzyme (Supplementary Table S6). None of these genes for glycosyltransferases showed differences in their expression under the different stress conditions assessed (Supplementary Figure S3). Finally, *SNP8165* (2A), associated with the MCARI chlorophyll index, was related to gene *TRITD2Av1G258530.1* (-30461bp), which encodes a MYB-related transcription factor, described by Zhao et al. (2018). This enzyme is involved in a plant's stress responses and increases its expression under PEG6 stress treatment, also showing a slight increase in its expression under T_C and T_S conditions (tetrad stage irrigated developing spike phenotype and tetrad stage drought-stressed developing spike phenotype, respectively) (Supplementary Figure S3).

Among the candidate HC genes results obtained using the bread wheat reference genome, the genes found within a ±50kbp window of three SNP markers (Supplementary Table S4) were of special interest. *SNP2648* (2D), mapped in durum wheat chromosome 2A and associated with the carotenoid indices CRI550 and CRI550m (both carotenoid indices), was found in the proximity of the HC genes *TraesCS2D03G0083900.1* and *TraesCS2D03G0084000.1*, both of which encode a flower-promoting factor 1-like protein 1. This protein regulates plant flowering, and is also involved in the gibberellin signaling pathway (Kania et al., 1997). The flowering locus has also been previously associated with seed dormancy processes (Chen et al., 2014; Chen and Penfield, 2018), germination (Chiang et al., 2009) and water use efficiency (McKay et al., 2003; Mohammadin et al., 2017),

among other plant processes. The orthologs in durum wheat for these genes were *TRITD2Av1G010590* and *TRITD2Bv1G013770*, both of which are located in durum wheat chromosome 2A and encode flowering-promoting factor 1-like proteins 1. Marker *SNP28567* (5B), linked to several HSIs related to photosynthetic processes (Supplementary Table S4), was found in the proximity (-5343 and -1454bp, respectively) of two HC genes, *TraesCS5B03G0023300.1* and *TraesCS5B03G0023400.1*, both of which encode an ERD (Early-responsive to dehydration stress) family protein. These ERD genes have been described as those with a rapid activation during drought stress conditions (Alves et al., 2011). The expression of the first gene slightly decreases with increasing stress levels under both field and PEG conditions. However, interestingly, this gene exhibits higher expression levels under PEG treatment compared to stress conditions in the field. *TraesCS5B03G0023400.1*, slightly increases its expression with increased PEG treatment (Supplementary Figure S4). The orthologs in durum wheat were *TRITD5Bv1G003930* and *TRITD5Av1G004810* (mapped on 5B and 5A, respectively), both of which encode an ERD family protein. *SNP34891* and *SNP34892* (both mapped on 6B), associated with several HSIs (Supplementary Table S4), were found in proximity to 6 HC genes (*TraesCS6B03G0102400.1, TraesCS6B03G0102500.1, TraesCS6B03G0102700.1, TraesCS6B03G0102800.1, TraesCS6B03G0103000.1* and *TraesCS6B03G0103100.1*) (Supplementary Table S6), all of which encode high affinity nitrate transporters, which, as their name suggests, play a key role in nitrate uptake (Crawford and Glass, 1998), as well as in nitrate transport and use, and stress resistance (Du et al., 2022). The ortholog genes in durum wheat were *TRITD6Av1G006050*, *TRITD6Av1G006030*, *TRITD6Bv1G008700* and *TRITD6Av1G006000* (mapped on 6A and 6B), all of which encode high affinity nitrate transporters.

Conclusions

The use of hyperspectral imagery as a high-throughput phenotypic tool to obtain vegetation indices, and their co-localization with final crop yield in GWAS analysis, opens up the possibility of using the HSIs to complement or replace certain field measurements in breeding programs, and of their use as estimators of final production. The GWAS results reported here showed marker-trait associations for final crop yield and HSIs related to photosynthesis processes and structural properties. These results contribute to a better understanding of the dissection of the HSIs assessed, which is directly or indirectly related to final yield or critical physiological processes in durum wheat. Candidate genes analysis revealed a number of gene models across all durum wheat chromosomes, among which we can highlight those related to photosynthetic processes and plant stress responses. The MTAs and candidate genes reported in this study could be of use in breeding programs focused on the use of HTP for driving yield improvements by selecting suitable genotypes. These results support the use of hyperspectral remote sensing imagery in the context of wheat breeding. Further research is needed to advance in our understanding of biophysical modelling to develop spectral plant traits specific to heat and drought resilience.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

RM-G: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis. SG: Writing – review & editing, Visualization, Methodology, Formal analysis, Data curation. IS: Writing – review & editing, Methodology. FM-M: Writing – review & editing, Methodology. CC: Writing – review & editing, Methodology, Formal analysis. JS: Writing – review & editing, Methodology, Formal analysis. CS: Writing – review & editing, Methodology, Formal analysis. KA: Writing – review & editing, Methodology. AB: Writing – review & editing, Formal analysis. VG-D: Writing – review & editing, Formal analysis. PZ-T: Writing – review & editing, Supervision, Funding acquisition, Formal analysis, Conceptualization. PH: Writing – review & editing, Writing – original draft, Supervision, Project administration, Funding acquisition, Conceptualization.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by Junta de Andalucia, projects P18-RT-992 (co-funded by FEDER) and Qualifica Project (QUAL21_023 IAS).

Acknowledgments

The help given by Rafael Romero, Alberto Hornero and Jesus Guillén with the image and data processing is gratefully acknowledged. JS is a Serra-Hunter fellow funded by the Generalitat de Catalunya.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2024.1470520/full#supplementary-material

SUPPLEMENTARY FIGURE S1
Physical position of the (57) associated SNP markers found in GWAS analysis.

SUPPLEMENTARY FIGURE S2
Manhattan and quantile-quantile plots for the 19 traits assessed in GWAS analysis.

SUPPLEMENTARY FIGURE S3
Heatmap for gene expression analysis under several stress conditions for candidate genes. IF: irrigated field conditions; MS: mild stress conditions; SS: severe stress conditions (Gálvez et al., 2019*); IS: seedling PEG shock control; PEG1: seedling 1 h PEG stress; PEG6: seedling 6 h PEG stress (Liu et al., 2015**); AD_S: anther stage irrigated shelter phenotype; AD_S: anther stage drought stressed shelter phenotype; T_C: tetra stage irrigated shelter phenotype; and T_S: tetrad stage drought shelter phenotype (Ma et al., 2017***).

References

Abd-Hamid, N. A., Ahmad-Fauzi, M. I., Zainal, Z., and Ismail, I. (2020). Diverse and dynamic roles of F-box proteins in plant biology. *Planta, Review* (Springer Berlin Heidelberg) 251:68. doi: 10.1007/s00425-020-03356-8

Alves, M. S., Fontes, E. P. B., and Fietto, L. G. (2011). EARLY RESPONSIVE to DEHYDRATION 15, a new transcription factor that integrates stress signaling pathways. *Plant Signal. Behav* 6, 1993–1996. doi: 10.4161/psb.6.12.18268

Anuarbek, S., Abugalieva, S., Pecchioni, N., Laidò, G., Maccaferri, M., Tuberosa, R., et al. (2020). Quantitative trait loci for agronomic traits in tetraploid wheat for enhancing grain yield in Kazakhstan environments. *PloS One* 15, 1–21. doi: 10.1371/journal.pone.0234863

Aparicio, N., Villegas, D., Araus, J. L., Casadesús, J., and Royo, C. (2002). Relationship between growth traits and spectral vegetation indices in durum wheat. *Crop Sci* 42, 1547–1555. doi: 10.2135/cropsci2002.1547

Araus, J. L., and Cairns, J. E. (2014). Field high-throughput phenotyping: The new crop breeding frontier. *Trends Plant Sci* 19, 52–61. doi: 10.1016/j.tplants.2013.09.008

Araus, J. L., Slafer, G. A., Reynolds, M. P., and Royo, C. (2002). Plant breeding and drought in C3 cereals: What should we breed for? *Ann. Bot* 89, 925–940. doi: 10.1093/aob/mcf049

Arif, M. A. R., Attaria, F., Shokat, S., Akram, S., Waheed, M. Q., Arif, A., et al. (2020). Mapping of QTLs associated with yield and yield related traits in durum wheat (*Triticum durum* desf.) under irrigated and drought conditions. *Int. J. Mol. Sci* 21, 7. doi: 10.3390/ijms21072372

Arriagada, O., Gadaleta, A., Marcotuli, I., Maccaferri, M., Campana, M., Reveco, S., et al. (2022). A comprehensive meta-QTL analysis for yield-related traits of durum wheat (*Triticum turgidum* L. var. *durum*) grown under different water regimes. *Front. Plant Sci* 13. doi: 10.3389/fpls.2022.984269

Arriagada, O., Marcotuli, I., Gadaleta, A., and Schwember, A. R. (2020). Molecular mapping and genomics of grain yield in durum wheat: A review. *Int. J. Mol. Sci* 21, 1–19. doi: 10.3390/ijms21197021

Babar, M. A., Reynolds, M. P., Van Ginkel, M., Klatt, A. R., Raun, W. R., and Stone, M. L. (2006). Spectral reflectance to estimate genetic variation for in-season biomass, leaf chlorophyll, and canopy temperature in wheat. *Crop Sci* 46, 1046–1057. doi: 10.2135/cropsci2005.0211

Barakat, M., Al-Doss, A., El-Hendawy, S., Al-Suhaibani, N., Abdella, K., and Al-Ashkar, I. (2021). Deciphering novel QTL for spectral reflectance indices in spring wheat. *Cereal Res. Commun* 49. doi: 10.1007/s42976-021-00131-7

Barakat, M., El-Hendawy, S., Al-Suhaibani, N., Elshafei, A., Al-Doss, A., Al-Ashkar, I., et al. (2016). The genetic basis of spectral reflectance indices in drought-stressed wheat. *Acta Physiol. Plant* 38, 227. doi: 10.1007/s11738-016-2249-9

Barnes, J. D., Balaguer, L., Manrique, E., Elvira, S., and Davison, A. W. (1992). A reappraisal of the use of DMSO for the extraction and determination of chlorophylls a and b in lichens and higher plants. *Environ. Exp. Bot* 32, 85–100. doi: 10.1016/0098-8472(92)90034-Y

Basnyat, P., McConkey, B., Lafond, G. P., Moulin, A., and Pelcat, Y. (2004). Optimal time for remote sensing to relate to crop grain yield on the Canadian prairies. *Can. J. Plant Sci* 84, 97–103. doi: 10.4141/p03-070

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988). The new S language: A programming environment for data analysis and graphics. *Wadsworth Brooks/Cole Pacific Grove CA U.S.A*. doi: 10.1201/9781351074988

Belkhadir, Y., Subramaniam, R., and Dangl, J. L. (2004). Plant disease resistance protein signaling: NBS-LRR proteins and their partners. *Curr. Opin. Plant Biol* 7, 391–399. doi: 10.1016/j.pbi.2004.05.009

Belozersky, M. A., Sarbakanova, S. T., and Dunaevsky, Y. E. (1989). Aspartic proteinase from wheat seeds: isolation, properties and action on gliadin. *Planta* 177, 321–326. doi: 10.1007/BF00403589

Bentley, A. R., Donovan, J., Sonder, K., Baudron, F., Lewis, J. M., Voss, R., et al. (2022).Near- to long-term measures to stabilize global wheat supplies and food security. *Nat Food* 3, 483–486. doi: 10.1038/s43016-022-00559-y

Berger, B., Parent, B., and Tester, M. (2010). High-throughput shoot imaging to study drought responses. *J. Exp. Bot* 61, 3519–3528. doi: 10.1093/jxb/erq201

Bhatta, M., Morgounov, A., Belamkar, V., and Baenziger, P. S. (2018). Genome-wide association study reveals novel genomic regions for grain yield and yield-related traits in drought-stressed synthetic hexaploid wheat. *Int. J. Mol. Sci* 19, 10. doi: 10.3390/ijms19103011

Blum, A. (2011). *Plant breeding for water-limited environments*. Springer New York. doi: 10.1007/978-1-4419-7491-4

Bort, J., Casadesus, J., Nachit, M. M., and Araus, J. L. (2005). Factors affecting the grain yield predicting attributes of spectral reflectance indices in durum wheat: growing conditions, genotype variability and date of measurement. *Int. J. Remote Sens* 26, 2337–2358. doi: 10.1080/01431160512331337808

Bowman, B. C., Chen, J., Zhang, J., Wheeler, J., Wang, Y., Zhao, W., et al. (2015). Evaluating grain yield in spring wheat with canopy spectral reflectance. *Crop Sci* 55, 1881–1890. doi: 10.2135/cropsci2014.08.0533

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635. doi: 10.1093/bioinformatics/btm308

Breseghello, F., and Sorrells, M. E. (2006). Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. *Genetics* 172, 1165–1177. doi: 10.1534/genetics.105.044586

Cabrera-Bosquet, L., Crossa, J., von Zitzewitz, J., Serret, M. D., and Araus, J. L. (2012). High-throughput phenotyping and genomic selection: the frontiers of crop breeding converge. *J. Integr. Plant Biol* 54, 312–320. doi: 10.1111/j.1744-7909.2012.01116.x

Camino, C., Calderón, R., Parnell, S., Dierkes, H., Chemin, Y., Román-Écija, M., et al. (2021). Detection of *Xylella fastidiosa* in almond orchards by synergic use of an epidemic spread model and remotely sensed plant traits. *Remote Sens. Environ* 260. doi: 10.1016/j.rse.2021.112420

Camino, C., Gonzalez-Dugo, V., Hernandez, P., and Zarco-Tejada, P. J. (2019). Radiative transfer Vcmax estimation from hyperspectral imagery and SIF retrievals to assess photosynthetic performance in rainfed and irrigated plant phenotyping trials. *Remote Sens. Environ* 231, 111186. doi: 10.1016/j.rse.2019.05.005

Casstevens, T., and Wang, Y. (2015). *First annual tassel hackathon*.

Chakraborty, S., Nguyen, B., Wasti, S. D., and Xu, G. (2019). Plant leucine-rich repeat receptor kinase (LRR-RK): Structure, ligand perception, and activation mechanism. *Molecules* 24, 17. doi: 10.3390/molecules24173081

Chandel, N. S., Tiwari, P. S., Singh, K. P., Jat, D., Gaikwad, B. B., Tripathi, H., et al. (2019). Yield prediction in wheat (*Triticum aestivum* L.) using spectral reflectance indices. *Curr. Sci* 116, 272–278. doi: 10.18520/cs/v116/i2/272-278

Chen, M., MacGregor, D. R., Dave, A., Florance, H., Moore, K., Paszkiewicz, K., et al. (2014). Maternal temperature history activates Flowering Locus T in fruits to control progeny dormancy according to time of year. *Proc. Natl. Acad. Sci. U. S. A* 111, 18787–18792. doi: 10.1073/pnas.1412274111

Chen, M., and Penfield, S. (2018). Feedback regulation of COOLAIR expression controls seed dormancy and flowering time. *Plant Sci. Res*. 12, 1014–1017. doi: 10.1126/science.aar7361

Chiang, G. C. K., Barua, D., Kramera, E. M., Amasino, R. M., and Donohue, K. (2009). Major flowering time gene, *FLOWERING LOCUS C*, regulates seed germination in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A* 106, 11661–11666. doi: 10.1073/pnas.0901367106

Chu, H. A., and Chiu, Y. F. (2016). The roles of cytochrome b559 in assembly and photoprotection of photosystem II revealed by site-directed mutagenesis studies. *Front. Plant Sci* 6. doi: 10.3389/fpls.2015.01261

Condorelli, G. E., Maccaferri, M., Newcomb, M., Andrade-sanchez, P., White, J. W., French, A. N., et al. (2018). Comparative aerial and ground based high throughput phenotyping for the genetic dissection of NDVI as a proxy for drought adaptive traits in durum wheat. *Front. Plant Sci* 9. doi: 10.3389/fpls.2018.00893

Crain, J., Mondal, S., Rutkoski, J., Singh, R. P., and Poland, J. (2018). Combining high-throughput phenotyping and genomic information to increase prediction and selection accuracy in wheat breeding. *Plant Genome* 11, 170043. doi: 10.3835/plantgenome2017.05.0043

Cramer, W. A. (2020). Structure-function of the cytochrome b6f lipoprotein complex: a scientific odyssey and personal perspective. *Photosynth. Res.* 139, 1-3, 53–65. doi: 10.1007/s11120-018-0585-x.Structure-Function

Crawford, N. M., and Glass, A. D. M. (1998). Molecular and physiological aspects of nitrate uptake in plants. *1Trends Plant Sci* 3, 389–395. doi: 10.1016/S1360-1385(98)01311-9

Datt, B. (1998). Remote sensing of chlorophyll a, chlorophyll b, chlorophyll a+b, and total carotenoid content in eucalyptus leaves. *Remote Sens. Environ* 66, 111–121. doi: 10.1016/S0034-4257(98)00046-7

Du, R. J., Wu, Z. X., Yu, Z. X., Li, P. F., Mu, J. Y., Zhou, J., et al. (2022). Genome-wide characterization of high-affinity nitrate transporter 2 (NRT2) gene family in *brassica napus*. *Int. J. Mol. Sci* 23, 9. doi: 10.3390/ijms23094965

Dufayard, J. F., Bettembourg, M., Fischer, I., Droc, G., Guiderdoni, E., Périn, C., et al. (2017). New insights on Leucine-Rich repeats receptor-like kinase orthologous relationships in angiosperms. *Front. Plant Sci* 8. doi: 10.3389/fpls.2017.00381

FAOSTAT (2023). Available online at: http://www.fao.org/faostat/ (Accessed 27th October 2023).

Farouk, I., Alsaleh, A., Motowaj, J., Gaboun, F., Belkadi, B., Filali Maltouf, A., et al. (2021). Detection of grain yield QTLs in the durum population Lahn/Cham1 tested in contrasting environments. *Turkish J. Biol* 45, 65–78. doi: 10.3906/biy-2008-41

Fromme, P., Jordan, P., and Krauß, N. (2001). Structure of photosystem I. *Biochim. Biophys. Acta - Bioenerg* 1507, 5–31. doi: 10.1016/S0005-2728(01)00195-5

Gálvez, S., Mérida-García, R., Camino, C., Borrill, P., Abrouk, M., Ramírez-González, R. H., et al. (2019). Hotspots in the genomic architecture of field drought responses in wheat as breeding targets. *Funct. Integr. Genomics* 19, 295–309. doi: 10.1007/s10142-018-0639-3

Gamon, J. A., Peñuelas, J., and Field, C. B. (1992). A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency. *Remote Sens. Env*. doi: 10.1016/0034-4257(92)90059-S

Gao, J., Wang, H., Yuan, Q., and Feng, Y. (2018). Structure and function of the photosystem supercomplexes. *Front. Plant Sci* 9. doi: 10.3389/fpls.2018.00357

Gaujoux, R., and Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. *BMC Inf* 11, 367. doi: 10.1186/1471-2105-11-367

Gitelson, A. A., Gritz, Y., and Merzlyak, M. N. (2003). Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *J. Plant Physiol* 160, 271–282. doi: 10.1078/0176-1617-00887

Gitelson, A. A., Keydan, G. P., and Merzlyak, M. N. (2006). Three-band model for noninvasive estimation of chlorophyll, carotenoids, and anthocyanin contents in higher plant leaves. *Geophys. Res. Lett* 33, 2–6. doi: 10.1029/2006GL026457

Gizaw, S. A., Garland-Campbell, K., and Carter, A. H. (2016). Evaluation of agronomic traits and spectral reflectance in Pacific Northwest winter wheat under rain-fed and irrigated conditions. *F. Crop Res* 196, 168–179. doi: 10.1016/j.fcr.2016.06.018

Gizaw, S. A., Godoy, J. G. V., Pumphrey, M. O., and Carter, A. H. (2018). Spectral reflectance for indirect selection and genome-wide association analyses of grain yield and drought tolerance in North American spring wheat. *Crop Sci* 58, 2289–2301. doi: 10.2135/cropsci2017.11.0690

Govta, N., Polda, I., Sela, H., Cohen, Y., Beckles, D. M., Korol, A. B., et al. (2022). Genome-wide association study in bread wheat identifies genomic regions associated with grain yield and quality under contrasting water availability. *Int. J. Mol. Sci* 23, 18. doi: 10.3390/ijms231810575

Haboudane, D., Miller, J. R., Pattey, E., Zarco-Tejada, P. J., and Strachan, I. B. (2004). Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture. *Remote Sens. Environ* 90, 337–352. doi: 10.1016/j.rse.2003.12.013

Haboudane, D., Miller, J. R., Tremblay, N., Zarco-Tejada, P. J., and Dextraze, L. (2002). Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. *Remote Sens. Environ* 81, 416–426. doi: 10.1016/S0034-4257(02)00018-4

Hassan, M. A., Yang, M., Rasheed, A., Yang, G., Reynolds, M., Xia, X., et al. (2019). A rapid monitoring of NDVI across the wheat growth cycle for grain yield prediction using a multi-spectral UAV platform. *Plant Sci* 282, 95–103. doi: 10.1016/j.plantsci.2018.10.022

Hernández-Clemente, R., Navarro-Cerrillo, R. M., and Zarco-Tejada, P. J. (2012). Carotenoid content estimation in a heterogeneous conifer forest using narrow-band indices and PROSPECT+DART simulations. *Remote Sens. Environ* 127, 298–315. doi: 10.1016/j.rse.2012.09.014

Horikoshi, M., and Tang, Y. (2016). ggfortify: data visualization tools for statistical analysis results. *R. J* 8, 474–489. doi: 10.32614/RJ-2016-060

Hu, C. H., Wei, X. Y., Yuan, B., Yao, L. B., Ma, T. T., Zhang, P. P., et al. (2018). Genome-wide identification and functional analysis of NADPH oxidase family genes in wheat during development and environmental stress responses. *Front. Plant Sci* 9. doi: 10.3389/fpls.2018.00906

Hussain, B., Akpınar, B. A., Alaux, M., Algharib, A. M., Sehgal, D., Ali, Z., et al. (2022). Capturing wheat phenotypes at the genome level. *Front. Plant Sci* 13. doi: 10.3389/fpls.2022.851079

IPPC report (2023). Available online at: http://www.ipcc.ch/ (Accessed 27th October 2023).

IWGSC (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361, eaar7191. doi: 10.1126/science.aar7191

Ji, L., and Peters, A. J. (2023). Assessing vegetation response to drought in the northern Great Plains using vegetation and drought indices. *Remote Sens. Env* 87, 85–98. doi: 10.1016/S0034-4257(03)00174-3

Jiang, L., Sun, L., Ye, M., Wang, J., Wang, Y., Bogard, M., et al. (2019). Functional mapping of N deficiency-induced response in wheat yield-component traits by implementing high-throughput phenotyping. *Plant J* 97, 1105–1119. doi: 10.1111/tpj.14186

Jin, X., Zarco-Tejada, P. J., Schmidhalter, U., Reynolds, M. P., Hawkesford, M. J., Varshney, R. K., et al. (2021). High-throughput estimation of crop traits: A review of ground and aerial phenotyping platforms. *IEEE Geosci. Remote Sens. Mag* 9, 200–231. doi: 10.1109/MGRS.2020.2998816

Jun, X. U., Xin-yu, W., and Wang-Zhen, G. U. O. (2015). The cytochrome P450 superfamily: Key players in plant development and defense. *J. Integr. Agric* 14, 1673–1686. doi: 10.1016/S2095-3119(14)60980-1

Kaan, F., Branlard, G., Chihab, B., and Borries, C. (1993). Relations between genes coding for grain storage protein and two pasta cooking quality criteria among world durum wheat (*Triticum durum* Desf.) genetic resources. *J. Genet. Breed* 47, 151–156.

Kang, J., Park, J., Choi, H., Burla, B., Kretzschmar, T., Lee, Y., et al. (2011). Plant ABC transporters. *Arab. B* 9, e0153. doi: 10.1199/tab.0153

Keegstra, K., and Raikhel, N. (2001). Plant glycosyltransferases. *Curr. Opin. Plant Biol* 4, 219–224. doi: 10.1016/S1369-5266(00)00164-3

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika* 30, 81–93. doi: 10.1093/biomet/30.1-2.81

Kendall, M. G. (1945). The treatment of ties in rank problems. *Biometrika* 33, 239–251. doi: 10.1093/biomet/33.3.239

Krishnappa, G., Khan, H., Krishna, H., Devate, N. B., Kumar, S., Singh, G. P., et al. (2023). Genome-wide association study for grain protein, thousand kernel weight, and normalized difference vegetation index in bread wheat (*Triticum aestivum* L.). *Genes (Basel)* 14, 637. doi: 10.3390/genes14030637

Kronenberg, Z. N., Rhie, A., Koren, S., Concepcion, G. T., Peluso, P., Munson, K. M., et al. (2021). Extended haplotype-phasing of long-read *de novo* genome assemblies using Hi-C. *Nat. Commun* 12, 1–10. doi: 10.1038/s41467-020-20536-y

Labus, M. P., Nielsen, G. A., Lawrence, R. I., Engel, R., and Long, D. S. (2010). Wheat yield estimates using multi-temporal NDVI satellite imagery. *Int. J. Remote Sens* 23, 4169–4180. doi: 10.1080/01431160110107653

Lee, H. A., and Yeom, S. I. (2015). Plant NB-LRR proteins: Tightly regulated sensors in a complex manner. *Brief. Funct. Genomics* 14, 233–242. doi: 10.1093/bfgp/elv012

Leegood, R. C., Evans, J. R., and Furbank, R. T. (2010). Food security requires genetic advances to increase farm yields. *Nature* 464, 891. doi: 10.1038/464831d

Leghari, S. J., Wahocho, N. A., Laghari, G. M., and Hafeez Laghari, A. (2016). Role of nitrogen for plant growth and development: A review. *Adv. Environ. Biol* 10, 209–218.

Li, X. M., Chen, X. M., Xiao, Y. G., Xia, X. C., Wang, D., He, Z. H., et al. (2014). Identification of QTLs for seedling vigor in winter wheat. *Euphytica* 198, 199–209. doi: 10.1007/s10681-014-1092-6

Li, L., Lin, D., Wang, J., Yang, L., and Wang, Y. (2020). Multivariate analysis models based on full spectra range and effective wavelengths using different transformation techniques for rapid estimation of leaf nitrogen concentration in winter wheat. *Front. Plant Sci* 11. doi: 10.3389/fpls.2020.00755

Li, D. M., Wang, Y., and Han, K. L. (2012). Recent density functional theory model calculations of drug metabolism by cytochrome P450. *Coord. Chem. Rev* 256, 1137–1150. doi: 10.1016/j.ccr.2012.01.016

Liu, W., Maccaferri, M., Rynearson, S., Letta, T., Zegeye, H., Tuberosa, R., et al. (2017). Novel sources of stripe rust resistance identified by genome-wide association mapping in Ethiopian durum wheat (*Triticum turgidum* ssp. *durum*). *Front. Plant Sci* 8. doi: 10.3389/fpls.2017.00774

Liu, C., Pinto, F., Cossani, C. M., Sukumaran, S., and Reynolds, M. P. (2019). Spectral reflectance indices as proxies for yield potential and heat stress tolerance in spring wheat: Heritability estimates and marker-trait associations. *Front. Agric. Sci. Eng* 6, 296–308. doi: 10.15302/J-FASE-2019269

Liu, Z., Xin, M., Qin, J., Peng, H., Ni, Z., Yao, Y., et al. (2015). Temporal transcriptome profiling reveals expression partitioning of homeologous genes contributing to heat and drought acclimation in wheat (*Triticum aestivum* L.). *BMC Plant Biol* 15, 1–20. doi: 10.1186/s12870-015-0511-8

Lowe, A., Harrison, N., and French, A. P. (2017). Hyperspectral image analysis techniques for the detection and classification of the early onset of plant disease and stress. *Plant Methods* 13, 1–12. doi: 10.1186/s13007-017-0233-z

Lozada, D. N., Godoy, J. V., Ward, B. P., and Carter, A. H. (2020). Genomic prediction and indirect selection for grain yield in US pacific northwest winter wheat using spectral reflectance indices from high-throughput phenotyping. *Int. J. Mol. Sci* 21, 1. doi: 10.3390/ijms21010165

Ma, J., Li, R., Wang, H., Li, D., Wang, X., Zhang, Y., et al. (2017). Transcriptomics analyses reveal wheat responses to drought stress during reproductive stages under field conditions. *Front. Plant Sci* 8. doi: 10.3389/fpls.2017.00592

Maccaferri, M., Cane, M. A., Sanguineti, M. C., Salvi, S., Colalongo, M. C., Massi, A., et al. (2014). A consensus framework map of durum wheat (*Triticum durum* Desf.) suitable for linkage disequilibrium analysis and genome-wide association mapping. *BMC Genomics* 15, 1–21. doi: 10.1186/1471-2164-15-873

Maccaferri, M., Harris, N. S., Twardziok, S. O., Pasam, R. K., Gundlach, H., Spannagl, M., et al. (2019). Durum wheat genome highlights past domestication signatures and future improvement targets. *Nat Genet* 51, 885–895.

Mahlein, A. K., Rumpf, T., Welke, P., Dehne, H. W., Plümer, L., Steiner, U., et al. (2013). Development of spectral indices for detecting and identifying plant diseases. *Remote Sens. Environ* 128, 21–30. doi: 10.1016/j.rse.2012.09.019

Mangini, G., Blanco, A., Nigro, D., Signorile, M. A., and Simeone, R. (2021). Candidate genes and quantitative trait loci for grain yield and seed size in durum wheat. *Plants* 10, 1–21. doi: 10.3390/plants10020312

McKay, J. K., Richards, J. H., and Mitchell-Olds, T. (2003). Genetics of drought adaptation in Arabidopsis thaliana: I. Pleiotropy contributes to genetic correlations among ecological traits. *Mol. Ecol* 12, 1137–1151. doi: 10.1046/j.1365-294X.2003.01833.x

McMullen, M. D., Kresovich, S., Sanchez Villeda, H., Bradbury, P., Li, H., Sun, Q., et al. (2009). Report genetic properties of the maize nested association mapping population. *Science New Ser* 325, 737–740. doi: 10.1126/science.1174320

Mérida-García, R., Bentley, A. R., Gálvez, S., Dorado, G., Solís, I., Ammar, K., et al. (2020). Mapping agronomic and quality traits in elite durum wheat lines under differing water regimes. *Agronomy* 10, 1–23. doi: 10.3390/agronomy10010144

Mérida-García, R., Liu, G., He, S., Gonzalez-Dugo, V., Dorado, G., Gálvez, S., et al. (2019). Genetic dissection of agronomic and quality traits based on association mapping and genomic selection approaches in durum wheat grown in Southern Spain. *PloS One* 14, 1–24. doi: 10.1371/journal.pone.0211718

Mi, J., Vallarino, J. G., Petřík, I., Novák, O., Correa, S. M., Chodasiewicz, M., et al. (2022). A manipulation of carotenoid metabolism influence biomass partitioning and fitness in tomato. *Metab. Eng* 70, 166–180. doi: 10.1016/j.ymben.2022.01.004

Mohammadin, S., Nguyen, T. P., Van Weij, M. S., Reichelt, M., and Schranz, M. E. (2017). Flowering locus C (FLC) is a potential major regulator of glucosinolate content across developmental stages of *Aethionema arabicum* (Brassicaceae). *Front. Plant Sci* 8. doi: 10.3389/fpls.2017.00876

Morisse, M., Wells, D. M., Millet, E. J., Lillemo, M., Fahrner, S., Cellini, F., et al. (2022). A European perspective on opportunities and demands for field-based crop phenotyping. *Field Crops Research* 276, 108371. doi: 10.1016/j.fcr.2021.108371

Moya, I., Camenen, L., Evain, S., Goulas, Y., Cerovic, Z. G., Latouche, G., et al. (2004). A new instrument for passive remote sensing: 1. Measurements of sunlight-induced chlorophyll fluorescence. *Remote Sens. Environ* 91, 186–197. doi: 10.1016/j.rse.2004.02.012

Mulugeta, B., Tesfaye, K., Ortiz, R., Johansson, E., Hailesilassie, T., Hammenhag, C., et al. (2023). Marker-trait association analyses revealed major novel QTLs for grain yield and related traits in durum wheat. *Front. Plant Sci* 13. doi: 10.3389/fpls.2022.1009244

Nigro, D., Gadaleta, A., Mangini, G., Colasuonno, P., Marcotuli, I., Giancaspro, A., et al. (2019). Candidate genes and genome-wide association study of grain protein content and protein deviation in durum wheat. *Planta* 249, 1157–1175. doi: 10.1007/s00425-018-03075-1

Ohkawa, H., Imaishi, H., Shiota, N., Yamada, T., Inui, H., and Ohkawa, Y. (1998). Molecular mechanisms of herbicide resistance with special emphasis on cytochrome P450 monooxygenases. *Plant Biotechnol* 15, 173–176. doi: 10.5511/plantbiotechnology.15.173

Panek, E., and Gozdowski, D. (2020). Analysis of relationship between cereal yield and NDVI for selected regions of Central Europe based on MODIS satellite data. *Remote Sens. Appl. Soc Environ* 17, 100286. doi: 10.1016/j.rsase.2019.100286

Paul, M. J. (2021). Improving photosynthetic metabolism for crop yields: what is going to work? *Front. Plant Sci* 12. doi: 10.3389/fpls.2021.743862

Plascyk, J. A., and Gabriel, F. C. (1975). The fraunhofer line discriminator MKII-an airborne instrument for precise and standardized ecological luminescence measurement. *IEEE Trans. Instrum. Meas* 24, 306–313. doi: 10.1109/TIM.1975.4314448

Rasheed, A., Xia, X., Ogbonnaya, F., Mahmood, T., Zhang, Z., Mujeeb-Kazi, A., et al. (2014). Genome-wide association for grain morphology in synthetic hexaploid wheats using digital imaging analysis. *BMC Plant Biol* 14, 1–21. doi: 10.1186/1471-2229-14-128

R Core Team (2020). *R: A language and environment for statistical computing* (Vienna, Austria: R Foundation for Statistical Computing). Available at: https://www.R-project.org/.

Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J., et al. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl. Acad. Sci* 98, 11479–11484. doi: 10.1073/pnas.201394398

Reynolds, M., Dreccer, F., and Trethowan, R. (2007). Drought-adaptive traits derived from wheat wild relatives and landraces. *J. Exp. Bot* 58, 177–186. doi: 10.1093/jxb/erl250

Reynolds, J. F., Kemp, P. R., Ogle, K., and Fernández, R. J. (2004). Modifying the "pulse-reserve" paradigm for deserts of North America: Precipitation pulses, soil water, and plant responses. *Oecologia* 141, 194–210. doi: 10.1007/s00442-004-1524-4

RIA (2023). Available online at: https://www.juntadeandalucia.es/agriculturaypesca/ifapa/riaweb/web/estaciones (Accessed February, 25, 2023).

Roth, L., Kronenberg, L., Aasen, H., Walter, A., Hartung, J., van Eeuwijk, F., et al. (2024). High-throughput field phenotyping reveals that selection in breeding has affected the phenology and temperature response of wheat in the stem elongation phase. *J. Exp. Bot* 75, 2084–2099. doi: 10.1093/jxb/erad481

Rouse, J. W., Hass, R. H., Schell, J. A., and Deering, D. W. (1973). Monitoring vegetation systems in the great plains with ERTS. *Third Earth Resour. Technol. Satell. Symp* 1, 309–317. doi: 10.1126/science.1174320

Royo, C., Ammar, K., Villegas, D., and Soriano, J. M. (2021). Agronomic, physiological and genetic changes associated with evolution, migration and modern breeding in durum wheat. *Front. Plant Sci* 12. doi: 10.3389/fpls.2021.674470

Rufo, R., López, A., Lopes, M. S., Bellvert, J., and Soriano, J. M. (2021a). Identification of QTL hotspots affecting agronomic traits and high-throughput vegetation indices in rainfed wheat. *Front. Plant Sci.* 12, 735192. doi: 10.1101/2021.06.25.449881

Rufo, R., López, A., Lopes, M. S., Bellvert, J., and Soriano, J. M. (2021b). Identification of quantitative trait loci hotspots affecting agronomic traits and high-throughput vegetation indices in rainfed wheat. *Front. Plant Sci* 12. doi: 10.3389/fpls.2021.735192

Rufo, R., Soriano, J. M., Villegas, D., Royo, C., and Bellvert, J. (2021c). Using unmanned aerial vehicle and ground-based RGB indices to assess agronomic performance of wheat landraces and cultivars in a mediterranean-type environment. *Remote Sens* 13, 1187. doi: 10.3390/rs13061187

Ruiz, M., and Carrillo, J. M. (1993). Linkage relationships between prolamin genes on chromosomes 1A and 1B of durum wheat. *Theor. Appl. Genet. Int. J. Plant Breed. Res* 87, 353–360. doi: 10.1007/BF01184923

Safdar, L. B., Dugina, K., Saeidan, A., Yoshicawa, G. V., Caporaso, N., Gapare, B., et al. (2023). Reviving grain quality in wheat through non-destructive phenotyping techniques like hyperspectral imaging. *Food Energy Secur* 12, 1–21. doi: 10.1002/fes3.498

Singh, D., Wang, X., Kumar, U., Gao, L., Noor, M., Imtiaz, M., et al. (2019). High-throughput phenotyping enabled genetic dissection of crop lodging in wheat. *Front. Plant Sci* 10. doi: 10.3389/fpls.2019.00394

Soriano, J. M., Malosetti, M., Roselló, M., Sorrells, M. E., and Royo, C. (2017). Dissecting the old Mediterranean durum wheat genetic architecture for phenology, biomass and yield formation by association mapping and QTL meta-analysis. *PloS One* 12, 1–19. doi: 10.1371/journal.pone.0178290

Suliman, M., Chateigner-Boutin, A. L., Francin-Allami, M., Partier, A., Bouchet, B., Salse, J., et al. (2013). Identification of glycosyltransferases involved in cell wall synthesis of wheat endosperm. *J. Proteomics* 78, 508–521. doi: 10.1016/j.jprot.2012.10.021

Sultana, S. R., Ali, A., Ahmad, A., Mubeen, M., Zia-Ul-Haq, M., Ahmad, S., et al. (2014). Normalized difference vegetation index as a tool for wheat yield estimation: A case study from Faisalabad, Pakistan. *Sci. World J.* doi: 10.1155/2014/725326

Tattaris, M., Reynolds, M. P., and Chapman, S. C. (2016). A direct comparison of remote sensing approaches for high-throughput phenotyping in plant breeding. *Front. Plant Sci* 7. doi: 10.3389/fpls.2016.01131

Tenikecier, H. S., and Genctan, T. (2020). Effect of endosperm and seed size on some yield and quality characteristics of wheat (*Triticum aestivum* L. Em thell). *Curr. Trends Nat. Sci* 9, 132–141. doi: 10.47068/ctns.2020.v9i17.015

Torii, K. U. (2004). Leucine-rich repeat receptor kinases in plants: structure, function, and signal transduction pathways. *Int. Rev. Cytol* 234, 1–46. doi: 10.1016/S0074-7696(04)34001-5

Valladares García, A. P., Desiderio, F., Simeone, R., Ravaglia, S., Ciorba, R., Fricano, A., et al. (2023). QTL mapping for kernel-related traits in a durum wheat x T. dicoccum segregating population. *Front. Plant Sci* 14. doi: 10.3389/fpls.2023.1253385

Vannoppen, A., Gobin, A., Kotova, L., Top, S., Cruz, L., Viksna, A., et al. (2020). Wheat yield estimation from NDVI and regional climate models in Latvia. *Remote Sens* 12, 1–20. doi: 10.3390/rs12142206

Wang, P., Hsu, C. C., Du, Y., Zhu, P., Zhao, C., Fu, X., et al. (2020). Mapping proteome-wide targets of protein kinases in plant stress responses. *Proc. Natl. Acad. Sci. U. S. A* 117, 3270–3280. doi: 10.1073/pnas.1919901117

Wei, J., Tang, X., Gu, Q., Wang, M., Ma, M., and Han, X. (2019). Using solar-induced chlorophyll fluorescence observed by OCO-2 to predict autumn crop production in China. *Remote Sens* 11, 1–14. doi: 10.3390/rs11141715

Weir, B. S. (1997). Genetic data analysis II. *Biometrics*. doi: 10.2307/2533134

Xiao, Q., Bai, X., Zhang, C., and He, Y. (2022). Advanced high-throughput plant phenotyping techniques for genome- wide association studies: A review. *J. Adv. Res* 35, 215–230. doi: 10.1016/j.jare.2021.05.002

Xiao, Y. G., Liu, J. J., Xia, X. C., Chen, X. M., Reynolds, M. P., and He, Z. H. (2013). Genetic analysis of early vigour in winter wheat using digital imaging. *Acta Agron. Sin* 39, 1935–1943. doi: 10.3724/SP.J.1006.2013.01935

Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., et al. (2021). rMVP: A memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Genomics Proteomics Bioinforma* 19, 619–628. doi: 10.1016/j.gpb.2020.10.007

Yousfi, S., Márquez, A. J., Betti, M., Araus, J. L., and Serret, M. D. (2016). Gene expression and physiological responses to salinity and water stress of contrasting durum wheat genotypes. *J. Integr. Plant Biol* 58, 48–66. doi: 10.1111/jipb.12359

Yu, R., Cao, X., Liu, J., Nie, R., Zhang, C., Yuan, M., et al. (2024). Using UAV-based temporal spectral indices to dissect changes in the stay-green trait in wheat. *Plant Phenomics* 6, 1–15. doi: 10.34133/plantphenomics.0171

Zadoks, J. C., Chang, T. T., and Konzak, C. F. (1974). A decimal code for the growth stages of cereals. *Weed Res* 14, 415–421. doi: 10.1016/s0262-1762(99)80614-2

Zarco-Tejada, P. J., Berjón, A., López-Lozano, R., Miller, J. R., Martín, P., Cachorro, V., et al. (2005). Assessing vineyard condition with hyperspectral indices: Leaf and canopy reflectance simulation in a row-structured discontinuous canopy. *Remote Sens. Environ* 99, 271–287. doi: 10.1016/j.rse.2005.09.002

Zarco-Tejada, P. J., Camino, C., Beck, P. S. A., Calderon, R., Hornero, A., Hernández-Clemente, R., et al. (2018). Previsual symptoms of *Xylella fastidiosa* infection revealed in spectral plant-trait alterations. *Nat. Plants* 4, 432–439. doi: 10.1038/s41477-018-0189-7

Zarco-Tejada, P. J., González-Dugo, M. V., and Fereres, E. (2016). Seasonal stability of chlorophyll fluorescence quantified from airborne hyperspectral imagery as an indicator of net photosynthesis in the context of precision agriculture. *Remote Sens. Environ* 179, 89–103. doi: 10.1016/j.rse.2016.03.024

Zarco-Tejada, P. J., Miller, J. R., Mohammed, G. H., and Noland, T. L. (2000). Chlorophyll fluorescence effects on vegetation apparent reflectance: I. Leaf-level measurements and model simulation. *Remote Sens. Environ* 74, 582–595. doi: 10.1016/S0034-4257(00)00148-6

Zendonadi dos Santos, N., Piepho, H. P., Condorelli, G. E., Licieri Groli, E., Newcomb, M., Ward, R., et al. (2021). High-throughput field phenotyping reveals genetic variation in photosynthetic traits in durum wheat under drought. *Plant Cell Environ* 44, 2858–2878. doi: 10.1111/pce.14136

Zhang, Z., Qu, Y., Ma, F., Lv, Q., Zhu, X., Guo, G., et al. (2024). Integrating high-throughput phenotyping and genome-wide association studies for enhanced drought resistance and yield prediction in wheat. *New Phytol* 243, 1758–1775. doi: 10.1111/nph.19942

Zhao, Y., Cheng, X., Liu, X., Wu, H., Bi, H., and Xu, H. (2018). The wheat MYB transcription factor taMYB31 is involved in drought stress responses in arabidopsis. *Front. Plant Sci* 9. doi: 10.3389/fpls.2018.01426

Zuhro, A., Tambunan, M. P., and Marko, K. (2020). Application of vegetation health index (VHI)to identify distribution of agricultural drought in Indramayu Regency, West Java Province. *IOP Conf. Ser. Earth Environ. Sci* 500, 1. doi: 10.1088/1755-1315/500/1/012047

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 14 March 2025
DOI 10.3389/fpls.2025.1485710

OPEN ACCESS

EDITED BY
Jinwu Wang,
Northeast Agricultural University, China

REVIEWED BY
Fei Liu,
Inner Mongolia Agricultural University, China
Yulong Chen,
Shandong University of Technology, China

*CORRESPONDENCE
Lijia Xu
wzp9526@163.com

†These authors have contributed equally to this work

RECEIVED 24 August 2024
ACCEPTED 05 February 2025
PUBLISHED 14 March 2025

CITATION
Han D, Li W, Wang Y, Wang Q, Wu Z, Wang Y, Xu Y and Xu L (2025) CFD-DEM coupling analysis of the negative pressure inlet structural parameters on the performance of integrated positive-negative pressure seed-metering device.
Front. Plant Sci. 16:1485710.
doi: 10.3389/fpls.2025.1485710

CFD-DEM coupling analysis of the negative pressure inlet structural parameters on the performance of integrated positive-negative pressure seed-metering device

Dandan Han[1†], Wei Li[1†], Yunxia Wang[2], Qing Wang[1], Zhijun Wu[1], Yuchao Wang[1], You Xu[1] and Lijia Xu[1*]

[1]College of Mechanical and Electrical Engineering, Sichuan Agricultural University, Ya'an, China, [2]Nanjing Institute of Agricultural Mechanization, Ministry of Agriculture and Rural Affairs, Nanjing, China

To assess the influence of the structural parameters of the negative pressure inlet pipe on the seeding performance of the integrated positive-negative pressure seed-metering device, angles α and β, along with taper θ, were selected as test variables for conducting coupling simulation tests and physical bench tests. The results indicate that the average differential pressure across the holes in the seed-filling zone (Δp_{sf}), the airflow rate at the interaction interface between the negative pressure inlet pipe and the negative pressure chamber (Q), the average drag force on seeds in the seed-filling zone ($F_{D,sf}$), and the average drag force on seeds in the seed-cleaning zone ($F_{D,sc}$) are all significantly affected by the test factors. The structural parameters of the negative pressure inlet pipe can be accurately predicted using the prediction model developed through regression analysis of the central composite test results. The optimal parameters combination is established as 17.1° for angle α, 81.3° for angle β, and 0.52° for taper θ. The maximum error between the predicted evaluation index values and those obtained from the coupling simulation verification test with the optimized negative pressure inlet pipe is less than 6.66%, indicating a strong correlation and reasonable prediction of the structural parameters. The results from the physical bench tests confirm that the optimal operational speed for the seed-metering device with the optimized negative pressure inlet pipe is 4 to 5 km/h, with an optimal working negative pressure of 4 to 6 kPa. Under these conditions, the qualified rate varies from 92.27% to 95.28%, the multiple rate from 2.43% to 5.52%, and the leakage rate from 1.22% to 5.2%.

KEYWORDS

maize, seed-metering device, integrated positive-negative pressure, air inlet, CFD-DEM coupling approach

1 Introduction

The soybean-maize banded composite planting technology represents a viable strategy for stabilizing maize yields while simultaneously expanding soybean cultivation, thereby providing a novel technical approach to bolster China's maize production capacity and enhance soybean self-sufficiency rates (Yang and Yang, 2019). A fundamental aspect of this planting pattern involves minimizing the spacing between maize and soybean seeds during sowing (Yang et al., 2022; Zhang et al., 2020). To address the demand for precision sowing in densely planted conditions, the development of compatible precision seed-metering devices is essential (Karayel and Özmerzi, 2001; Vianna et al., 2014; Wang et al., 2021).

The air-suction type serves as a pneumatic seed-metering device, offering numerous advantages such as improved seed spacing accuracy, reduced seed damage, and adaptability to a diverse array of seeds with only minor modifications to the device (Shafii and Holmes, 1990; Guarella et al., 1996; Karayel et al., 2022). Jack et al. (2013) identified a correlation between the total area of the vacuum holes and the operational efficiency of the air-suction seed-metering device, determining that optimal seeding performance occurs within the 400~650 mm^2 range when examining the interplay of rotational speed and pressure. To enhance the seeding efficacy of air-suction seed-metering devices at elevated operating speeds in densely planted scenarios, various researchers have explored the effects of seeding plate design on seed population disturbance and mobility (Wang et al., 2020; Shi et al., 2020; Liu et al., 2022a).

Kostić et al. employed both theoretical and experimental methodologies to investigate the causes of maize seeding failures, revealing that observable variables significantly impact seeding distribution accuracy (Kostić et al., 2018). During the validation of seeding performance, high-speed cameras have proven to be effective tools for replicating the dynamic processes and capturing the movement of seeds within the seed-metering device, surpassing the observational capabilities of the human eye (Pezzuolo et al., 2018; Tang et al., 2023; Zhao et al., 2024). The operation of the air-suction seed-metering device involves continuous collisions and friction among seeds and between seeds and the device, resulting in highly complex mechanical behaviors during seed movement that are challenging to analyze as previously described (Lai et al., 2016; Pareek et al., 2023).

To further improve seeding performance, a positive pressure inlet pipe was integrated into the seed-casting zone of the air-suction seed-metering device designed in the previous stage (Han et al., 2024). This modification aimed to facilitate seed voting, resulting in a positive-negative pressure combined seed-metering device. The addition of the positive pressure inlet pipe and chamber altered the operational area of the negative pressure chamber, thereby influencing the uniformity of pressure within the airflow field and enhancing the adsorption effect. The uniformity of the airflow pressure within the negative pressure chamber is directly affected by the structural parameters of the negative pressure inlet pipe, necessitating an optimization of these parameters.

A standard gas-solid two-phase flow coupling field, comprising an airflow field and a seed particle field, operates inside the seed-metering device (Ei-Emam et al., 2021). This analysis not only provides a comprehensive understanding of the key variables affecting the operational efficiency of the seed-metering device and establishes a reasonable range for the working parameters, but it also mitigates the uncertainty associated with prototype trial production through the application of CFD-DEM coupling simulations (Han et al., 2018; Liu et al., 2022b; Darabi et al., 2011). Consequently, a simulation test to assess the influence of the structural parameters of the negative pressure inlet pipe on seeding performance was conducted utilizing the CFD-DEM coupling simulation and analysis methodology in this study. To identify the ideal operating parameters for the seed-metering device equipped with the optimized negative pressure inlet pipe, the adaptability to various operational conditions was evaluated.

2 Structure and working principle of the seed-metering device

2.1 Structure of the seed-metering device

The structural schematic of the positive-negative pressure combined maize precision seed-metering device designed in this study is depicted in Figure 1. It primarily consists of a front shell, internal and external cleaning blades, a seeding plate, positive and negative pressure inlet pipes, a seed layer height adjustment plate, a rear shell, and a seeding tube, among other components. The internal and external cleaning blades are positioned adjacent to the seeding plate to eliminate extraneous seeds near the holes, guaranteeing a consistent single seeding rate. The seed layer height adjustment plate is situated beneath the seed inlet to regulate and control the height of the seed pile. A sealing ring is integrated

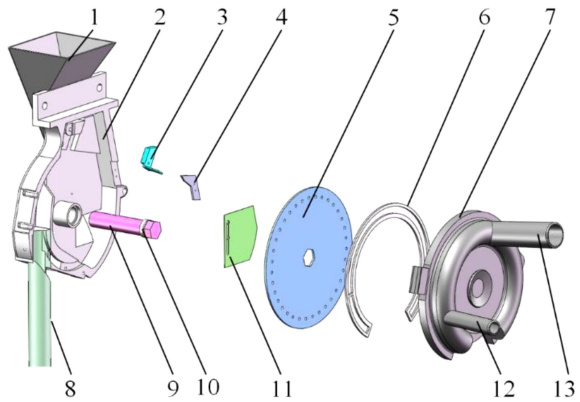


FIGURE 1
Structural diagram of the positive-negative pressure combined seed-metering device. 1. Seed inlet; 2. Front shell; 3. Internal cleaning blade; 4. External cleaning blade; 5. Seeding plate; 6. Sealing ring; 7. Rear shell; 8. Seeding tube; 9. Seeding shaft; 10. Bearing; 11. Seed layer height adjustment plate; 12. Positive pressure inlet pipe; 13. Negative pressure inlet pipe.

between the negative pressure chamber and the seeding plate to maintain the integrity of the negative pressure chamber.

The working principle of the seed-metering device can be delineated into four distinct stages: filling, cleaning, holding, and casting. The respective zones traversed by the seeds include the seed-filling, seed-cleaning, seed-holding, and seed-casting zones, as illustrated in Figure 2. The seed-filling zone extends from the base of the seed-filling chamber to the external cleaning blade, connected by a 90° arc that links the working zones. The seed-cleaning zone is defined as the area between the internal and external cleaning blades, featuring a working arc of approximately 45°. The seed-holding zone spans from the internal cleaning blade to the terminus of the negative pressure chamber, with a working interval of 110°. The positive pressure chamber encompasses the seed-casting zone, characterized by a 35° working interval arc.

When the seed-metering device is operational, seeds are introduced into the seed-filling zone via the seed inlet, while the seed layer height adjustment plate ensures the maintenance of the appropriate seed layer. As the seeding plate rotates counterclockwise, seeds located near the holes in the seed-filling area are drawn into the holes by the adsorption force generated by the negative pressure chamber. The internal and external cleaning blades progressively eliminate excess seeds as the adsorbed seeds transition to the seed-cleaning zone, leaving only one seed that remains stably adsorbed. The holes containing the adsorbed seeds gradually rotate from the seed-holding zone into the seed-casting zone. At this juncture, the air pressure on both sides of the hole shifts from negative to positive, causing the seed to transition from an adsorption state to a blowing. Consequently, the seed that has lost its adsorption force will fall from the hole due to the combined effects of gravity and centrifugal force, subsequently entering the seed guide pipe.

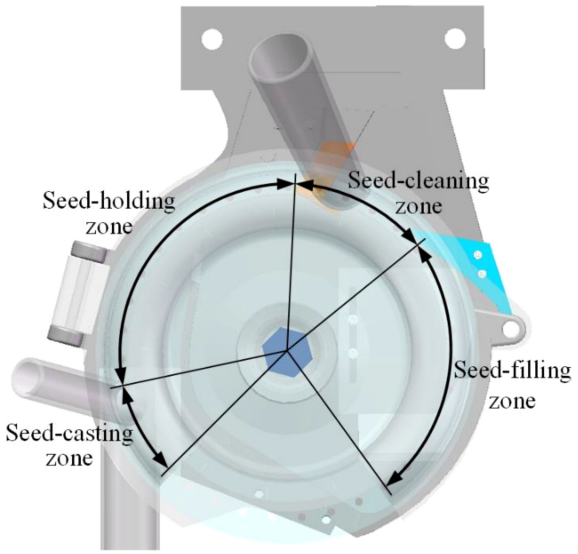


FIGURE 2
Schematic illustration of the seed-metering device's working process.

2.2 Kinematic properties of seeds in the seed-metering device

The seeding performance is ultimately dictated by the movement dynamics of the seed. Therefore, analyzing the movement characteristics of the seeds not only offers a more comprehensive and intuitive understanding of the working principle but also identifies the factors that directly influence seeding performance through the examination of the seeds' movement process. This analysis enables to exploration of the relationship between structural parameters, operational parameters, and the influencing factors. Thus, investigating the movement of seeds within the seed-metering device proves to be highly beneficial.

2.2.1 Force analysis of the seed-filling process

The seed-filling process is divided into two distinct phases to examine the forces acting on the seed, with a spatial coordinate system established using the seed's center as the origin, as depicted in Figure 3. In this spatial coordinate system, the horizontal direction is designated as the x-axis, the direction perpendicular to the seeding plate is the y-axis, and the vertical direction is represented by the z-axis. When the seed is positioned within the seed layer, it experiences the turbulent force (F_t) from the turbulent seed cam, in addition to its gravitational force. The seeds tend to migrate in the same direction as the rotation of the seeding plate. During the seed-filling process, seeds are subjected to airflow drag force (F_D), gradually moving closer to the designated hole. However, they also encounter friction and crowding pressure from the surrounding population, which hinders their adsorption process. Once the seed overcomes the total resistance force (F_r) from the population and becomes adsorbed to the hole, it will then experience the adsorption force (F_a).

Based on the force state of the seed within the seed pile, the condition necessary for the seed to detach from the seed pile and advance toward the hole is established as Equation 1.

$$\begin{cases} \sum F_x \geq 0 : F_D \cos\varphi_{Dx} + F_t cos\varphi_{tx} \geq F_r \cos\varphi_{rx} \\ \sum F_y \geq 0 : F_D \cos\varphi_{Dy} + F_t cos\varphi_{ty} \geq F_r \cos\varphi_{ry} \\ \sum F_z \geq 0 : F_{N1} + F_D \cos\varphi_{Dz} + F_t cos\varphi_{tz} \geq mg + F_r \cos\varphi_{rz} \end{cases} \quad (1)$$

Where F_D is the drag force in N, φ_{Dx}, φ_{Dy}, and φ_{Dz} are the angles of F_D with respect to the x-axis, y-axis, and z-axis in (°), F_t is the turbulent force in N, φ_{tx}, φ_{ty}, and φ_{tz} are the angles of F_t relative to the x-axis, y-axis, and z-axis in (°), F_r is the total resistance force acting on the seed during the seed-filling process in N, φ_{rx}, φ_{ry}, and φ_{rz} are the angles of F_r concerning the x-axis, y-axis, and z-axis in (°), F_{N1} is the support force exerted on the seed by the seed pile in N, m is the mass of the seed in kg, and g is the gravitational acceleration in m/s^2.

The drag force (F_D) varies with the pressure exerted in the negative pressure chamber. The drag force can be estimated using Equation 2 provided below (Han et al., 2017; Su et al., 2023):

$$F_D = \frac{1}{2}\rho_g A_p C_D |v_g - v_p|(v_g - v_p) \quad (2)$$

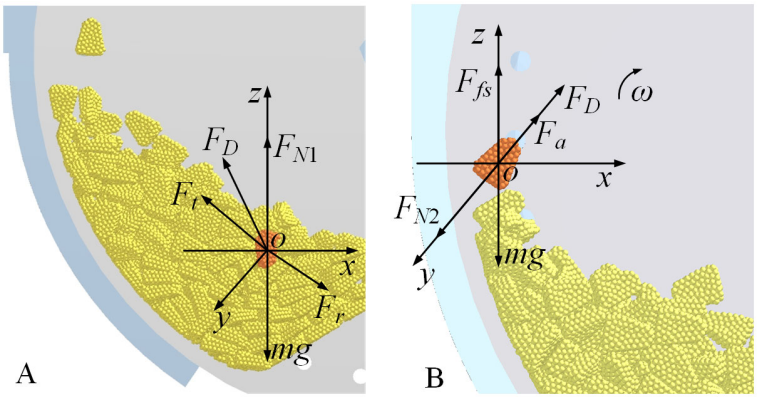


FIGURE 3
Force analysis of seeds during the seed-filling process **(A)** is the seed in the seed pile stage; **(B)** is the seed adsorbed to the hole.

Where ρ_g is the gas density in kg/m^3, A_p is the windward area of the particle in m^2, C_D is the drag force coefficient, v_g is the gas velocity in m/s, and v_p is the particle velocity in m/s.

A Rayleigh curve plotted against the Reynolds number (Re_p) (C_D=f (Re_p) can be utilized using Equation 3 to determine the aerodynamic drag coefficient (C_D) of the particles (Mudarisov et al., 2017).

$$C_D = \begin{cases} \frac{24}{\mathrm{Re}_p}, (\mathrm{Re}_p \le 0.5) \\ \frac{24(1.0+0.25\,\mathrm{Re}_p^{0.687})}{\mathrm{Re}_p}, (0.5 < \mathrm{Re}_p \le 1000) \\ 0.44, (\mathrm{Re}_p > 1000) \end{cases} \quad (3)$$

In this scenario, the Reynolds number of the particle is calculated using Equation 4.

$$\mathrm{Re}_p = \frac{\rho_g d_p (v_g - v_p)}{\mu} \quad (4)$$

Where d_p is the equivalent diameter of the particles in m, and μ is the dynamic viscosity of air in PA·s (at 20°C, μ=1.82×10^{-5} PA·s).

Upon detachment from the adhering population and stabilization within the hole, the forces reach equilibrium that is shown in Equation 5:

$$\begin{cases} \sum F_y = 0 : F_D + F_a - F_{N2} = 0 \\ \sum F_z = 0 : F_{fs} - mg = 0 \end{cases} \quad (5)$$

Where F_a is the adsorption force of the seed by the hole in N, F_{N2} is the support force exerted by the seeding plate in N, and F_{fs} is the frictional force between the seed and the seeding plate in N.

The force analysis during the seed-filling process reveals that an increase in both the drag force (F_D) and the adsorption force (F_a) enhances the seed-filling efficiency. The adsorption force is directly proportional to the differential pressure across two sides of the hole.

2.2.2 Force analysis of the seed-cleaning process

Figure 4 depicts the findings of the force analysis on seeds within both the internal and external seed-cleaning zones. As the seeds are rotated into these zones by the seeding plate, they experience the airflow drag force F_D, the support force from the

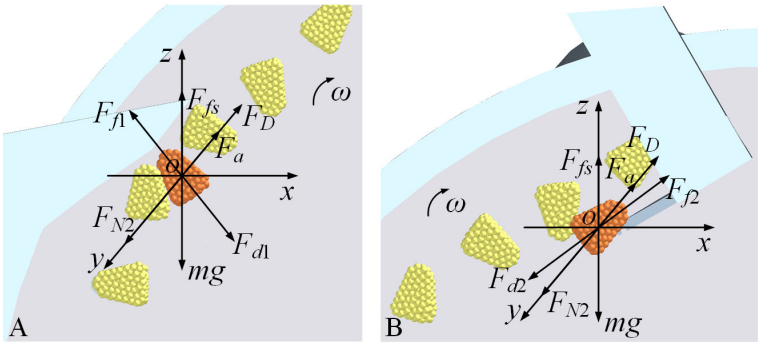


FIGURE 4
Force analysis of seeds during the seed-cleaning process **(A)** is the outer seed-cleaning zone; **(B)** is the inner seed-cleaning zone.

seeding plate F_{N2}, the driving forces from the internal and external seed-cleaning blades F_{d2} and F_{d1}, the frictional forces between the seed and both the seeding plate and the seed-cleaning blades, as well as the adsorption force from the holes F_a, in addition to the gravitational force acting on the seeds.

Seeds in an unstable state near the holes are dislodged by the action of the external and internal seed-cleaning blades, as exemplified by the forces acting on the seed (Equation 6) during the cleaning process of the external seed-cleaning blade:

$$\begin{cases} \sum F_x \geq 0 : F_{d1} cos\varphi_{dx} \geq F_{f1} \cos \varphi_{fx} \\ \sum F_y = 0 : F_D + F_a - F_{N2} = 0 \\ \sum F_z \geq 0 : mg + F_{d1} cos\varphi_{dz} - F_{fs} - F_{f1} \cos \varphi_{fz} \geq 0 \end{cases} \quad (6)$$

Where F_{d1} is the driving force of the external seed-cleaning blade in N, φ_{dx} and φ_{dz} are the angles of F_{d1} with respect to the x-axis and z-axis in (°), F_{f1} is the frictional force between the seed and the external seed-cleaning blade in N, φ_{fx} and φ_{fz} are the angles of F_{f1} concerning the x-axis and z-axis in (°).

The force analysis conducted during the seed-cleaning process, as illustrated in Figure 4, indicates that an increase in the drag force (F_D) acting on the seed correlates with improved seed-filling efficiency. It is crucial to maintain the adsorption force within the seed-cleaning zone, or differential pressure, at an optimal level. The excessive force may hinder the effectiveness of the seed-cleaning function.

2.2.3 Force analysis of the seed-holding process

The results of the force analysis on the seeds within the seed-holding zone are depicted in Figure 5. During the seed-holding process, the seed experiences identical forces as illustrated in Figure 3b when it is separated from the population and adheres to the hole. The force analysis can be referenced in Equation 5.

The force analysis of the seed during the seed-holding process reveals that an increase in both the drag force (F_D) and the adsorption force (F_a) enhances the seed-holding capability.

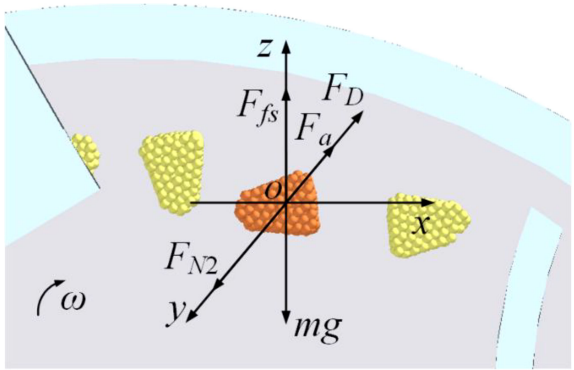


FIGURE 5
Force analysis of seeds during the seed-holding process.

2.3 Structural optimization of the negative-pressure air inlet

The positive-negative pressure combination seed-metering device primarily facilitates the adhesion of seeds to the seeding plate through the differential pressure generated by the negative-pressure chamber flanking the perforations. A more constant pressure within the airflow field significantly contributes to improved seeding uniformity and reduced leakage rate (Ding et al., 2018). The structural and positional features of the negative-pressure inlet pipe directly influence the pressure homogeneity of the airflow field across the chamber, thereby affecting overall seeding performance. Thus, optimizing the structural parameters of the negative pressure inlet pipe is critical for enhancing seeding efficacy. The force analysis of seeds across various working areas, as discussed in section 2.2, indicates that the direction of incoming airflow plays a pivotal role in the seeds' adsorption capacity. The negative pressure inlet pipe is strategically positioned at the center of the negative pressure chamber to ensure a consistent adsorption force for each perforation, as a larger area within the negative pressure chamber necessitates a greater negative pressure.

Based on the investigation of seed dynamics and force conditions during the operation of the seed-metering device, alongside insights from other researchers on air-suction seed-metering devices (Ding et al., 2018; Xu et al., 2022), it is evident that the angle (α) between the negative-pressure inlet pipe in the *XOZ* plane and the vertical axis, the angle (β) of the inlet pipe in the *YOZ* plane relative to the vertical axis, and the inlet pipe's taper (θ) all significantly impact seeding performance. Figure 6 presents a schematic representation of these structural parameters. The objective of this research is to identify the optimal structural parameters of the negative pressure inlet pipe by examining its impact on seeding performance through CFD-DEM coupling simulation analysis. Subsequently, the optimal operating conditions for the seed-metering device and the associated seeding performance will be established by evaluating its adaptability to various operational scenarios.

3 Materials and methods

3.1 CFD-DEM coupling simulation

The interior of the seed-metering device constitutes a standard coupling field comprising both the seed particle field and the airflow field during operation. Therefore, to successfully conduct the simulation test of the coupling process, the following prerequisites must be satisfied: (1) the seed-metering device must be geometrically modeled to facilitate the creation of the airflow field, (2) a particle model of the maize seed must be developed to create the particle field by generating particles via the particle factory, (3) the requisite simulation conditions and parameters must be configured to guarantee the execution of the coupling

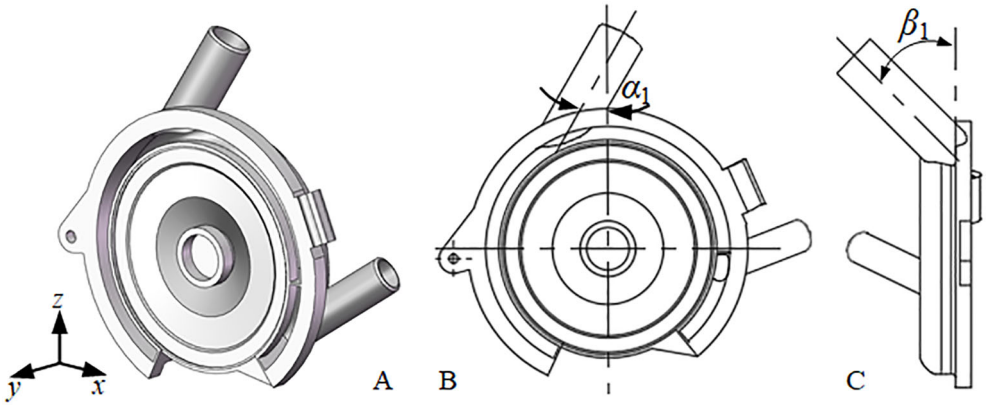


FIGURE 6
Structural parameter diagram of the negative pressure inlet pipe **(A)** is the axonometric drawing; **(B)** is the xoz plane; **(C)** is the yoz plane.

simulation test, and (4) all coupling simulation tests must adhere to the established testing protocol.

3.1.1 Particle modeling of maize seeds

The "Zhongyu 3" maize variety was selected as the subject of this study. Based on its external dimension characteristics, this variety was categorized into flat and spherical shapes, and a 3D model of the maize seeds was constructed using Solidworks (version 2022, Dassault Systèmes SolidWorks Corporation, Waltham, Massachusetts, United States). Previous studies (Su et al., 2022; Long et al., 2022; Li et al., 2022) have effectively simulated the deformation and fracture behaviors of maize (Kozhar et al., 2015) by employing bonded particle model (BPM) to create maize simulation particles. Consequently, the BPM approach was utilized in this paper to model the imported 3D structure in EDEM 2018 (DEM Solution Ltd., Edinburgh, Scotland), and the API particle replacement method was employed to generate the particle model required for simulation. Supplementary Figure S1 displays the maize seed particle model, showcasing the particle bonding model, the 3D representation, and the physical image of the seed, arranged sequentially from top to bottom.

3.1.2 Geometry modeling of the seed-metering device

To enhance the computational efficiency of the CFD-DEM coupling simulation, components that did not contribute to the seeding effect were removed. The simplified DEM model of the seed-metering device included the seed inlet, front and rear shells, internal and external cleaning blades, seed layer height adjustment plate, seeding plate, and both positive and negative pressure inlet pipes. The 3D model was finalized in Solidworks, saved in.step format, and imported into EDEM.

Given the intricate internal structure of the seed-metering device, meshing the entire CFD domain presents challenges. It is essential to streamline and modify the structure of the CFD domain without compromising the integrity of key components concerning structural features and seeding efficacy. To accelerate solving efficiency and accuracy, the CFD domain encompassing positive and negative pressure air inlets, holes, individual chambers, a seeding tube, and other elements meshed using a hexahedral structure with the sweeping method. The seeding plate must be rotated around its axis during operation. Accordingly, within the CFD computational domain, the holes should be aligned to rotate per the seeding plate's rotation in EDEM.

To replicate the relative rotation between the holes and the internal chamber, the moving mesh technique was employed to construct a rotational model of the holes, enabling real-time monitoring of their spatial positional changes. Simultaneously, the holes were designated as dynamic regions, while the remaining CFD domains were fixed as static regions. The contact surfaces between each component of the CFD domain were configured as interfaces, with the corresponding relationships defined in the solver for data transfer. To ensure the precision of the CFD-DEM coupling simulation, the mesh quality was checked using the Examine Mesh function in Gambit. The proportion of meshes exhibiting skewness between 0 and 0.8 reached 100%, indicating exceptional mesh quality. The CFD domain mesh model is depicted in Figure 7.

The seeding plate in EDEM was set to rotate in alignment with the rotation of the holes in the CFD domain, ensuring that both shared the same rotational speed. A particle factory was created within EDEM, utilizing an API particle replacement method to create the maize seed bonded particle model. The simplified DEM model of the seed-metering device and the particle replacement process are depicted in Supplementary Figure S2.

3.1.3 Computational conditions and parameters

The simulation parameters are categorized into two groups: material mechanical parameters and material contact parameters, with their values derived from the actual materials used. In this investigation, the seeding plate is constructed from stainless steel, the front shell from transparent Plexiglas, and the remaining components from aluminum alloy. The mechanical properties of

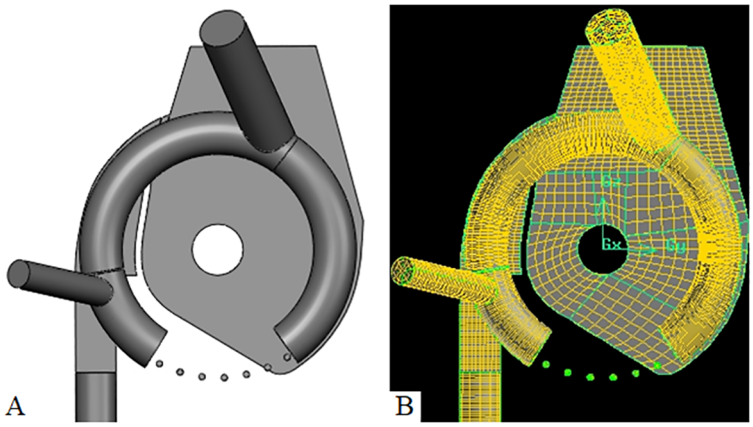


FIGURE 7
CFD dominated mesh model **(A)** is the CFD domain modeling; **(B)** is the CFD domain meshing.

maize seeds, stainless steel, Plexiglas, aluminum alloy, and other materials, along with their physical interactions, were utilized to define the physical properties of the seeds and the contact components, as presented in Table 1. The Hertz-Mindlin no-slip contact model is based on the contact characteristics of the seeds and the seed-metering device in the coupling tests.

To accurately simulate the seed-filling process, 74 flat seeds and 26 spherical seeds were generated in proportion to the ungraded seeds. The time step size in EDEM was set to 5×10^{-6} s, while the Fluent time step size was established at 5×10^{-4} s. The number of time steps in Fluent was configured to 6000, resulting in a total simulation time of 3 s. The total simulation duration in EDEM was also set to 3 s, allowing for the rotation of at least 30 holes in the seeding plate at a forward speed of 6 km/h. To capture the motion data of particles in detail, information was recorded in both EDEM and Fluent every 0.02 s.

3.1.4 CFD-DEM coupling procedure

The coupling simulation calculation process is divided into two segments: the airflow field and the particle field. The drag force affects particle motion, while the motion of the particles also interacts with the airflow field, affecting its flow. These elements are interconnected but must be computed separately within two distinct models. Consequently, two separate simulation models for particles and fluids are required. Supplementary Figure S3 provides a concise overview of the linked flow utilized in this study.

3.2 CFD-DEM coupling simulation test scheme

To investigate the influence of the structural parameters of the negative pressure inlet pipe on seeding performance and to identify the optimal combination of specified levels, the structural parameters outlined in section 2.3 were utilized as test factors. These include the angle (α) between the negative pressure inlet pipe in the *XOZ* plane and the vertical axis, the angle (β) between the inlet pipe in the *YOZ* plane and the vertical axis, and the taper (θ) of the inlet pipe. An orthogonal table $L_9(3^4)$ was chosen to conduct a three-factor, three-level orthogonal experiment. The schematic representations of the test factors and their respective levels are shown in Supplementary Figure S4, with the corresponding values detailed in Table 2.

The evaluation indices are significantly affected by the test factors, and the optimal level combinations for these factors at specified levels can be derived from the ANOVA results of the orthogonal test. To further ascertain the optimal combination of the structural parameters of the negative pressure inlet pipe, a central composite test was continued. In this phase, the optimal combination of the test factors identified from the orthogonal test was employed as the central values, and the corresponding level values were calculated. The evaluation indices that were notably affected by the test factors were designated as test indices.

To determine the optimal structural parameters of the negative pressure inlet pipe, a regression model linking the test factors to the evaluation index—referred to as the prediction model—was established through multiple regression analysis of the central composite test results. The prediction model was then solved, with the maximum value of the evaluation index serving as the target value. The solutions derived represent the optimal predicted structural parameters of the negative pressure inlet pipe. Ultimately, the reliability of these optimal structural parameters was validated by comparing the evaluated index values obtained from the coupling simulation tests of the optimized negative pressure inlet pipe with the predicted evaluation index values.

The ANOVA for both the orthogonal and central composite tests was performed using Design-Expert 13 (version 13, Stat-Ease Ltd., Godward St NE, Minneapolis, USA), while the multiple regression analysis for the central composite test results and the resolution of the prediction model for the optimal structural parameters were executed using its Optimization module.

TABLE 1 The input parameters of CFD-DEM coupling simulation.

Type	Parameters	Maize kernel	Organic glass	Aluminum alloy
Solid phase	Poisson's ratio	0.4	0.50	0.25
	Shear modulus (Pa)	1.37×108	1.77×108	2.70×1010
	Density (kg/m3)	1197	1180	2700
	Restitution coefficient (with maize kernel)	0.182	0.709	0.620
	Static friction coefficient (with maize kernel)	0.0338	0.4590	0.3420
	Rolling friction coefficient (with maize kernel)	0.0021	0.0931	0.0515
	DEM time step (s)	5×10-6		
Gas phase	Fluid	Air		
	Gravitational acceleration (m/s3)	9.81		
	Density (kg/cm3)	1.225		
	Viscosity (kg/m/s)	1.7984×10-5		
	CFD time step (s)	5×10-4		
	Data resources	(Wang et al., 2016; Han et al., 2023; Zhang et al., 2022)		

3.3 Evaluation index design

Outcome indices such as qualified rate, multiple rate, and leakage rate are utilized to assess seeding performance and are not applicable for evaluating the findings from the coupling simulation tests. Section 3.2 indicates that the coupling simulation is divided into two parts: the airflow field and the particle field. The analysis presented in section 2.2 indicates that in the context of the airflow field, the differential pressure and airflow rate are critical factors influencing seeding performance, necessitating extraction from the Fluent software. Concerning the particle field, the drag force emerges as the primary factor affecting seeding performance, which must be derived from the EDEM software.

3.3.1 Differential pressure

The suction force exerted on seeds is a pivotal aerodynamic parameter that significantly impacts seed filling and transport processes. The seed-metering device proposed in this study is fundamentally reliant on the differential pressure between the negative pressure chamber and the seed-filling compartment to effectively adsorb maize seeds to the seeding plate, thereby underscoring the substantial influence of differential pressure on the seeds' adsorption efficacy.

TABLE 2 Factors and levels of the orthogonal test.

Levels	Factors		
	A, α (°)	*B*, β (°)	*C*, θ (°)
1	0	15	0
2	15	45	2
3	30	75	4

Section 2.2 reveals that the adsorption force, represented by the differential pressure during the seed-filling and seed-holding phases, facilitates the stable adhesion of maize seeds to the seeding plate. To enhance the analysis of differential pressure across various operational zones, the holes within the area encompassed by the negative pressure chamber have been systematically numbered, as illustrated in Supplementary Figure S5. The differential pressures associated with each working zone were extracted and averaged for both the seed-filling and seed-holding zones, denoted as Δp_{sf} and Δp_{sh}, respectively.

During the meshing of the CFD domain for the seed-metering device, interfaces for each fluid body were established. Consequently, the pressures on either side of each hole were individually post-processed in Fluent before calculating the differential pressure. The locations for pressure extraction are depicted in Figure 8. To analyze the pressure variations at the holes within each working zone, pressure contours for the seed-filling and seed-holding zones were obtained. The positions from which these pressure contours were extracted are shown in Figure 9. The pressure contours for the holes in each working zone are depicted in Supplementary Figure S6.

Given that the negative pressure inlet pipe is situated in proximity to the seed-cleaning zone, the negative pressure within the negative pressure chamber progressively increases in the seed-filling zone, peaks near the seed-cleaning zone, and subsequently declines in the seed-holding zone, as evidenced by the pressure maps for each working zone. The pressure variation at each hole transitions from "0 pa" on the seed-filling side to the maximum negative pressure, indicating that the simulation test is valid and can be utilized for optimization testing.

3.3.2 Airflow rate

In this design, airflow is introduced into the negative pressure chamber from the seed-filling chamber and exits through the outlet. The flow rate (Q) at the interface between the negative pressure

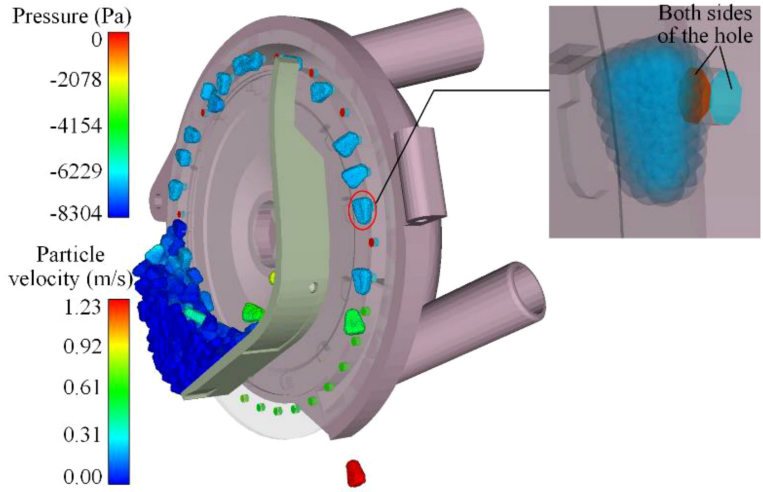


FIGURE 8
Schematic of pressure extraction position.

chamber and the suction pipe greatly influences the overall adsorption seed airflow velocity, thus serving as a crucial evaluation metric for structural optimization. Supplementary Figure S7 depicts the extraction point for the airflow rate at the interface of the air inlet and the negative pressure chamber.

3.3.3 Drag force

During the operation of the seed-metering device, maize seeds are subjected to a multifaceted environment characterized by airflow, particle, and gravitational forces. The seeds experience a confluence of forces from the airflow field, including drag, buoyancy, and pressure gradients, as well as Basset, Magnus, and Saffman lift forces. Given the relatively low velocity of the seeds, the influence of drag force on the seeds is accentuated, taking into account the unique characteristics of each force. Drag force is defined as the force exerted by a moving airflow on a solid object with a relative velocity. The interaction between the airflow and the seed generates a force. The seed, moving at a lower velocity, exerts a resistance effect on the faster-moving airflow, while simultaneously experiencing a drag force from the airflow (Pasha et al., 2015).

An analysis was conducted on the variations in drag force and particle velocity throughout the operational phases of the seed-metering device. The relationships between drag force and particle velocity over time are plotted in Figure 10. The drag force exhibited fluctuations as it transitioned from the seed-filling zone to the seed-holding zone, generally trending upwards, before dissipating in the seed-voting zone when the negative pressure airflow was obstructed. During the initial stage in the seed-filling zone, the seeds encountered a complex array of forces as they transitioned from the seed pile to being adhered to the holes, resulting in some variability in particle velocity. Once the seeds were secured by the holes, they maintained a relatively stable state, leading to a consistent particle velocity. As the negative-pressure adsorption force diminished in the seed-voting zone, the seeds descended into

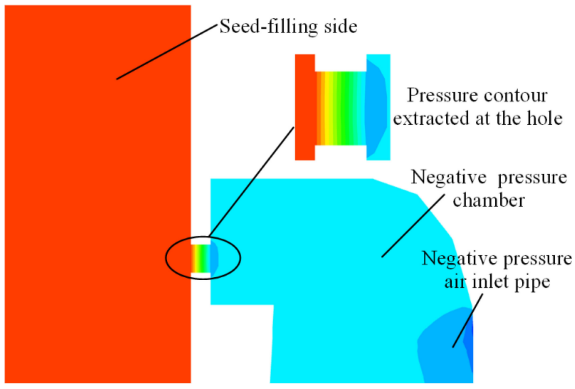


FIGURE 9
Schematic diagram of the pressure contour extracted location of the hole.

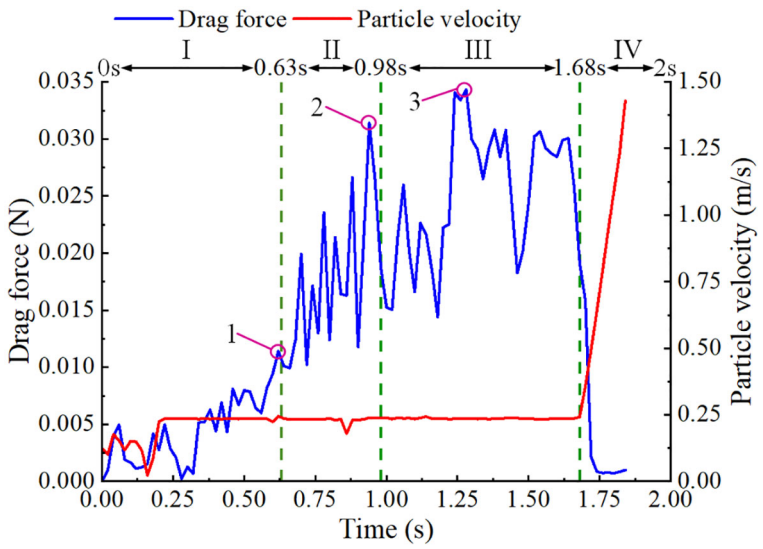


FIGURE 10
The drag force of seeds in each region. I is the seed-filling zone, II is the seed-cleaning zone, III is the seed-carrying zone, and IV is the seed-voting zone.

the tube, accelerating under the action of gravitational force and positive-pressure auxiliary airflow, which resulted in a rapid increase in particle velocity.

Analysis of the kinematic behavior of the seeds across each working zone, as detailed in section 2.3, revealed that drag force played a crucial role in facilitating the adsorption of maize seeds during the seed-filling, seed-cleaning, and seed-holding processes. The maximum drag forces experienced by the same seed in the seed-filling, seed-cleaning, and seed-holding zones are denoted as 1, 2, and 3 in Figure 10, respectively. Following each coupling simulation, the maximum drag forces for ten seeds were extracted from each working zone and averaged, yielding the average drag force in the seed-filling zone ($F_{D,sf}$), the average drag force in the seed-cleaning zone ($F_{D,sc}$), and the average drag force in the seed-holding zone ($F_{D,sh}$).

To thoroughly assess the outcomes of the coupling simulation tests, corresponding evaluation indices were derived from both the airflow and particle fields, as presented in Table 3.

3.4 The adaptability of the seed-metering device to different working conditions

To ascertain the seeding performance and optimal working conditions of the seed-metering device equipped with the optimized negative pressure inlet pipe under varying working conditions, a bench test is proposed for testing. The rear shell was fabricated and shaped utilizing 3D printing technology, employing the predicted optimal structural parameters of the negative pressure inlet pipe integrated into the seed-metering device. This assembly was subsequently affixed to the PZQCSY-2 seeding performance tester in preparation for the experimental evaluation (Figure 11).

Concerning the optimized structural parameters of the negative pressure inlet pipe, an initial comparative bench test was conducted against the pre-optimized air-suction seed-metering device. In this comparative assessment, the theoretical seed spacing was set at 10 cm, the working pressure of the negative pressure inlet pipe was set at -5 kPa, and the working speed of the seeder ranged from 3 to 7 km/h. Furthermore, to achieve a comprehensive understanding of

TABLE 3 Table of evaluation indexes.

Physical field	Evaluation index	Symbol	Unit
Airflow field in CFD	Averaged differential pressure between the two sides of the holes in the seed-filling zone	Δp_{sf}	Pa
	Averaged differential pressure between the two sides of the holes in the seed-holding zone	Δp_{sh}	Pa
	The airflow rate in the interaction surface between the negative pressure inlet pipe and the negative pressure chamber	Q	m^3/min
Particle field in EDEM	Averaged drag force on seeds in the seed-filling zone	$F_{D,sf}$	N
	Averaged drag force on seeds in the seed-cleaning zone	$F_{D,sc}$	N
	Averaged drag force on seeds in the seed-holding zone	$F_{D,sh}$	N

seeding performance under varying operational conditions, a full-factorial test encompassing both working speed and negative pressure will be executed. This will validate the accuracy of the simulation tests and determine the optimal working parameters.

In the full-factorial bench test, the theoretical seed spacing was also established at 10 cm, with the operating speed ranging from 3 to 7 km/h and the operating negative pressure set between 4 to 8 kPa. The maize seed varieties utilized in the experiment were consistent with those employed for particle modeling in section 2.3.1. Following the agronomic requirements for densely planted maize cultivation, the theoretical spacing was set at 10 cm. The evaluation metrics included the qualified rate, multiple rate, and leakage rate. Data collection adhered to the GB/T 6973-2005 (Müller et al., 2006) testing methodologies for single seed drills (precision drills) (National Technical Committee for the Standardization of Agricultural Machinery (2006)), with 250 seeds collected for each trial. Each evaluation index was computed as Equation 7:

$$\begin{cases} Q = \frac{n_1}{N} \times 100\,\% \\ M = \frac{n_2}{N} \times 100\,\% \\ L = \frac{n_3}{N} \times 100\,\% \end{cases} \tag{7}$$

Where Q is the qualified rate in %, M is the multiple rate in %, L is the leakage rate in %, n_1 is the number of single seed holes in one trial, n_2 is the number of holes containing two or more seeds in one trial, n_3 is the count of unseeded holes in one trial, and N is the total number of consecutively recorded seeded holes in one trial, with N equaling 250.

FIGURE 11
Bench test apparatus.

4 Results and analysis

4.1 Orthogonal test results and analysis

The extent to which the structural parameters of the negative pressure inlet pipe affect the evaluation indexes can be ascertained from the ANOVA (Table 4) of the orthogonal coupling simulation test results presented in Table 5, which also reveals the optimal combination of test elements at specified levels.

From the analysis detailed in Table 4, it is evident that the angle (α) has a highly significant impact on the airflow rate (Q) and $F_{D,sc}$. Additionally, angle (α) also generally influences Δp_{sf} and $F_{D,sf}$. Furthermore, the angle (β) demonstrates a generally significant effect on Δp_{sf}, the airflow rate (Q), and $F_{D,sc}$. The taper (θ) extremely significantly influenced Δp_{sf}, the airflow rate (Q), and $F_{D,sc}$, while its impact on $F_{D,sf}$ was also generally notable. Conversely, the effects of the three test factors on Δp_{sh} and $F_{D,sh}$ were not statistically significant.

To further elucidate the optimal level combination that substantially affects the evaluation indices of the test factors, a polar analysis of the test results was conducted. Table 6 presents the outcomes of this analysis after excluding the evaluation indices (Δp_{sh}, $F_{D,sh}$) that were not substantially impacted by the test factors.

The polar analysis indicates that the test factors affect Δp_{sf} in the sequence of C>A>B, with the optimal combination being $C_1A_1B_3$. For Q, the order remains C>A>B, with the best combination identified as $C_2A_2B_2$. Regarding $F_{D,sf}$, the sequence is again C>A>B, with the best combination being $C_1A_2B_3$. In the case of $F_{D,sc}$, the order shifts to A>C>B, with the optimal combination being $A_1C_2B_2$.

Table 6 further illustrates that the optimal levels for each test factor varied across the four evaluation indices. Given the substantial impact of these indices on seeding performance, each was assigned a score of 25 points to facilitate the determination of their optimal levels. Additionally, test factors that did not significantly influence the evaluation indicators were excluded from the calculations. The optimal structural parameters for the

TABLE 4 Variance analysis of orthogonal test results.

Dependent variables	Sources of variance	Sum of deviation squares	*df*	Mean square	*F*-value	*P*-value	Significance
Δp_{sf}	*A*	968535.924	2	484267.962	62.720	0.016	*
	B	1044786.352	2	522393.176	67.658	0.015	*
	C	3276844.675	2	1638422.338	212.202	0.005	**
	Error	15442.121	2	7721.061			
Δp_{sh}	*A*	804077.428	2	402038.714	0.308	0.765	–
	B	171479.251	2	85739.625	0.066	0.938	–
	C	785110.478	2	392555.239	0.301	0.769	–
	Error	2611332.682	2	1305666.341			
Q	*A*	0.025	2	0.012	111.063	0.009	**
	B	0.008	2	0.004	34.429	0.028	*
	C	0.102	2	0.051	456.438	0.002	**
	Error	0.000	2	0.000			
$F_{D,sf}$	*A*	5.180E-5	2	2.590E-5	19.958	0.048	*
	B	6.836E-6	2	3.418E-6	2.634	0.275	–
	C	0.000	2	5.586E-5	43.044	0.023	*
	Error	2.596E-6	2	1.298E-6			
$F_{D,sc}$	*A*	0.000	2	9.782E-5	177.492	0.006	**
	B	3.030E-5	2	1.515E-5	27.492	0.035	*
	C	0.000	2	0.000	229.266	0.004	**
	Error	1.102E-6	2	5.511E-7			
$F_{D,sh}$	*A*	2.274E-5	2	1.137E-5	0.216	0.823	–
	B	9.854E-5	2	4.927E-5	0.934	0.517	–
	C	0.000	2	0.000	4.404	0.185	–
	Error	0.000	2	5.275E-5			

* means significant influence in a 95% confidence interval, ** means significant influence in a 99% confidence interval, - means no significant influence. The same is below.

TABLE 5 Simulation results of the orthogonal test.

No.	A	B	C	CFD simulation results			DEM simulation results		
				Δp_{sf} (Pa)	Δp_{sh} (Pa)	Q (m^3/min)	$F_{D,sf}$ (N)	$F_{D,sc}$ (N)	$F_{D,sh}$ (N)
1	1	1	1	4385.95	498.23	1.231	0.0179	0.0274	0.0333
2	1	2	2	3446.78	634.36	1.310	0.0139	0.0298	0.0320
3	1	3	3	3585.59	1650.18	1.063	0.0106	0.0214	0.0270
4	2	1	2	2609.12	1592.01	1.241	0.0184	0.0210	0.0360
5	2	2	3	2167.75	975.84	1.082	0.0109	0.0100	0.0240
6	2	3	1	4387.25	1231.28	1.290	0.0200	0.0250	0.0347
7	3	1	3	2841.20	2011.06	0.901	0.0050	0.0060	0.0090
8	3	2	1	4179.53	2578.07	1.173	0.0131	0.0180	0.0337
9	3	3	2	4010.32	388.12	1.200	0.0137	0.0210	0.0409

negative pressure inlet pipe at the specified levels were determined to be 7.5° for the angle (α), 55° for the angle (β), and 1° for the inlet pipe taper (θ).

4.2 Response surface test results and analysis

4.2.1 CCD test results and analysis

The coding levels for the central composite test are presented in Table 7, while the results of the central composite test and the corresponding ANOVA analysis are shown in Tables 8–10.

From the ANOVA of the airflow field simulation results in Table 9, it is evident that the effects of angle β (x_2), taper θ (x_3), and their interactions (x_2x_3), as well as ${x_1}^2$, ${x_2}^2$, and ${x_3}^2$ on Δp_{sf} (y_1), were extremely significant. The influence of angle α (x_1) was generally significant, whereas the effects of the other factors were not statistically significant. The influence of angle α (x_1), taper θ (x_3), the interaction (x_2x_3) between angle β (x_2) and taper θ (x_3), as well as ${x_1}^2$ and ${x_2}^2$ on the flow rate Q (y_2) was found to be highly significant. Conversely, the interaction (x_1x_2) between angle α (x_1) and angle β (x_2) exhibited a generally significant effect, while the impacts of other factors were deemed insignificant.

From the ANOVA analysis of the particle field simulation results presented in Table 10, it is evident that angle α (x_1) and angle β (x_2) exerted extremely significant effects on $F_{D,sf}$ (y_3). The taper θ (x_3) and the interaction (x_2x_3) between angle β (x_2) and taper θ (x_3) demonstrated generally significant effects, whereas the influence of other factors was not significant. The taper θ (x_3) and ${x_1}^2$ had extremely significant effects on $F_{D,sc}$(y_4), while angle α (x_1), angle β (x_2), and the interaction (x_1x_3) between angle α (x_1) and taper θ (x_3) showed generally significant effects, with other factors being insignificant.

TABLE 6 Range analysis of the significant factors influencing evaluation indicators.

Evaluation indicators	Factors	Levels			Extreme Difference	Optimal level	Optimal combination
		1	2	3			
Δp_{sf}	A	3806.107	3054.707	3677.017	751.40	1	$C_1A_1B_3$
	B	3278.757	3264.687	3994.387	729.70	3	
	C	4317.577	3355.407	2864.847	1452.73	1	
Q	A	1.201	1.204	1.091	0.113	2	$C_2A_2B_2$
	B	1.124	1.188	1.184	0.064	2	
	C	1.231	1.250	1.015	0.235	2	
$F_{D,sf}$	A	0.0141	0.0164	0.0106	0.0058	2	$C_1A_2B_3$
	B	0.0138	0.0126	0.0148	0.0021	3	
	C	0.0170	0.0153	0.0089	0.0082	1	
$F_{D,sc}$	A	0.0262	0.0164	0.0106	0.0156	1	$A_1C_2B_2$
	B	0.0181	0.0229	0.0225	0.0048	2	
	C	0.0235	0.0239	0.0125	0.0115	2	

The ANOVA results from the simulation tests concerning airflow and particle fields suggest that the outcomes from the central composite tests can be utilized for developing a predictive model for the structural parameters of the negative pressure inlet pipe.

4.2.2 Comprehensive analysis

To enhance the understanding of how the structural characteristics of the negative pressure inlet pipe affect the evaluation index, the CFD-DEM coupling simulation process warrants an assessment. For instance, Figure 12 depicts the adsorption state of seeds across various working zones, demonstrating that the seeding process generated by the coupling simulation aligns with actual conditions.

Supplementary Figure S8 presents the fifteen sets of CFD-DEM coupling simulation results derived from the central composite test indicating that the speed and adsorption effects of the seeds differ when they adhere to the seeding plate, influenced by varying structural parameters of the negative pressure inlet pipe. Although tests conducted under identical operating parameters (negative pressure of 7 kPa) yielded a relatively similar degree of seed population boiling and velocity changes within the seed pile, the findings collectively affirm that these test results can inform the establishment of a predictive model for the structural parameters of the negative pressure inlet pipe.

4.3 Parameters optimization and validation

To forecast the structural parameters of the negative pressure inlet pipe, the central composite test results in Table 8 were subjected to multivariate regression analysis. The regression Equation 8 correlating the structural parameters of the negative pressure inlet pipe (α, β, and θ) with the evaluation indices Δp_{sf}, Q, $F_{D,sf}$, and $F_{D,sc}$ were established accordingly.

$$\begin{cases} y_1 = 4981.31 - 85.08x_1 + 122.77x_2 - 139.19x_3 + 43.51x_1x_2 - 7.05x_1x_3 - 169.27x_2x_3 \\ \quad - 170.81x_1^2 - 212.63x_2^2 - 226.56x_3^2 \\ y_2 = 1.28 - 0.011x_1 - 0.0032x_2 - 0.0142x_3 + 0.0096x_1x_2 - 0.0004x_1x_3 - 0.0139x_2x_3 \\ \quad - 0.0106x_1^2 + 0.0147x_2^2 - 0.0019x_3^2 \\ y_3 = 0.0147 + 0.0008x_1 + 0.0011x_2 - 0.0006x_3 + 0.0006x_1x_2 - 0.0005x_1x_3 + 0.0008x_2x_3 \\ \quad + 0.0001x_1^2 + 0.0001x_2^2 - 0.0000x_3^2 \\ y_4 = 0.0241 + 0.0009x_1 + 0.0011x_2 + 0.0013x_3 + 0.0001x_1x_2 - 0.0012x_1x_3 + 0.0003x_2x_3 \\ \quad + 0.0013x_1^2 + 0.0002x_2^2 - 0.0007x_3^2 \end{cases} \tag{8}$$

TABLE 7 Factors and levels of central composite design.

Levels	Factors		
	α (°)	β (°)	θ (°)
−1.682	−7.50	25.00	0.00
−1	−1.42	37.16	0.41
0	7.50	55.00	1.00
1	16.42	72.84	1.59
1.682	22.50	85.00	2.00

Where x_1 is the angle (α) between the negative pressure inlet pipe at the XOZ plane and the vertical axis, x_2 is the angle (β) between the inlet pipe at the YOZ plane and vertical axis, x_3 is the taper of the inlet pipe (θ), y_1 is the average differential pressures across the two sides of the holes in the seed-filling zone (Δp_{sf}) in Pa, y_2 is the airflow rate at the interaction interface between the negative pressure inlet pipe and the negative pressure chamber (Q) in m^3/min, y_3 is the average drag force in the seed-filling zone ($F_{D,sf}$) in N, and y_4 is the average drag force in the seed-cleaning zone ($F_{D,sc}$) in N.

The ANOVA results presented in Tables 9, 10 demonstrate that the prediction models exhibit a high goodness of fit (p<0.0068), with the lack of fit terms showing no significant effects (p>0.1251). This indicates that the prediction models are suitable for determining the optimal structural parameters of the negative pressure inlet pipe. The optimal parameter solution model for the negative pressure inlet pipe was constructed using the maximum values of the evaluation indices Δp_{sf}, Q, $F_{D,sf}$, and $F_{D,sc}$ as target values, while the maximum and minimum levels of the test factors listed in Table 6 served as constraints.

$$\begin{cases} \max(y_1, y_2, y_3, y_4) \\ s.t. \begin{cases} -7.5° \le x_1 \le 22.5° \\ 25° \le x_2 \le 85° \\ 0° \le x_3 \le 2° \end{cases} \end{cases} \tag{9}$$

The solution to Equation 9 was obtained using the Optimization module of the Design-Expert software, yielding the predicted optimal parameters for the negative pressure inlet pipe, which are as follows: the angle (α) is 17.1°, the angle (β) is 81.3°, and the taper of the inlet pipe (θ) is 0.52°.

Table 11 provides a comparison between the predicted values of the evaluation indices and those obtained from the coupling simulation verification test of the seed-metering device equipped with the enhanced negative pressure inlet pipe. The validation test results indicate that Δp_{sf} is 4526 Pa, the airflow rate Q is 1.26 m^3/min, $F_{D,sf}$ is 0.0168 N, and $F_{D,sc}$ is 0.029 N. These results are largely consistent with the anticipated theoretical optimal search outcomes, with the maximum error being less than 6.66%, thereby affirming the validity of the prediction model for the structural parameters of the negative pressure inlet pipe.

4.4 Analysis of the adaptability of the seed-metering device to different working conditions

4.4.1 Improved analysis of the seeding performance with negative pressure inlet pipe

For the structural parameters of the optimized negative pressure inlet pipe, comparative bench tests were conducted alongside the pre-optimized air-suctioin seed-metering device, with results illustrated in Figure 13. The qualified rate of the seed-metering device equipped with the optimized negative pressure inlet pipe has improved to a certain extent, while both the multiple rate and leakage rate have been reduced. This indicates that the optimized

10.3389/fpls.2025.1485710

TABLE 8 Scheme and results of CCD tests.

No.	Factors			CFD simulation results		DEM simulation results	
	x_1	x_2	x_3	Δp_{sf} (Pa)	Q (m^3/min)	$F_{D,sf}$ (N)	$F_{D,sc}$ (N)
1	−1	−1	−1	4259.73	1.310	0.0133	0.0209
2	1	−1	−1	4136.09	1.274	0.0152	0.0253
3	−1	1	−1	4934.43	1.313	0.0141	0.0205
4	1	1	−1	4721.32	1.294	0.0171	0.0268
5	−1	−1	1	4376.65	1.318	0.0125	0.0246
6	1	−1	1	3961.30	1.259	0.0114	0.0256
7	−1	1	1	4110.77	1.244	0.0152	0.0268
8	1	1	1	4132.96	1.245	0.0175	0.0267
9	−1.682	0	0	4686.24	1.257	0.0139	0.0282
10	1.682	0	0	4429.35	1.235	0.0168	0.0283
11	0	−1.682	0	4287.63	1.311	0.0141	0.022
12	0	1.682	0	4591.43	1.324	0.0163	0.0281
13	0	0	−1.682	4528.26	1.291	0.0167	0.0203
14	0	0	1.682	4271.96	1.250	0.0133	0.0246
15	0	0	0	4836.68	1.287	0.0135	0.0257
16	0	0	0	4925.34	1.263	0.0141	0.0242
17	0	0	0	5087.42	1.285	0.0152	0.0229
18	0	0	0	4982.35	1.275	0.0155	0.0233
19	0	0	0	5048.23	1.282	0.0149	0.0241
20	0	0	0	4987.39	1.278	0.0146	0.0245

x_1, x_2, and x_3 are the levels of α, β, and θ, respectively.

negative pressure inlet pipe enhances seeding performance. Additionally, as depicted in Figure 13, the qualified rate and multiple rate of the seed-metering device with the optimized negative pressure inlet pipe tend to decrease as the working speed of the seeder increase, whereas the leakage rate shows a gradual increase with the rising working speed. Therefore, to optimize the performance metrics of qualified rate, multiple rate and leakage rate, a full factorial test was carried out on the seed-metering device featuring the optimized negative pressure inlet pipe.

4.4.2 Seeding performance under various operating conditions

The results of the adaptation test are displayed in Figure 14, showcasing the performance of the seed-metering device with the optimized negative pressure input pipe under varying operational conditions.

As the working speed ranges from 3 to 7 km/h and the working negative pressure varies between 4 to 8 kPa, it is observed that the qualified rate exhibits a gradual decrease and increase in response to the rising working speed and working negative pressure. As the working negative pressure increases, the multiple rate also rises, while the leakage rate tends to increase with the working speed but decreases as the working negative pressure escalates.

Consequently, the optimal operational parameters for the seed-metering device should be set within a working speed range of 4 to 5 km/h and a working negative pressure range of 4 to 6 kPa, considering the energy losses and operational efficiency during field applications. Under these specified conditions, the seed-metering device achieves a qualified rate ranging from 92.27% to 95.28%, a multiple rate between 2.43% and 5.52%, and a leakage rate from 1.22% to 5.2%.

5 Conclusions

In this study, a coupling simulation test methodology was employed to investigate the effect of the structural parameters of the negative pressure inlet pipe on the seeding performance of the positive-negative pressure combined seed-metering device. The adaptability of the optimized seed-metering device to varying operational conditions was assessed to identify the optimal working conditions and their corresponding seeding performance, leading to the following conclusions:

TABLE 9 Variance analysis of CFD simulation results of the CCD tests.

Sources	y_1 (Δp_{sf})					y_2 (Q)				
	Sum of squares	df	*F*-value	*P*-value	Significance	Sum of squares	df	*F*-value	*P*-value	Significance
Model	2.332E+06	9	16.2700	<0.0001	**	0.0121	9	9.8200	0.0007	**
x_1	98860.10	1	6.2100	0.0319	*	0.0016	1	12.0100	0.0061	**
x_2	2.058E+05	1	12.9200	0.0049	**	0.0001	1	0.9935	0.3424	–
x_3	2.646E+05	1	16.6100	0.0022	**	0.0028	1	20.0800	0.0012	**
x_1x_2	15144.09	1	0.9506	0.3526	–	0.0007	1	5.4000	0.0424	*
x_1x_3	397.76	1	0.0250	0.8776	–	1.125E-06	1	0.0082	0.9296	–
x_2x_3	2.292E+05	1	14.3900	0.0035	**	0.0015	1	11.2300	0.0074	**
x_1^2	4.205E+05	1	26.3900	0.0004	**	0.0016	1	11.7100	0.0065	**
x_2^2	6.515E+05	1	40.9000	<0.0001	**	0.0031	1	22.7800	0.0008	**
x_3^2	7.397E+05	1	46.4400	<0.0001	**	0.0001	1	0.3773	0.5528	–
Residual	1.593E+05	10				0.0014	10			
Lack of Fit	1.195E+05	5	3.0100	0.1261	–	0.0010	5	2.6200	0.1574	–
Error	39756.44	5				0.0004	5			
Cor Total	2.492E+06	19				0.0135	19			

* means significant influence in a 95% confidence interval, ** means significant influence in a 99% confidence interval, - means no significant influence.

1. Orthogonal tests indicate that the structural parameters of the negative pressure inlet pipe significantly affect evaluation metrics such as Δp_{sf}, the airflow rate (Q), $F_{D,sf}$, and $F_{D,sc}$. The evaluation metrics that are notably influenced yield optimal performance when the angle (α) is set at 7.5°, the angle (β) is 55°, and the taper of the inlet pipe (θ) is 1°.
2. Results from the central composite test results demonstrate that the coupling simulation accurately reflects the seeding process of the seed-metering device, and these test

TABLE 10 Variance analysis of DEM simulation results of the CCD tests.

Sources	y_3 ($F_{D,sf}$)					y_4 ($F_{D,sc}$)				
	Sum of squares	df	*F*-value	*P*-value	Significance	Sum of squares	df	*F*-value	*P*-value	Significance
Model	0.0000	9	5.4900	0.0068	**	0.0001	9	5.2200	0.0065	**
x_1	8.823E-06	1	10.7600	0.0083	**	0.0000	1	5.1400	0.0450	*
x_2	0.0000	1	20.6300	0.0011	**	0.0000	1	7.9800	0.0172	*
x_3	5.694E-06	1	6.9400	0.0250	*	0.0000	1	9.8100	0.0069	**
x_1x_2	2.531E-06	1	3.0900	0.1095	–	8.000E-08	1	0.0406	0.8429	–
x_1x_3	1.711E-06	1	2.0900	0.1792	–	0.0000	1	6.0900	0.0319	*
x_2x_3	4.651E-06	1	5.6700	0.0385	*	6.050E-07	1	0.3069	0.5882	–
x_1^2	1.745E-07	1	0.2128	0.6544	–	0.0000	1	12.5300	0.0048	**
x_2^2	4.686E-08	1	0.0571	0.8159	–	4.549E-07	1	0.2308	0.6173	–
x_3^2	2.701E-09	1	0.0033	0.9554	–	6.832E-06	1	3.4700	0.0741	–
Residual	8.202E-06	10				0.0000	10			
Lack of Fit	5.489E-06	5	2.0200	0.2289	–	0.0000	5	3.10	0.1251	–
Error	2.713E-06	5				4.808E-06	5	5.22		
Cor Total	0.0000	19				0.0001	19	5.14		

* means significant influence in a 95% confidence interval, ** means significant influence in a 99% confidence interval, - means no significant influence.

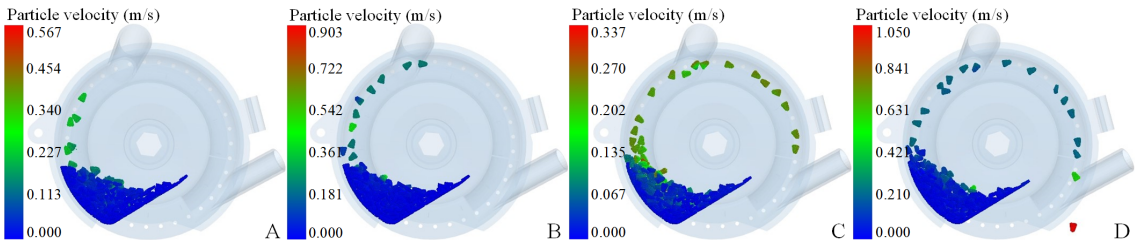


FIGURE 12
The adsorption state of seeds in each working area **(A)** is seed-filling process; **(B)** is seed-cleaning process; **(C)** is seed-carrying process; **(D)** is seed-voting process.

TABLE 11 Simulation verification test results.

Parameter	Evaluation metrics			
	Δp_{sf} (N)	Q (m^3/min)	$F_{D,sf}$ (N)	$F_{D,sc}$ (N)
Predicted value	4651	1.32	0.0180	0.028
Measured value	4526	1.26	0.0168	0.029
Error (%)	2.68	4.55	6.66	3.57

outcomes can be utilized to develop a predictive model for the structural parameters of the negative pressure inlet pipe.

3. The predictive model for the structural parameters of the negative pressure inlet pipe exhibits strong goodness of fit, with the impacts of the lack of fit terms being negligible, thereby rendering it suitable for predicting structural parameters. The ideal combination of predicted structural parameters for the negative pressure inlet pipe includes an angle (α) of 17.1°, an angle (β) of 81.3°, and a taper (θ) of 0.52°. The evaluation index values derived from the coupling simulation tests utilizing this optimal parameters combination align closely with the predicted values, exhibiting a maximum deviation of no more than 6.66%, thereby affirming the plausibility of the optimal prediction parameters.
4. The seed-metering device, equipped with the optimized negative pressure inlet pipe, operates most effectively at a speed of 4 to 5 km/h and a negative pressure range of 4 to 6 kPa. Under these operational conditions, the device achieves a qualified rate between 92.27% and 95.28%, a multiple rate from 2.43% to 5.52%, and a leakage rate ranging from 1.22% to 5.2%.

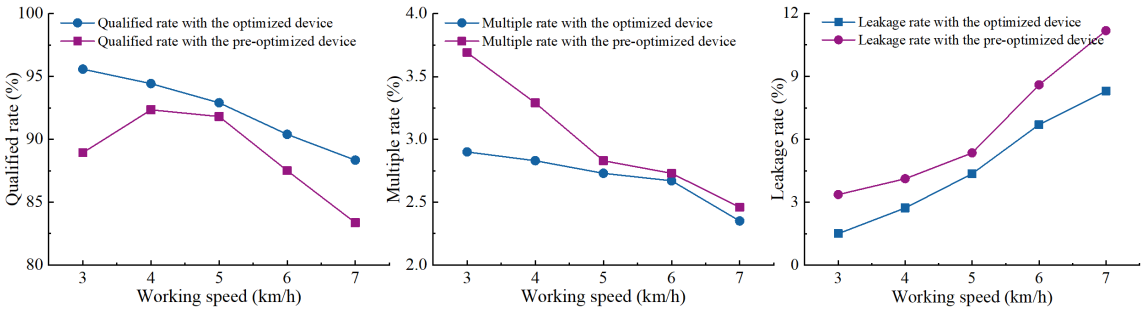


FIGURE 13
Results of the negative pressure inlet pipe on seeding performance enhancement.

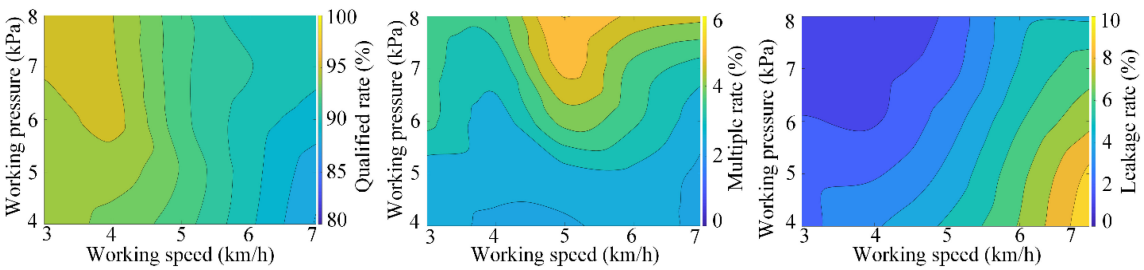


FIGURE 14
Bench test results.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author/s.

Author contributions

DH: Conceptualization, Data curation, Funding acquisition, Methodology, Writing – original draft. WL: Data curation, Software, Writing – original draft. YXW: Data curation, Software, Validation, Writing – review & editing. QW: Methodology, Writing – review & editing. ZW: Formal Analysis, Software, Writing – review & editing. YCW: Methodology, Resources, Writing – review & editing. YX: Data curation, Software, Validation, Writing – review & editing. LX: Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The authors acknowledge the support from the Natural Science Foundation of Sichuan Province (2022NSFSC0138), Technological Innovation R&D Projects of Chengdu Science and Technology Bureau (2022YF0501141SN), and Listing Project of Rural Revitalization Research Institute of Sichuan Tianfu District (XZY1-11).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2025.1485710/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Modeling of the maize kernel bonded particles.

SUPPLEMENTARY FIGURE 2
DEM geometrical model.

SUPPLEMENTARY FIGURE 3
CFD-DEM coupling flowchart.

SUPPLEMENTARY FIGURE 4
Schematics of the factors and levels for the air inlet position.

SUPPLEMENTARY FIGURE 5
Schematic diagram of the hole numbering.

SUPPLEMENTARY FIGURE 6
Pressure contours of the holes in each working zone.

SUPPLEMENTARY FIGURE 7
Schematic diagram of airflow retrieved position. I is the seed-filling zone, II is the seed-cleaning zone, III is the seed-carrying zone, and IV is the seed-voting zone.

SUPPLEMENTARY FIGURE 8
Simulation renderings of CCD test 1~15 series.

References

Darabi, P., Pougatch, K., Salcudean, M., and Grecov, D. (2011). DEM investigations of fluidized beds in the presence of liquid coating. *Powder. Technol.* 214, 365–374. doi: 10.1016/j.powtec.2011.08.032

Ding, L., Yang, L., Wu, D. H., Li, D., Zhang, D. X., and Liu, S. R. (2018). Simulation and experiment of corn air suction seed metering device based on DEM-CFD coupling method. *Trans. Chin. Soc Agric. Mach.* 49, 48–57. doi: 10.6041/j.issn.1000-1298.2018.11.006

Ei-Emam, M. A., Zhou, L., Shi, W. D., Han, C., Bai, L., and Agarwal, R. (2021). Theories and applications of CFD-DEM coupling approach for granular flow: A review. *Arch. Comput. Method. E.* 28, 4979–5020. doi: 10.1007/s11831-021-09568-9

Guarella, P., Pellerano, A., and Pascuzzi, S. (1996). Experimental and theoretical performance of a vacuum seed nozzle for vegetable seeds. *J. Agric. Eng. Res.* 64, 29–36. doi: 10.1006/jaer.1996.0043

Han, D. D., He, B., Wang, Q., Zhang, R. C., Tang, C., Li, W., et al. (2024). Optimization and experiment of seed-filling performance of the air-suction densely planted seed-metering device based on DEM. *Comput. Part. Mech.* In press. doi: 10.1007/s40571-024-00734-x

Han, D. D., Xu, Y., Huang, Y. X., He, B., Dai, J. W., Lv, X. R., et al. (2023). DEM parameters calibration and verification for coated maize particles. *Comput. Part. Mech.* 10, 1931–1941. doi: 10.1007/s40571-023-00598-7

Han, D. D., Zhang, D. X., Jing, H. R., Yang, L., Cui, T., Ding, Y. Q., et al. (2018). DEM-CFD coupling simulation and optimization of an inside-filling air-blowing maize precision seed-metering device. *Comput. Electron. Agric.* 150, 426–438. doi: 10.1016/j.compag.2018.05.006

Han, D. D., Zhang, D. X., Yang, L., Cui, T., Ding, Y. Q., and Bian, X. H. (2017). Optimization and experiment of inside-filling air-blowing seed metering device based on EDEM-CFD. *Trans. Chin. Soc Agric. Mach.* 48, 43–51. doi: 10.6041/j.issn.1000-1298.2017.11.006

Jack, D. S., Hesterman, D. C., and Guzzomi, A. L. (2013). Precision metering of Santalum spicatum (Australian Sandalwood) seeds. *Biosyst. Eng.* 115, 171–183. doi: 10.1016/j.biosystemseng.2013.03.004

Karayel, D., Güngör, O., and Šarauskis, E. (2022). Estimation of optimum vacuum pressure of air-Suction seed-metering device of precision seeders using artificial neural network models. *Agronomy* 12, 1600. doi: 10.3390/agronomy12071600

Karayel, D., and Özmerzi, A. (2001). Effect of forward speed and seed spacing on seeding uniformity of a precision vaccum metering unit for melon and cucumber seeds. *J. Fac. Agr.* 14, 63–67.

Kostić, M., Rakić, D., Radomirović, D., Savin, L., Dedović, N., Crnojević, V., et al. (2018). Corn seeding process fault cause analysis based on a theoretical and experimental approach. *Comput. Electron. Agric.* 151, 207–218. doi: 10.1016/j.compag.2018.06.014

Kozhar, S., Dosta, M., Antonyuk, S., Heinrich, S., and Bröckel, U. (2015). DEM simulation of amorphous irregular shaped micrometer-sized titania agglomerates at compression. *Adv. Powder. Technol.* 26, 767–777. doi: 10.1016/j.apt.2015.05.005

Lai, Q. H., Ma, W. P., Su, W., Hu, Z. W., and Xing, J. L. (2016). Design and experiment of pneumatic disc seed-metering device for mini-tuber. *Trans. Chin. Soc Agric. Mach.* 47, 30–37. doi: 10.6041/j.issn.1000-1298.2016.12.005

Li, X. Y., Du, Y. F., Liu, L., Mao, E. R., Wu, J., Zhang, Y. N., et al. (2022). A rapid prototyping method for crop models using the discrete element method. *Comput. Electron. Agric.* 203, 107451. doi: 10.1016/j.compag.2022.107451

Liu, R., Liu, Z. J., Liu, L. J., and Li, Y. H. (2022a). Design and experiment of corn high speed air suction seed metering device with disturbance assisted seed-filling. *Trans. Chin. Soc Agric. Mach.* 53, 50–59. doi: 10.6041/j.issn.1000-1298.2022.09.005

Liu, R., Liu, Z. J., Zhao, J. L., Lu, Q., Liu, L., and Li, Y. H. (2022b). Optimization and experiment of a disturbance-assisted seed filling high-speed vacuum seed-metering device based on DEM-CFD. *Agriculture* 12, 1304. doi: 10.3390/agriculture12091304

Long, S. F., Xu, S. M., Zhang, Y. J., Zhang, J., and Wang, J. (2022). Effect of modeling parameters on the mechanical response of macroscopic crushing of agglomerate. *Powder. Technol.* 408, 117720. doi: 10.1016/j.powtec.2022.117720

Mudarisov, S., Khasanov, E., Rakhimov, Z., Gabitov, I., Badretdinov, I., Farchutdinov, I., et al. (2017). Specifying two-phase flow in modeling pneumatic systems performance of farm machines. *J. Mech. Eng. Res. Dev.* 40, 706–715. doi: 10.7508/jmerd.2017.04.018

Müller, J., Wiesehoff, M., and Hörner, R. (2006). Seed spacing evaluation for seed drills-ISO-Standard 7256/2 is in need of revision. *Landtechnik* 61, 374–375.

National Technical Committee for the Standardization of Agricultural Machinery. (2006). *GB/T6973—2005 Testing methods of single seed drills (precision drills).* (Beijing: Standards Press of China).

Pareek, C. M., Tewari, V. K., and Machavaram, R. (2023). Multi-objective optimization of seeding performance of a pneumatic precision seed metering device using integrated ANN-MOPSO approach. *Eng. Appl. Artif. Intel.* 117, 105559. doi: 10.1016/j.engappai.2022.105559

Pasha, M., Hassanpour, A., Ahmadian, H., Tan, H. S., Bayly, A., and Ghadiri, M. (2015). A comparative analysis of particle tracking in a mixer by discrete element method and positron emission particle tracking. *Powder. Technol.* 270, 569–574. doi: 10.1016/j.powtec.2014.09.007

Pezzuolo, A., Guarino, M., Sartori, L., Gonz´alez, L. A., and Marinello, F. (2018). On-barn pig weight estimation based on body measurements by a Kinect v1 depth camera. *Comput. Electron. Agric.* 148, 29–36. doi: 10.1016/j.compag.2018.03.003

Shafii, S., and Holmes, R. G. (1990). Air jet seed metering a theoretical and experimental study. *Trans. ASAE.* 33, 1432–1438. doi: 10.13031/2013.31489

Shi, S., Liu, H., Wei, G. J., Zhou, J. L., Jian, S. C., and Zhang, R. F. (2020). Optimization and experiment of pneumatic seed metering device with guided assistant filling based on EDEM-CFD. *Trans. Chin. Soc Agric. Mach.* 51, 54–66. doi: 10.6041/j.issn.1000-1298.2020.05.006

Su, Y., Xu, Y., Cui, T., Gao, X. J., Xia, G. Y., Li, Y. B., et al. (2022). A combined experimental and DEM approach to optimize the centrifugal maize breakage tester. *Powder. Technol.* 397, 117008. doi: 10.1016/j.powtec.2021.11.052

Su, W., Zhao, Q. H., Lai, Q. H., Xie, G. F., Tian, B. N., and Wang, Y. J. (2023). Design and experiment of air-suction broad bean seed metering device with flat belt auxiliary seed-filling. *Trans. Chin. Soc Agric. Mach.* 54, 144–155. doi: 10.6041/j.issn.1000-1298.2023.07.014

Tang, H., Xu, F. D., Gaun, T. Y., Xu, C. S., and Wang, J. W. (2023). Design and test of a pneumatic type of high-speed maize precision seed metering device. *Comput. Electron. Agric.* 211, 107997. doi: 10.1016/j.compag.2023.107997

Vianna, L. R., Dos Reis, A. V., and MaChado, A. L. T. (2014). Development of a horizontal plate meter with double seed outlets. *Rev. Bras. Eng. Agr. Amb.* 18, 1086–1091. doi: 10.1590/1807-1929/agriambi.v18n10p1086-1091

Wang, Y. X., Liang, Z. J., Zhang, D. X., Cui, T., Shi, S., Li, K. H., et al. (2016). Calibration method of contact characteristic parameters for corn seeds based on EDEM. *Trans. Chin. Soc Agric. Eng.* 32, 36–42. doi: 10.11975/j.issn.1002-6819.2016.22.005

Wang, J. W., Qi, X., Xu, C. S., Wang, Z. M., Jiang, Y. M., and Tang, H. (2021). Design evaluation and performance analysis of the inside-filling air-Assisted high-speed precision maize seed-metering device. *Sustainability* 13, 5483. doi: 10.3390/su13105483

Wang, G. W., Xia, X. M., Zhu, Q. H., Yu, H. Y., and Huang, D. Y. (2020). Design and experiment of soybean high-speed precision vacuum seed metering with auxiliary filling structure based on DEM-CFD. *J. Jilin. Univ.* 52, 1209–1221. doi: 10.13229/j.cnki.jdxbgxb20211366

Xu, J., Hou, J. W., Wu, W. B., Han, C. Y., Wang, X. M., Tang, T., et al. (2022). Key structure design and experiment of air-suction vegetable seed-metering device. *Agronomy* 12, 675. doi: 10.3390/agronomy12030675

Yang, W. Y., and Yang, F. (2019). Developing maize-soybean strip intercropping for demand security of national food. *Sci. Agric. Sin.* 52, 3748–3750. doi: 10.3864/j.issn.0578-1752.2019.21.003

Yang, H., Zhou, Y., Chen, P., Du, Q., Zheng, B. C., Pu, T., et al. (2022). Effects of nutrient uptake and utilization on yield of maize-legume strip intercropping system. *Acta Agron. Sin.* 48, 1476–1487. doi: 10.3724/SP.J.1006.2022.13017

Zhang, L. H., Cai, J. X., Li, Y. H., Wang, X. C., and Yang, W. Y. (2020). Research progress of mechanization technology and equipment for whole process of corn-soybean strip compound planting. *J. Xihua. Univ.* 39, 91–97. doi: 10.12198/j.issn.1673-159X.3742

Zhang, R. F., Zhou, J. L., Liu, H., Shi, S., Wei, G., and He, T. F. (2022). Determination of interspecific contact parameters of corn and simulation calibration of discrete element. *Trans. Chin. Soc Agric. Mach.* 53, 69–77. doi: 10.6041/j.issn.1000-1298.2022.S1.008

Zhao, P. F., Gao, X. J., Su, Y., Xu, Y., and Huang, Y. X. (2024). Investigation of seeding performance of a novel high-speed precision seed metering device based on numerical simulation and high-speed camera. *Comput. Electron. Agric.* 217, 108563. doi: 10.1016/j.compag.2023.108563

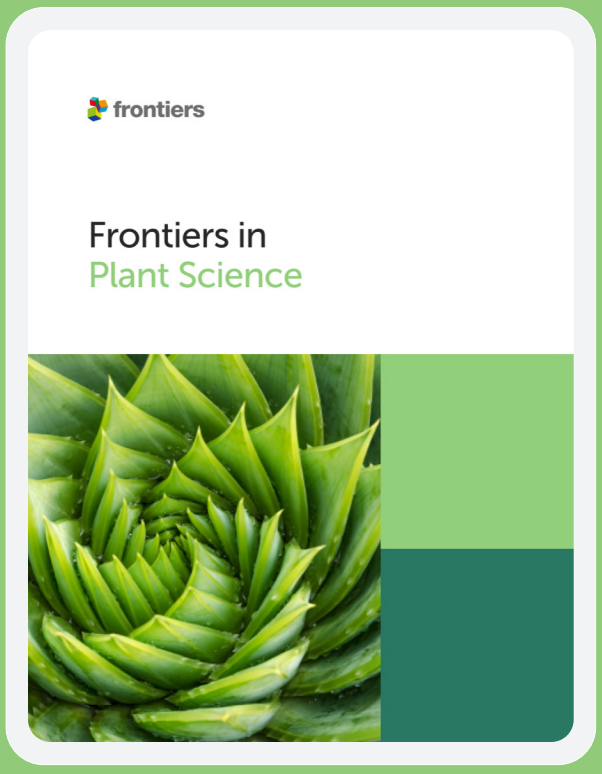
Frontiers in
Plant Science

Cultivates the science of plant biology and its applications

The most cited plant science journal, which advances our understanding of plant biology for sustainable food security, functional ecosystems and human health.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact