Computers and Electronics in Agriculture 217 (2024) 108551

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

NIR-hyperspectral imaging and machine learning for non-invasive chemotype classification in *Cannabis sativa* L

M. San Nicolas[a,b,c,*], A. Villate[a,b], I. Alvarez-Mora[a,b], M. Olivares[a,b], O. Aizpurua-Olaizola[c], A. Usobiaga[a,b], J.M. Amigo[a,d]

[a] *Department of Analytical Chemistry, Faculty of Science and Technology, University of the Basque Country (UPV/EHU), 48940 Leioa, Basque County, Spain*
[b] *Research Centre for Experimental Marine Biology and Biotechnology (PIE), University of the Basque Country (UPV/EHU), 48620 Plentzia, Basque County, Spain*
[c] *Sovereign Fields S.L., 20006 San Sebastian, Basque County, Spain*
[d] *IKERBASQUE, Basque Foundation for Science, Plaza Euskadi 5, 48009 Bilbao, Spain*

ARTICLE INFO

Keywords:
Non-destructive analysis
Hemp
Cannabinoids
Quality control
PLS-DA
External validation

ABSTRACT

The current public acceptance rate towards medical cannabis feasibility has led to a worldwide increase in this plant species production. Nevertheless, the currently transforming legal framework does not prevent the originally unlawful knowledge around cannabis breeding, which lacks quality control regulations or standards for correct manufacturing processes, a fact that could subsequently lead to uncontrolled and even harmful crop products. In this line, the objective of this work was to develop a non-invasive methodology for cannabis chemotype classification in different cultivars during the plant cultivation process, in order to keep undoubtful production control over cannabis crops. Hence, hyperspectral imaging (HSI), coupled with various multivariate data analysis approaches, such as principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA), enabled the non-invasive in-situ analysis of the plants. Hence, two PLS-DA classification models were trained with the plant spectral data for three chemotypes, based on the cannabinoid content of the plant inflorescences, with the difference between both approaches being the regard of the stem part of the plant as a bias. Thus, obtained sensitivity and specificity values in the inflorescences were 0.845/0.845 for Chemotype I, 0.954/0.920 for Chemotype II, and 0.888/0.925 for Chemotype III. At last, a hierarchical PLS-DA, which considered the stem as a bias, presented an overall 94.7 % trueness in the external validation of 57 different plant individuals, divided as 92.3 % trueness for chemotype I, 100.0 % trueness for chemotype II and 88.9 % trueness for chemotype III. Based on these results, the proof of concept for comprehensive agricultural control of cannabis crops through a non-invasive analytical technique was demonstrated, a previously unproven fact. Therefore, this work could further pave the way for non-invasive technology development for horticultural quality control in medical cannabis productions, as this emerging industry will require strict control over the cannabis chemotypes, with the strong advantage of avoiding destructive and time-consuming analytical techniques such as chromatography.

1. Introduction

Over the last few years, the trend of institutional acceptance towards the potential use of cannabis as a medical alternative to synthetic drugs for diverse illnesses or palliative pain treatment has undergone favourable progress (Cristino et al., 2020; Black, 2019; Cox-Georgian et al., 2019; Aizpurua-Olaizola et al., 2017). Examples of this are the authorisation of medical cannabis use in 38 US states through the approval of the Medical Marijuana Laws (MML) (‘State Medical Cannabis Laws’. Accessed: Jun. 02, 2023) or the declassification of cannabis and cannabis resin from the Schedule IV of the Single Convention of Narcotic Drugs (Recommendation of removal for cannabis and cannabis resin from Schedule IV of, 1961), thus recognising its therapeutic potential. As a result, various European countries have currently adopted favourable positionings towards the therapeutic application of cannabis, such as Germany or Switzerland (Abgabe, 2022; Act, 2022). This interest in the adoption of cannabis-based products in medicine mainly lies in the cannabinoids, a nearly exclusive class of terpenoid bioactive compounds

* Corresponding author at: Barrio Sarriena s/n, 48940 Leioa, Vizcaya, Spain.
E-mail address: markel.sannicolas@ehu.eus (M. San Nicolas).

https://doi.org/10.1016/j.compag.2023.108551
Received 29 August 2023; Received in revised form 17 November 2023; Accepted 13 December 2023
Available online 23 December 2023

present in this plant species that interact with the human endocannabinoid system (ECS) (Booth and Bohlmann, 2019; Maccarrone, 2015; Di Marzo, 2018). The ECS is part of the human nervous system that plays an important physiological role, as it is involved in various processes such as brain plasticity regulation, neuronal development, energy balance or appetite regulation, among others (Aizpurua-Olaizola et al., 2017). However, within the class of cannabinoids, different compounds are biosynthesised in the cannabis plant, and each of them interacts differently with the receptors of the ECS, thus, possibly leading to different medical outcomes (Muller et al., 2019; Bonini, 2018; Zou and Kumar, 2018). This is the reason why the growth in the interest in medical cannabis has highlighted the need for accurate and efficient methods for ensuring quality control compliance in the plant production process.

Currently, two cannabinoids are commonly acknowledged to be potential active pharmaceutical ingredients (API), which are also the ones that appear in higher concentrations in cannabis inflorescences (Aizpurua-Olaizola, 2016): Δ^9-Tetrahydrocannabinol (THC) and cannabidiol (CBD) (Gülck and Møller, 2020; Fraguas-Sánchez and Torres-Suárez, Nov. 2018). Therefore, since their discovery, cannabis strains have been cross-bred following the Mendelian genetics laws, in order to find resulting breeds with optimal THC/CBD ratios that improve therapeutic efficacy and safety. As a consequence, countless cultivars can be now found within the cannabis species, but most of them can be classified into three chemotypes, depending on the content of these major cannabinoids (Small and Beckstead, 1973; Lewis et al., 2018): chemotype I, which can be defined as THC predominant; chemotype II, which possesses THC and CBD levelled contents; and chemotype III, which is predominant in CBD (Aizpurua-Olaizola, 2016). Therefore, in medical cannabis, breeds must be correctly classified through accurate cannabinoid quantification, which traditionally has been mostly done by liquid chromatography coupled to diode array detector (LC-DAD), liquid chromatography and mass spectrometry (LC-MS) or gas chromatography coupled to flame ionisation detector (GC-FID) (Aizpurua-Olaizola et al., 2014; San Nicolas, 2020; Chandra et al., 2019; Citti et al., 2018; Berman, et al., 2018; Citti et al., 2018; San Nicolas, et al., 1279). This means that the process involves destructive sampling, laborious sample preparation or costly analytical techniques, which are time-consuming and resource intensive. Moreover, in recent years, near-infrared spectroscopy (NIRS) has gained general appreciation for the analysis of cannabinoids, due to less labour-intensive and costly procedures, but without compromising precision and accuracy (Sánchez-Carnerero Callado et al., 2018; Duchateau et al., 2020; Deidda, 2021; Espel Grekopoulos, 2019; Su, 2022; Yao et al., 2022). Nevertheless, the presence of moisture in the plant samples can be a procedural limitation when using NIRS, as the chemical bonds of the water molecule cause significant banding in the NIR region. Therefore, it could be stated that, overall, moisture can cause either quantitative or qualitative interference in the analysis of cannabinoids by this method. Due to this fact, the European Union (EU) delegated a commission regulation (N° 2017/1155) which described an experimental procedure involving drying as pre-treatment of hemp samples for the analysis of cannabinoids (Regulation, 2017), a procedure that is generally followed (Sánchez-Carnerero Callado et al., 2018; Duchateau et al., 2020; Su, 2022; Yao et al., 2022; Jarén et al., 2022; Valinger, 2021). Therefore, although NIRS requires an easier sample pre-treatment than chromatographic analysis, it does demand an invasive experimental procedure on the cannabis plant for a reliable analysis of cannabinoids.

In this line, hyperspectral imaging (HSI) appears as a potential non-invasive analytical technique for chemotype determination in cannabis, which combines the advantages of imaging and spectroscopy for capturing both spatial and spectral information from the plant (Amigo, 2019). Through optical imaging, HSI enables two-dimensional object visualisation as a normal image, whereas a wide electromagnetic spectrum from each pixel is retrieved. Therefore, instead of just capturing the primary colours (red, green, blue) from each pixel, the wide spectrum from the pixels is broken down into spectral bands, and deeper information than what can be observed at first glance can be retrieved (Amigo et al., 2015). Hence, NIR-HSI not only enables the determination of the chemical composition of a sample by spectroscopic means but also permits the two-dimensional visualisation of its distribution throughout the sample surface.

Thus, a hyperspectral image can be observed as a three-dimensional array which has two spatial dimensions, divided into pixels (x and y), and one spectral dimension (λ) (Fig. 1) (Amigo et al., 2015), which contains all the chemical information of the measured plant. This is why hyperspectral images usually present multicomponent mixture dependence and they rarely contain selective spectral variables related to specific components (Amigo et al., 2015). For this reason, hyperspectral images often exhibit multicomponent mixture dependence, and they rarely contain selective spectral variables related to specific components; however, this is the reason why HSI may deal with selective variables related to moisture bands that cause interference in NIRS analysis, enabling a non-invasive analysis in a living plant. Moreover, hyperspectral images can also contain spatial or spectral noise and redundant data, so the data must be adequately treated for the correct determination of the sought conclusions. In this line, machine learning approaches can fill this gap and handle multivariate challenges extracting meaningful patterns from hyperspectral data, in order to provide automated and efficient chemotype classification for quality control compliance at the plant production site, allowing rapid and non-destructive analysis of cannabis.

In addition, it is also worth the ad hoc mentioning of some recent works focused on the quantification of the main cannabinoids (THC and CBD, alongside their THCA and CBDA acidic conjugates) by NIR-HSI and other machine learning approaches. In this line, in *S. K. Abeysekera et al.* (Abeysekera, et al., 2023) it was proved that NIR-HSI technology is a valuable tool for the accurate estimation of THCA content for high-throughput phenotyping of cannabis. This was also the conclusion of *W. S. Holmes et al.* (Holmes, 2023) regarding the quantification of CBDA content in inflorescences and leaves through this technique. Also in the case of *Y. Lu et al.* (Lu et al., 2022), it was determined that NIR-HSI, as opposed to conventional analytical methods, is a potentially useful tool for non-destructive, rapid quantification of the stated major cannabinoids in floral material of cannabis. Thus, recent empirical studies are demonstrating the potential of this technology in cannabis for various objectives, simplifying the procedure without compromising analytical capabilities. For this reason, innovative non-invasive techniques such as NIR-HSI should be more frequently introduced in crop quality control, seeking to reduce field and laboratory workloads.

Hence, the objective of this study was to determine the feasibility of HSI, coupled with machine learning approaches, for non-invasive chemotype classification in cannabis. The scope is the development of an accurate and reliable method that could streamline the analysis process in-situ, excluding the need for any sample preprocessing for chemotype classification in cannabis, thus, facilitating quality control in the medical cannabis industry. The proposed methodology holds great promise for improving cultivation practices and enabling well-informed decision-making in cannabis-related research and production.

2. Materials and methods

2.1. Cannabis plants cultivation

The cannabis plants were cultivated in the facilities of Sovereign Fields S.L. (Larramendi 3 str., Donostia-San Sebastian 20006, Spain). Plants were grown in 11 L black pots containing a soil/hummus/nutrient mixture. Specifically, the mixture consisted of 80 % of Light mix soil of Biobizz Worldwide S.L. (Lezama-Leguizamon industrial park, Gorbeia 11 str., Etxebarri, Spain), 20 % of hummus, and 10 g/L of farmer mix nutrient solution by Lurpe Natural Solutions (Zubiate 3 str., Lemoa 48330, Spain), which is composed of bat guano, bone meal, kelp meal, Azomite®, organic alfalfa, insect frass, blood meal, dolomite,

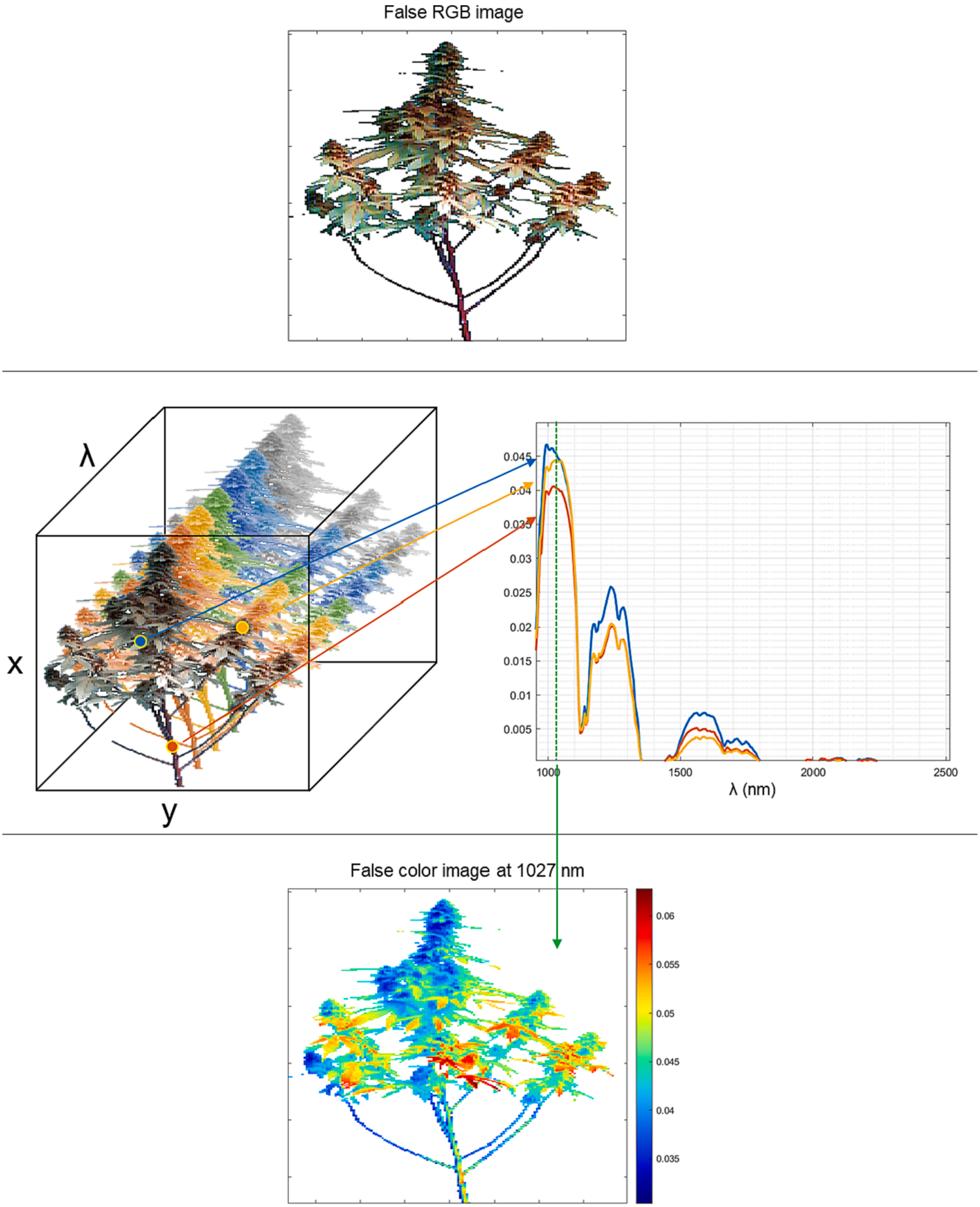

Fig. 1. NIR-Hyperspectral image visualisation of a cannabis plant captured in the 930–2500 nm wavelength range.

langbeinite humic and fulvic acids, and a complex blend of *rhizobacteria* and *trichoderma*. The total cultivation time was 12 weeks: the first 4 weeks corresponded to the vegetative stage, the period during which the plants grew; and the next 8 weeks were the flowering stage when inflorescences developed. The photoperiod regime defined the vegetative and flowering stages: At 18 h light/6h dark for the vegetative stage, it changed to 12 h light/12 h dark for the flowering stage. NIR-Hyperspectral images of the plants were acquired in their tenth cultivation week.

NIR-hyperspectral images of 57 plant individuals were taken for this work, divided into different varieties: 10 Dairy Queen (DQ), 9 Futura 75 (FT), 10 Remedy (RE), 8 Roma (RO), 10 Tel Aviv (TA) and 10 White Widow (WW) individuals. Each one of the individuals was enumerated.

2.2. Analysis of cannabinoids

As a reference method for the definition of each variety's chemotype, the corresponding cannabinoid content was determined in inflorescences of the stated cultivars using Liquid Chromatography with Diode Array Detector (LC-DAD) according to Aizpurua-Olaizola et al. (Aizpurua-Olaizola, 2016) in the Sovereign Fields S.L. facilities. It was assessed that Tel Aviv, Roma and Dairy Queen varieties belonged to Chemotype I ($C_{Total\ THC}/C_{Total\ CBD} > 10$), Remedy and White Widow varieties belonged to Chemotype II ($0.3 < C_{Total\ THC}/C_{Total\ CBD} < 3$), and Futura 75 variety belonged to Chemotype III ($C_{Total\ THC}/C_{Total\ CBD} < 0.1$) (Small and Beckstead, 1973).

2.3. Nir-hyperspectral image acquisition

Images were taken in situ, using a HySpex SWIR 384 (HySpex by NEO, Østensjøveien 34, N-0667 Oslo, Norway) hyperspectral camera with the field setup, under sunlight illumination. Images were obtained on the same summer day (02/07/2021), between 09:00 and 14:00.

The spectral range of the hyperspectral camera was set between 930 nm and 2500 nm, divided into 288 spectral channels with a 5.45 nm spectral resolution, and the spatial resolution was 384 x 286 pixels. The camera was equipped with the 3 m distance lens and the rotation angle of the rotor of the tripod of the field setup was set at 16°, which provided a linear field of view (FOV) of 841 mm, resulting in a pixel size across-/along the track of 2.19/2.19 mm.

For the acquisition of the images, the corresponding plant individual was placed in front of a white wall inside the greenhouse, at the facilities of Sovereign Fields S.L. Then, the tripod was placed at a distance of 3 m from the plant pot, perpendicularly to the wall. As the rotation of the tripod rotor moved from left to right according to the established rotation angle (16°), the plant was aligned so that it was centered in the linear FOV. Once all the parameters were set, the image was acquired with the ENVI® hyperspectral image processing software (NV5 Geospatial Solutions, Inc., Broomfields, Colorado, USA) and then radiometrically calibrated. Images of one plant individual at a time were taken.

2.4. Hyperspectral image analysis

NIR-hyperspectral images were imported into MATLAB environment (The MathWorks Inc.) and handled with HYPER-Tools 3.0 (freely available at https://www.hypertools.org) (Mobaraki and Amigo, 2018).

A total of 502 flower spectra were extracted from the NIR-hyperspectral images by manual flower-pixel picking: 249 of chemotype I, 178 of chemotype II and 75 of chemotype III. In addition, a cumulative number of 219 stem spectra were as well extracted from all images by manually picking pixels from the plant stem regions, as it can be observed in Fig. 2. Retrieved spectra were preprocessed with Savitzky-Golay derivative (Window width 7; Polynomial order 2; Derivative order 1) (Fig. 2) and mean-centered for the classification model training.

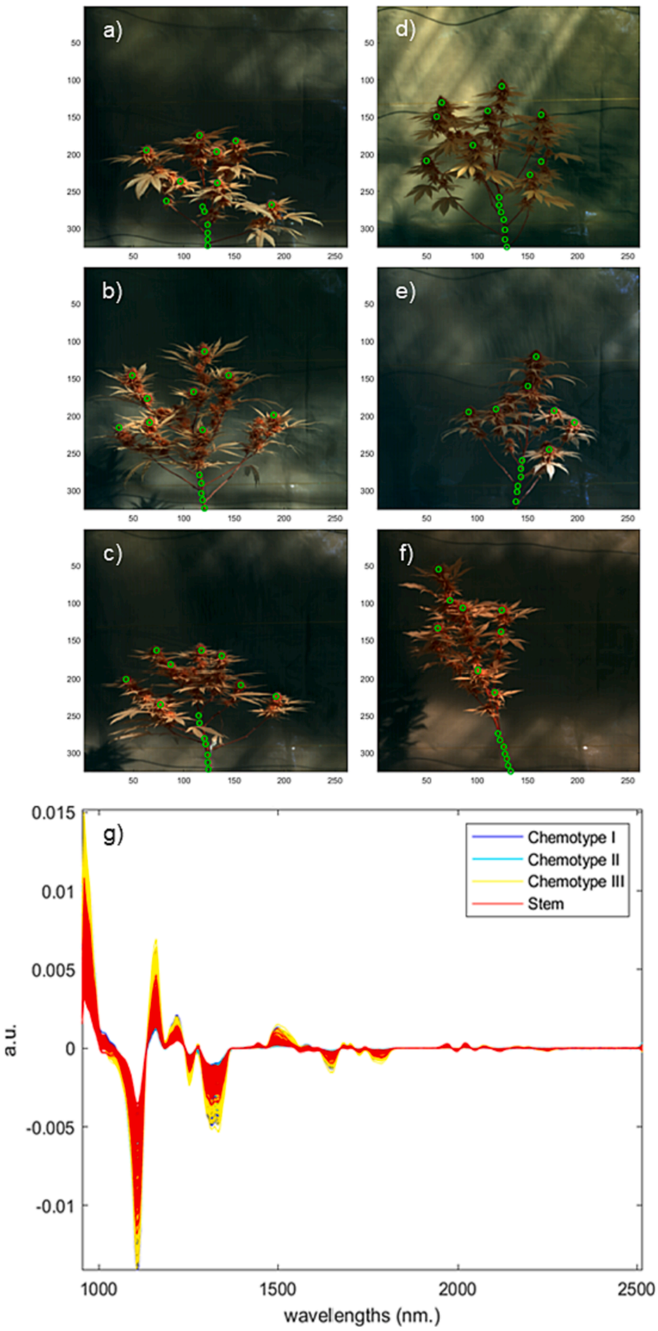

Fig. 2. Manual flower-pixel and stem-pixel picking in the hyperspectral images of various plant individuals for spectra retrieving to train the corresponding classification model (a) DQ5 individual (b) TA6 individual (c) RO6 individual (d) RE2 individual (e) WW5 individual (f) FT9 individual g) SWIR spectra of the flower-pixels and stem-pixels used for the classification model training after Savitzky-Golay derivative.

For correct chemotype prediction and appropriate data visualisation, the background was eliminated for each plant. For doing so, Principal Component Analysis (PCA) was performed on each image, after mean centering (Pearson, Nov. 1901; Hotelling, 1933). Scores of PC2 provided morphological fit of the plant individuals, focused on inflorescences, allowing the irregular background removal from the images, as shown in Fig. 3.

Two classification strategies based on Partial Least Squares-Discriminant Analysis (PLS-DA) were proposed in this work (Ståhle and Wold, 1987). The first approach was conducted by directly training the calibration model with the 502 flower spectra of the three

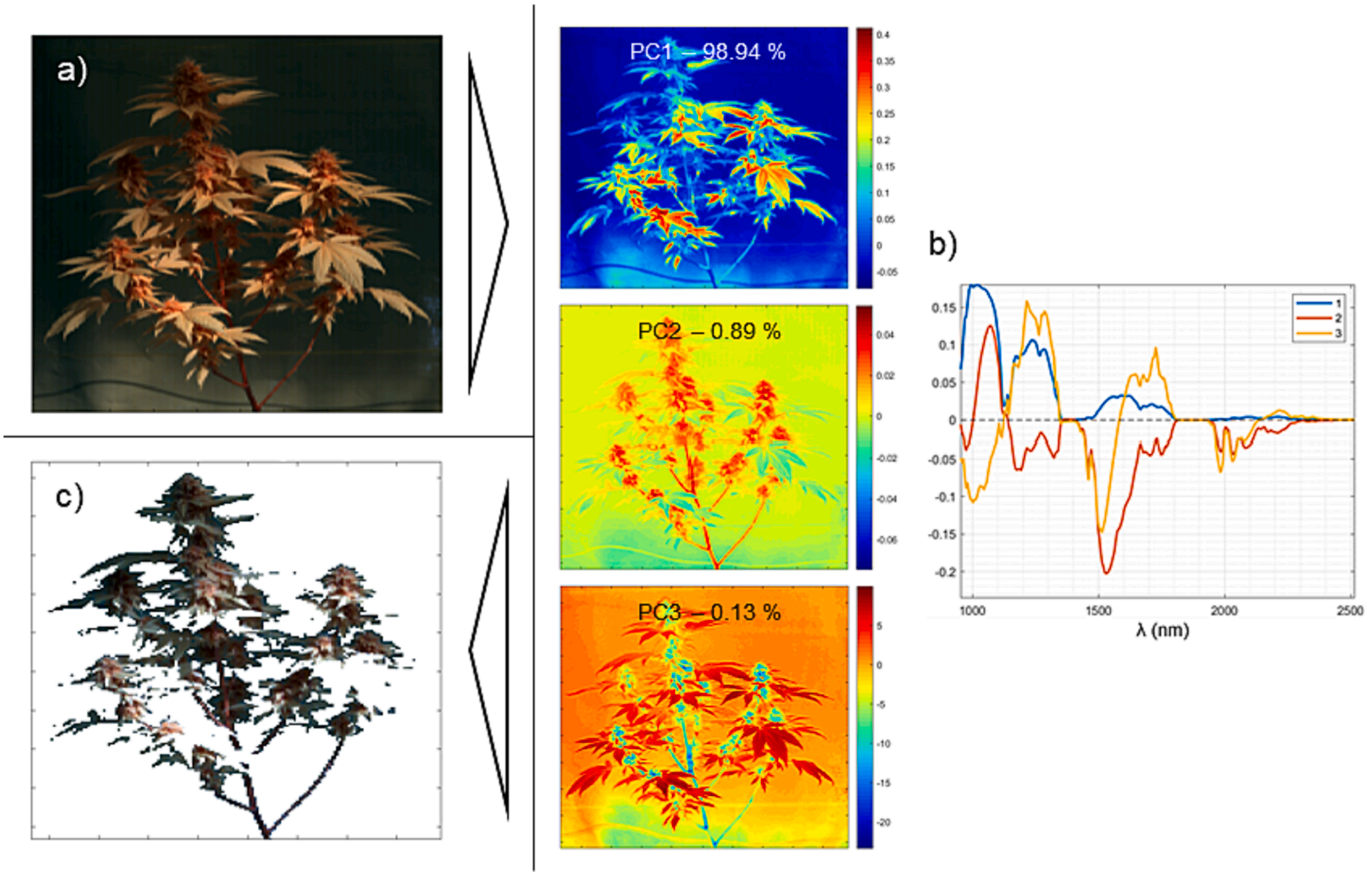

Fig. 3. Morphological masking on Dairy Queen 4 plant individual through PCA (a) False RGB image of the plant (b) Scores and Loadings of PC1, PC2 and PC3 (c) Masked false RGB image of the plant individual.

chemotypes. On the other hand, the second strategy was based on a two-layer hierarchical model: in the first layer, the stem was classified from the rest of the plant using the 502 flower spectra and 219 stem spectra; in the second level, the non-stem part of the plant was classified as the corresponding chemotype using the 502 flower spectra. The models, developed using the PLS_Toolbox 9.0 (Eigenvector Reseach, WA, USA), were cross-validated with random subsets, with 10 data splits and 5 iterations.

The pixels not used in the training of the classification models were predicted with these in every hyperspectral image.

3. Results and discussion

3.1. Major cannabinoid characterisation

The chemotype of each cultivar was defined through destructive analysis of cannabinoids in their pooled dry inflorescences. Corresponding major cannabinoid concentrations are shown in Table 1, from which Dairy Queen, Tel Aviv and Roma cultivars were defined as chemotype I belonging; White Widow and Remedy were defined as chemotype II; and Futura 75 was defined as chemotype III.

Table 1
THC and CBD concentrations in different cannabis cultivars' dry inflorescences determined by HPLC-DAD in and their respective chemotype.

	Cannabinoid concentration (w/w. %)			
Variety	Total THC	Total CBD	$C_{Total\ THC}/C_{Total\ CBD}$	Class
White Widow	5.6	7.5	0.75	Chemotype II
Dairy Queen	20.2	0.7	28.86	Chemotype I
Tel Aviv	16.9	0.2	84,50	Chemotype I
Roma	17.3	0.2	86,50	Chemotype I
Futura 75	0.7	15.6	0.04	Chemotype III
Remedy	6.8	8.8	0.77	Chemotype II

3.2. PLS-DA model performance

3.2.1. PLS-DA model training

The classification model for the three chemotypes was trained using 8 Latent Variables. Cross-validation of the model was performed through random subsampling, since the three defined classes possessed different statistical weights in the model, with 10 data splits permutated in 5 iterations. Cross-validated optimised sensitivity and specificity values were the following for each class: 0.845/0.845 for Chemotype I, 0.954/0.920 for Chemotype II, and 0.888/0.925 for Chemotype III. ROC curves for each class are shown in Fig. 4.

To ensure the reliability of the model, a 100-iteration permutation test was performed with the data used for the model training, to test the statistical significance of the model. This way, it is determined whether the predictive capacity of the model is overfitted, if it is product of chance or if it is product of the corresponding sample-variable correlations (Lopez et al., Sep. 2023). The probability of chance of the model was tested through the Wilcoxon test (Wilcoxon, 1945), where both the self-predicted and cross-validated result happened to be 0 for each one of the chemotype classes. Thus, it was determined that the model is statistically robust and significant, meaning neither cross-validated parameters of the model nor its predictive capacity are affected by random factors or product of chance, therefore implying that every prediction result would be the truthful product of the correlation between the variables and the classes of the data.

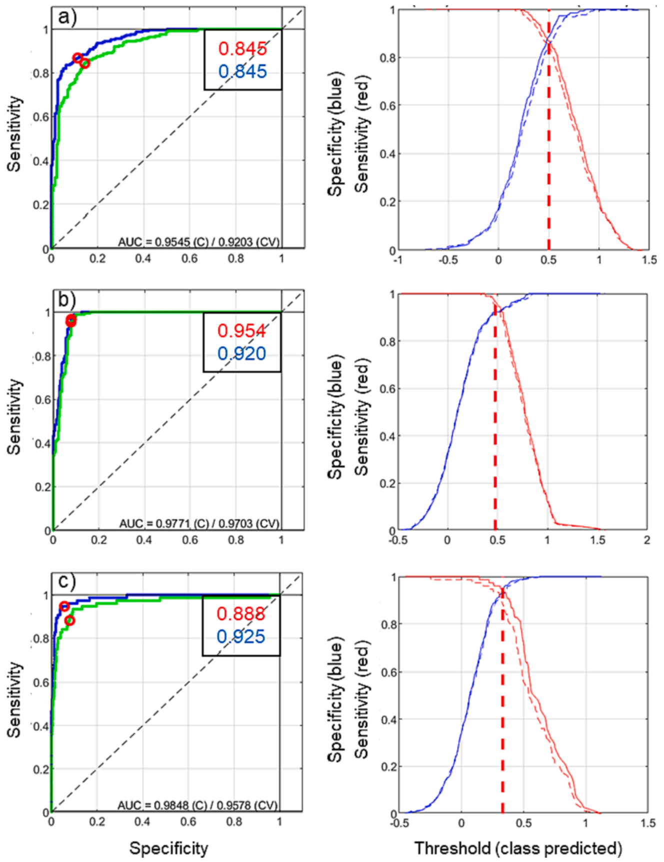

Fig. 4. Estimated (blue) and Cross-Validated (green) receiving operating characteristic (ROC) curves and estimated (solid) and cross-validated (dashed) response curves (a) Chemotype I (b) Chemotype II (c) Chemotype III. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2.2. PLS-DA model external validation

The predicted chemotype probabilities for each plant are presented in *Table S1* of Supplementary Material (SM). The minimum threshold for predicting the corresponding chemotype in each pixel of the images was set at 50 % probability. Thus, for the external prediction of the chemotype of a plant individual, a restriction rule was set, stating that at least 50 % of the pixels that constitute an image must be classified in one of the chemotypes. According to this statement, the predicted chemotypes on the true plant individual classes are shown in Table 2.

From the results presented in Table 2 it could be acknowledged that 4 individuals were incorrectly classified (7.02 % misclassification ratio). In comparison, the other two individuals were not classified in any of the classes (3.51 % non-classification ratio), so 51 out of 57 plant individuals were correctly classified (89.47 % trueness). In this line, it is worth highlighting that every individual belonging to chemotype II was correctly classified (100 % trueness). The best results were expected to correspond to this class, as it showed the best overall sensitivity and specificity values in the model training (0.954 and 0.920, respectively). Furthermore, the volume of analysed chemotype I individuals was the largest between the three classes (n = 28). This class was composed of three different cultivars, which could lead to a higher misclassification probability due to greater biological variability among individuals. This could be why, out of the 28 individuals, there were 25 individuals correctly classified (89.29 % trueness). Finally, the model performed worst in predicting chemotype III plants. In this case, only 6 out of 9 individuals were correctly classified (66.67 % trueness). In general terms, this result would not be adequate for the development of a representative classification model. However, this fact could result from a bias in its predictive capacity. Taking a closer look at the predictions, it could be noted that, overall, the stem part of the plants was predicted as chemotype I, regardless of the class each plant belonged to. Indeed, this bias seems more noticeable in chemotype III plants, where pixels predicted as chemotype I comprise a notoriously large portion of the images. Therefore, the model was hierarchised at two classification levels to leverage the stem area of the plants in the images. This modification was expected to remove this prediction bias, so a model with better predictive capacity could be achieved.

3.3. HPLS-DA model performance

3.3.1. HPLS-DA model training

The first classification level of the HPLS-DA consisted of a classification between pixels corresponding to the stem and non-stem part. This model showed cross-validated sensitivity and specificity values of 0.986 and 0.995 for the non-stem part of the plant, respectively, and 0.995 and 0.986 for the case of the stem part, with 4 Latent Variables (LV) (see Fig. 5). The second classification level of the HPLS-DA corresponded to

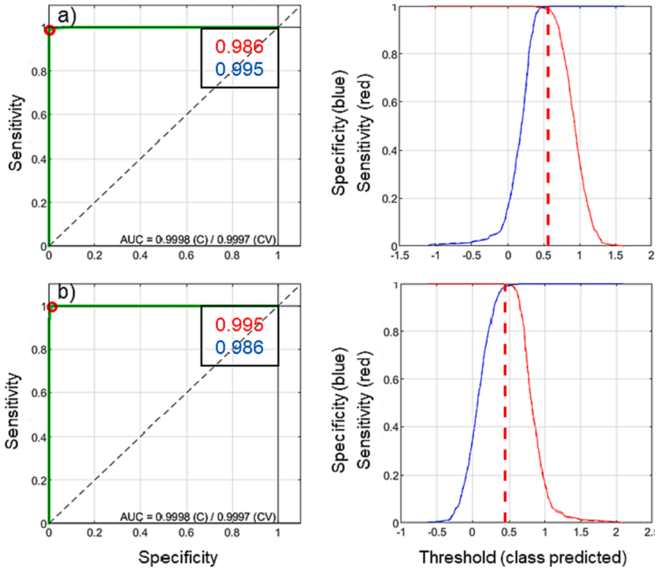

Fig. 5. Estimated (blue) and cross-validated (green) ROC curves and estimated (solid) and cross-validated (dashed) response curves (a) Non-stem class (b) Stem class. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
True classes and predicted classes of 57 plant individuals of 6 different varieties belonging to three chemotypes through trained PLS-DA classification model.

		Predicted class (Chemotype Probability > 50 %)				
		Chem. I	Chem. II	Chem. III	Not classified	
True class	Chemotype I (n = 28)	25	2	0	1	28
	Chemotype II (n = 20)	0	20	0	0	20
	Chemotype III (n = 9)	2	0	6	1	9
		27	22	6		Sum = 57

the previously trained PLS-DA (section 3.2.).

3.3.2. HPLS-DA model external validation

The predicted chemotype probabilities for each plant individual through HPLS-DA are presented in *Table S2* of SM. The minimum threshold for predicting the corresponding chemotype in each pixel was also set at 50 % probability, as in the previous PLS-DA predictions. Nevertheless, in this case, for chemotype prediction in a plant, the leverage of stem pixels among the total pixels constituting each image was assessed, as either the morphology or size of the plant varied between individuals and varieties. Hence, a restriction rule was established in the second classification level, stating that at least 50 % of the leaf pixels must be classified in one of the three chemotypes. By this statement, the predicted chemotypes on the true plant individual classes are shown in Table 3.

According to the results presented in Table 3, no individual was unclassified in chemotype prediction through HPLS-DA, which indicates a classification with a greater probability than 50 % in one of the chemotypes in every plant individual (0.00 % non-classification ratio). However, 5 out of 57 individuals were incorrectly classified (91.23 % trueness, 8.77 % misclassification ratio). As in the case of PLS-DA, all chemotype II individuals were correctly classified (100 % trueness), but in the case of chemotype III plants, the results improved, with correct classification of 8 out of 9 individuals (88.89 % trueness). Considering the number of chemotype III plant individuals, this was a remarkable improvement over the PLS-DA results (22.22 % improvement). Nonetheless, the trueness ratio of the chemotype I plant was downgraded (3.58 % downgrade), as 24 out of 28 individuals were correctly classified (85.71 % trueness).

3.4. Comparison between classification models

Comparing the PLS-DA and HPLS-DA models overall, it could be stated that the hierarchical model offered better results than the first one, shown in Table 4, as no individual was left unclassified. Some of the visual prediction examples can be observed in Fig. 6, where it can be observed that, for instance, Futura75 no. 2 individual, which belongs to chemotype I, was classified as chemotype I through the PLS-DA model, whereas it was correctly classified using the HPLS-DA model. This later one enhanced overall predictive capacity, as trueness for the chemotype III class was improved due to the removal of the stem bias. Consequently, the achieved results were closer to true classes.

3.5. Proof of concept confirmation

In both model cases, two certain individuals were undoubtfully misclassified. Those individuals were Tel Aviv 3 and Tel Aviv 4, theoretically belonging to chemotype I, which has been undoubtedly classified to chemotype II. In the case of the Tel Aviv 3 individual, the probability results obtained by PLS-DA model were 8.05 % for chemotype I, 91.95 % for chemotype II and 0.00 % for chemotype III, while through HPLS-DA, prediction probabilities of 1.33 % for chemotype I, 98.67 % for chemotype II and 0.00 % for chemotype III were obtained. In the case of the Tel Aviv 4 individuals, the prediction results were similar. PLS-DA provided resulting probabilities of 9.02 % for chemotype I, 90.98 % for chemotype II and 0.00 % for chemotype III, while using HPLS-DA, these values were 0.09 % for chemotype I, 99.91 % for chemotype II and 0.00 % for chemotype III. Both individuals were significantly classified as chemotype II belonging and, as unlikely as it may appear, this result could be correct, derived from a plant labelling error in the plant cultivation facilities. This fact was later confirmed, since, at the time of the image acquisition, some White Widow individuals (chemotype II) were incorrectly labelled and lost control over their tracking. Therefore, to ascertain the comparability of these precise individuals with the other Tel Aviv and White Widow individuals, a *Student t*-test was applied between the prediction results of these groups (Student, 1908). The statistic *t*-tests were calculated assuming equal variances between the two groups. On the one hand, the belonging of Tel Aviv 3 and Tel Aviv 4 individuals to Tel Aviv variety was proposed as a null hypothesis, thus comparing chemotype I resulting in predictions in the corresponding individuals of both groups. On the other hand, the resemblance of Tel Aviv 3 and Tel Aviv 4 individuals to White Widow cultivar was proposed as a null hypothesis, so predicted probabilities for chemotype II were compared between the stated groups. Both calculations, shown in Table 5, were done with predictions resulting from the HPLS-DA model.

Table 4
Classification model predictive capacity comparison.

	PLS-DA (3 classes)	HPLS-DA (2 levels)
Overall trueness	89.47 %	91.23 %
Chemotype I	89. 29 %	85.71 %
Chemotype II	100.00 %	100.00 %
Chemotype III	66.67 %	88.89 %
Misclassification ratio	7.02 %	8.77 %
Non-classification ratio	3.51 %	0.00 %

The results of the statistical *t*-test rejected the null hypothesis of TA3 and TA4 individuals belonging to T.A. variety, as the calculated *t* resulted to be −9.18 (*t* Critical for one-tail = 1.86), while it accepted their similarity to W.W. variety (calculated *t* = 1.43; *t* Critical for one-tail = 1.81). Being this so, if these individuals were to be correctly relabeled as W.W. cultivar belonging, the HPLS-DA prediction results would considerably improve to 92.31 % trueness for chemotype I, 100.00 % trueness for chemotype II and 88.89 % trueness for chemotype III, alongside 5.26 % misclassification ratio, as just 3 individuals would be incorrectly predicted. Therefore, the overall prediction trueness of the HPLS-DA would be 94.74 %, which would give this classification model an accurate predictive capacity. Accepting this hypothesis, the final average predicted probabilities for each chemotype would be those shown in Table 6. According to those results, the classification performance for chemotype II plants is excellent, showing a predicted average probability of 96.63 % with 5.59 % precision. In the other cases, the predicted average probability and precision results were 67.85 % and 15.10 % for chemotype I plants, and 61.75 % and 17.82 %, respectively, for chemotype III plants. This fact could occur due to chemotype II plants containing both major cannabinoids in significant contents. In contrast, either of the other two classes only contains one of them (*section 3.1*), so it could be deduced that the presence of both cannabinoids at significant concentrations considerably enables representative chemotype classification by NIR-hyperspectral imaging in cannabis cultivars.

Table 3
True classes and predicted classes of 57 plant individuals of 6 different varieties belonging to three chemotypes through trained HPLS-DA classification model.

		Predicted class (Chemotype Probability > 50 %)				
		Chem. I	Chem. II	Chem. III	Not classified	
True class	Chemotype I (n = 28)	24	4	0	0	28
	Chemotype II (n = 20)	0	20	0	0	20
	Chemotype III (n = 9)	1	0	8	0	9
		25	24	8		Sum = 57

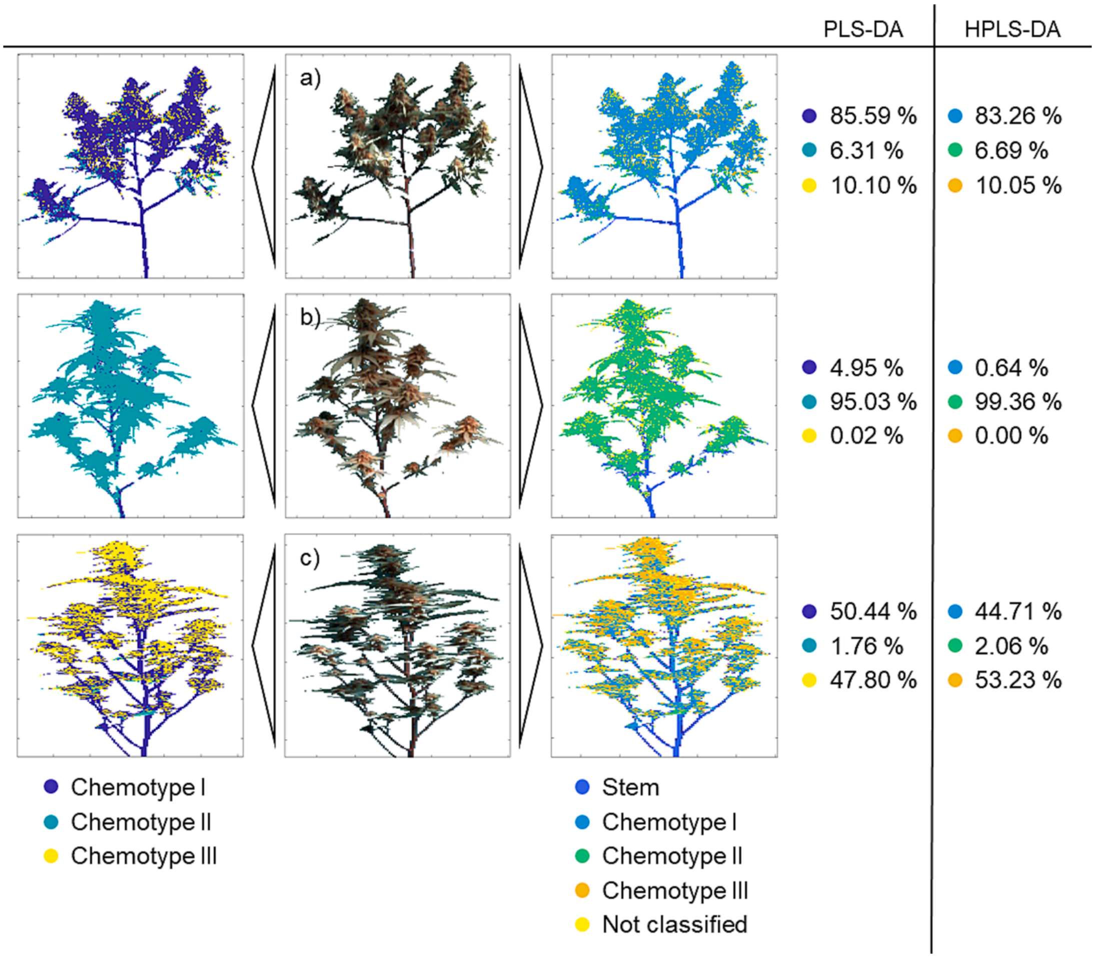

Fig. 6. PLS-DA (left) and HPLS-DA (right) prediction examples in different plants: (a) Roma no. 3 individual (Chemotype I) (b) White Widow no. 10 individual (Chemotype II) (c) Futura75 no. 2 individual (Chemotype III).

Table 5
Statistic t-tests between TA3 and TA4 individuals with the rest of T.A. group and TA3 and TA4 individuals with the W.W. group.

	t-Test: Two-groups assuming equal variances (α = 0.05)			
	Chemotype I		Chemotype II	
	TA3 and TA4	Rest of TAs	TA3 and TA4	WWs
Mean	0.71	69.66	99.29	92.80
Variance	0.77	103.10	0.77	38.01
Observations	2	8	2	10
Pooled Variance	90.31		34.29	
P(T ≤ t) two-tail	1.61E-05		0.18	
t Critical two-tail	2.31		2.23	

Table 6
Predicted average probability, standard deviation and precision in each of the three chemotypes, accepting the hypothesis of TA3 and TA4 individuals belonging to WW variety.

	Chemotype I (%)	Chemotype II (%)	Chemotype III (%)
Trueness	92.31	100.00	88.89
Predicted average probability	67.85	96.63	61.75
Standard Deviation	10.24	5.40	11.01
Precision (RSD)	15.10	5.59	17.82

*RSD: Relative standard deviation

4. Conclusion

This work concluded that, through NIR-hyperspectral image analysis of *Cannabis sativa* L., different cultivars belonging to chemotypes I, II and III could be representatively classified, avoiding invasive analytical techniques. This was made possible due to a two-level hierarchical classification model in cannabis plants: on the first level stem and non-stem parts of the plants could be classified, while on the second level plants were classified in one of the three chemotypes based on SWIR spectral information. According to the HPLS-DA model, the results showed that plants belonging to chemotype II were perfectly classified, as in the cases of chemotype I and chemotype III, 1 and 2 individuals, respectively, were misclassified. This translates into an overall classification trueness of 94.74 %, a correct proof of concept classification performance. Moreover, as formerly stated, TA3 and TA4 individuals, theoretically belonging to chemotype I, were representatively classified to chemotype II. Even though it was an anomaly, it was later confirmed that, during the cultivation time, a mislabelling error happened in the cultivation facilities, as some individuals belonging to White Widow variety (chemotype II) were incorrectly tagged. Rather than worsening the results, this error empirically demonstrated the purpose of this work, as in-field analysis enabled procedural control over the plants avoiding invasive characterisation of cannabinoids in plant inflorescences. Nevertheless, it is also true that, in any other scenario, this could be a potential source of error in the data, so it is imperative that adequate traceability is kept towards the identification of the plant individuals for future field work involving larger sample sizes, with rigorous labelling and tracking procedures. Moreover, it would also be beneficial to assess and demonstrate the classification model's performance across different cultivation environments, apart from the specific conditions of the facilities of Sovereign Fields S.L., so as to ensure the transferability of the results in different growth sites.

Therefore, the methodology based on NIR-HSI with a HPLS-DA for chemotype classification properly dealt with the main handicap of the analysis of cannabinoids by NIRS, which was the moisture present in the fresh plant tissue, enabling representative analysis directly in a complete living plant individual. Nonetheless, it is essential to highlight and not confuse the main aim of this work, which was the classification of cannabis plant individuals into their corresponding chemotype through an easy and non-invasive methodology for quality control in production, with the quantification of cannabinoids present in the plants, for which the traditional methodology would be the drying of the corresponding plant material for the subsequent analysis by NIRS or other means such as chromatography.

Thus, as the main objective of this work was achieved, the proof of concept for comprehensive breeding control of crops, and more certainly of cannabis, through a non-invasive analytical technique was demonstrated. This fact could pave the way for non-invasive technology development in agricultural quality control, as the stated aim was achieved by avoiding usual analytical techniques such as chromatography or conventional SWIR spectroscopy. However, to make it possible, the classification model would need much more extensive training data. In order to enhance the generalizability of the results, a much larger sample number would be needed, with greater cultivar diversity coverage through different cultivation times. Therefore, an ideal proposal would be to extent the model calibration data with more cultivars from the three chemotypes, ideally balancing the statistical weight of the classes, and adding spectra from different plant tissues, such as leaves, and complementing it with spectra at different growing stages. This way, a regressive tendency would be deduced according to different times of growth, making the prediction of the corresponding chemotype in individuals at early growth stages possible.

Funding

This work was financially supported by the Education Department of the Basque Country as a consolidated group of the Basque Research System (IT1213-19).

CRediT authorship contribution statement

M. San Nicolas: Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft. **A. Villate:** Investigation. **I. Alvarez-Mora:** Methodology. **M. Olivares:** Conceptualization. **O. Aizpurua-Olaizola:** Conceptualization, Investigation, Project administration, Supervision, Writing – review & editing. **A. Usobiaga:** Conceptualization, Investigation, Project administration, Supervision, Writing – review & editing. **J.M. Amigo:** .

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

M. San Nicolas and A. Villate are grateful to the IBeA consolidated research group and to the UPV/EHU for their predoctoral fellowships.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2023.108551.

Bibliography

'State Medical Cannabis Laws'. Accessed: Jun. 02, 2023. [Online]. Available: https://www.ncsl.org/health/state-medical-cannabis-laws.

Abeysekera, S.K., Robinson, A., Ooi, M.P.L., Kuang, Y.C., Manley-Harris, M., Holmes, W., Hirst, E., Nowak, J., Caddie, M., Steinhorn, G., Demidenko, S., 2023. Sparse reproducible machine learning for near infrared hyperspectral imaging: Estimating the tetrahydrocannabinolic acid concentration in Cannabis sativa L. Industr. Crops Prod. 192, 116137 https://doi.org/10.1016/j.indcrop.2022.116137.

Kontrollierte Abgabe von Cannabis: Eckpunktepapier der Bundesregierung liegt vor. 2022. Accessed: Jun. 02, 2023. [Online]. Available: https://www.bundesgesundheitsministerium.de/ministerium/meldungen/kontrollierte-abgabe-von-cannabis-eckpunktepapier-der-bundesregierung-liegt-vor.html.

Federal Act on Narcotics and Psychotropic Substances (Narcotics Act, NarcA). 2022. Accessed: Jun. 02, 2023. [Online]. Available: https://www.fedlex.admin.ch/eli/cc/1952/241_241_245/en.

Aizpurua-Olaizola, O., et al., 2016. Evolution of the Cannabinoid and terpene content during the growth of Cannabis sativa plants from different chemotypes. J. Nat. Prod. 79 (2), 324–331. https://doi.org/10.1021/acs.jnatprod.5b00949.

Aizpurua-Olaizola, O., Omar, J., Navarro, P., Olivares, M., Etxebarria, N., Usobiaga, A., 2014. Identification and quantification of cannabinoids in Cannabis sativa L. plants by high performance liquid chromatography-mass spectrometry. Anal. Bioanal. Chem. 406 (29), 7549–7560. https://doi.org/10.1007/s00216-014-8177-x.

Aizpurua-Olaizola, O., Elezgarai, I., Rico-Barrio, I., Zarandona, I., Etxebarria, N., Usobiaga, A., 2017. Targeting the endocannabinoid system: future therapeutic strategies. Drug Discov. Today 22 (1), 105–110. https://doi.org/10.1016/j.drudis.2016.08.005.

Amigo, J.M., Babamoradi, H., Elcoroaristizabal, S., 2015. Hyperspectral image analysis. A tutorial. Anal. Chim. Acta 896, 34–51. https://doi.org/10.1016/j.aca.2015.09.030.

J. M. Amigo, 'Chapter 1.1 - Hyperspectral and multispectral imaging: setting the scene', in *Data Handling in Science and Technology*, vol. 32, J. M. Amigo, Ed., in Hyperspectral Imaging, vol. 32, Elsevier, 2019, pp. 3–16. doi: 10.1016/B978-0-444-63977-6.00001-8.

Berman, P., Futoran, K., Lewitus, G.M., Mukha, D., Benami, M., Shlomi, T., Meiri, D., 2018. A new ESI-LC/MS approach for comprehensive metabolic profiling of phytocannabinoids in Cannabis. Sci. Rep. 8 (1), 14280. https://doi.org/10.1038/s41598-018-32651-4.

Black, N., et al., 2019. Cannabinoids for the treatment of mental disorders and symptoms of mental disorders: a systematic review and meta-analysis. Lancet Psychiatry 6 (12), 995–1010. https://doi.org/10.1016/S2215-0366(19)30401-8.

Bonini, S.A., et al., 2018. Cannabis sativa: A comprehensive ethnopharmacological review of a medicinal plant with a long history. J. Ethnopharmacol. 227, 300–315. https://doi.org/10.1016/j.jep.2018.09.004.

Booth, J.K., Bohlmann, J., 2019. Terpenes in Cannabis sativa – From plant genome to humans. Plant Sci. 284, 67–72. https://doi.org/10.1016/j.plantsci.2019.03.022.

Chandra, S., Radwan, M.M., Majumdar, C.G., Church, J.C., Freeman, T.P., ElSohly, M.A., 2019. New trends in cannabis potency in USA and Europe during the last decade (2008–2017). Eur. Arch. Psychiatry Clin. Neurosci. 269 (1), 5–15. https://doi.org/10.1007/s00406-019-00983-5.

Citti, C., Pacchetti, B., Vandelli, M.A., Forni, F., Cannazza, G., 2018. Analysis of cannabinoids in commercial hemp seed oil and decarboxylation kinetics studies of cannabidiolic acid (CBDA). J. Pharm. Biomed. Anal. 149, 532–540. https://doi.org/10.1016/j.jpba.2017.11.044.

Citti, C., Braghiroli, D., Vandelli, M.A., Cannazza, G., 2018. Pharmaceutical and biomedical analysis of cannabinoids: A critical review. J. Pharm. Biomed. Anal. 147, 565–579. https://doi.org/10.1016/j.jpba.2017.06.003.

Cox-Georgian, D., Ramadoss, N., Dona, C., Basu, C., 2019. 'Therapeutic and medicinal uses of terpenes', in *Medicinal Plants*. From Farm to Pharmacy 333–359. https://doi.org/10.1007/978-3-030-31269-5_15.

Cristino, L., Bisogno, T., Di Marzo, V., 2020. Cannabinoids and the expanded endocannabinoid system in neurological disorders. Nat. Rev. Neurol. 16 (1), 9–29. https://doi.org/10.1038/s41582-019-0284-z.

Deidda, R., et al., 2021. New perspective for the in-field analysis of cannabis samples using handheld near-infrared spectroscopy: A case study focusing on the determination of Δ9-tetrahydrocannabinol. J. Pharm. Biomed. Anal. 202 https://doi.org/10.1016/j.jpba.2021.114150.

Di Marzo, V., 2018. New approaches and challenges to targeting the endocannabinoid system. Nat. Rev. Drug Discov. 17 (9), 623–639. https://doi.org/10.1038/nrd.2018.115.

Duchateau, C., Kauffmann, J.-M., Canfyn, M., Stévigny, C., De Braekeleer, K., Deconinck, E., 2020. Discrimination of legal and illegal Cannabis spp. according to European legislation using near infrared spectroscopy and chemometrics. Drug Test. Anal. 12 (9), 1309–1319. https://doi.org/10.1002/dta.2865.

Espel Grekopoulos, J., 2019. Construction and Validation of Quantification Methods for Determining the Cannabidiol Content in Liquid Pharma-Grade Formulations by Means of Near-Infrared Spectroscopy and Partial Least Squares Regression. Medical Cannabis and Cannabinoids 2 (1), 43–55. https://doi.org/10.1159/000500266.

Fraguas-Sánchez, A.I., Torres-Suárez, A.I., Nov. 2018. Medical Use of Cannabinoids. Drugs 78 (16), 1665–1703. https://doi.org/10.1007/s40265-018-0996-1.

Gülck, T., Møller, B.L., 2020. Phytocannabinoids: Origins and Biosynthesis. Trends Plant Sci. 25 (10), 985–1004. https://doi.org/10.1016/j.tplants.2020.05.005.

Holmes, W.S., et al., 2023. On machine learning methods to estimate cannabidiolic acid content of Cannabis sativa L. from near-infrared hyperspectral imaging. In: Presented at the Conference Record - IEEE Instrumentation and Measurement Technology Conference. https://doi.org/10.1109/I2MTC53148.2023.10175994.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24, 417–441. https://doi.org/10.1037/h0071325.

Jarén, C., Zambrana, P.C., Pérez-Roncal, C., López-Maestresalas, A., Ábrego, A., Arazuri, S., 2022. Potential of NIRS technology for the determination of cannabinoid content in industrial hemp (Cannabis sativa L.). Agronomy 12 (4), 938. https://doi.org/10.3390/agronomy12040938.

Lewis, M.A., Russo, E.B., Smith, K.M., 2018. Pharmacological Foundations of Cannabis Chemovars. Planta Med. 84 (4), 225–233. https://doi.org/10.1055/s-0043-122240.

Lopez, E., Etxebarria-Elezgarai, J., Amigo, J.M., Seifert, A., Sep. 2023. The importance of choosing a proper validation strategy in predictive models. A tutorial with real examples. Anal. Chim. Acta 1275, 341532. https://doi.org/10.1016/j.aca.2023.341532.

Lu, Y., Li, X., Young, S., Li, X., Linder, E., Suchoff, D., 2022. Hyperspectral imaging with chemometrics for non-destructive determination of cannabinoids in floral and leaf materials of industrial hemp (Cannabis sativa L.). Comput. Electron. Agric. 202 https://doi.org/10.1016/j.compag.2022.107387.

Maccarrone, M., et al., 2015. Endocannabinoid signaling at the periphery: 50 years after THC. Trends Pharmacol. Sci. 36 (5), 277–296. https://doi.org/10.1016/j.tips.2015.02.008.

Mobaraki, N., Amigo, J.M., 2018. HYPER-Tools. A graphical user-friendly interface for hyperspectral image analysis. Chemom. Intel. Lab. Syst. 172, 174–187. https://doi.org/10.1016/j.chemolab.2017.11.003.

Muller, C., Morales, P., Reggio, P.H., 2019. Cannabinoid ligands targeting TRP channels. Front. Mol. Neurosci. 11 https://doi.org/10.3389/fnmol.2018.00487.

Pearson, K., Nov. 1901. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2 (11), 559–572. https://doi.org/10.1080/14786440109462720.

Recommendation of removal for cannabis and cannabis resin from Schedule IV of 1961 Single Convention on Narcotic Drugs', presented at the United Nations Comission on Narcotic Drugs, Vienna, Dec. 2020. Accessed: Jun. 02, 2023. [Online]. Available: https://www.who.int/news/item/04-12-2020-un-commission-on-narcotic-drugs-reclassifies-cannabis-to-recognize-its-therapeutic-uses.

Commission Delegated Regulation (EU) 2017/1155 of 15 February 2017 amending Delegated Regulation (EU) No 639/2014 as regards the control measures relating to the cultivation of hemp, certain provisions on the greening payment, the payment for young farmers in control of a legal person, the calculation of the per unit amount in the framework of voluntary coupled support, the fractions of payment entitlements and certain notification requirements relating to the single area payment scheme and the voluntary coupled support, and amending Annex X to Regulation (EU) No 1307/2013 of the European Parliament and of the Council, vol. 167. 2017. Accessed: Nov. 08, 2023. [Online]. Available: http://data.europa.eu/eli/reg_del/2017/1155/oj/eng.

San Nicolas, M., et al., 2020. Analysis of cannabinoids in plants, marijuana products and biological tissues. Compr. Anal. Chem. 90, 65–102. https://doi.org/10.1016/bs.coac.2020.04.002.

San Nicolas, M., Villate, A., Olivares, M., Etxebarria, N., Zuloaga, O., Aizpurua-Olaizola, O., Usobiaga, A., 2023. Exploratory optimisation of a LC-HRMS based analytical method for untargeted metabolomic screening of Cannabis Sativa L. through Data Mining. Analytica Chimica Acta 1279, 341848. https://doi.org/10.1016/j.aca.2023.341848.

Sánchez-Carnerero Callado, C., Núñez-Sánchez, N., Casano, S., Ferreiro-Vera, C., 2018. The potential of near infrared spectroscopy to estimate the content of cannabinoids in Cannabis sativa L.: A comparative study. Talanta 190, 147–157. https://doi.org/10.1016/j.talanta.2018.07.085.

Small, E., Beckstead, H.D., 1973. Cannabinoid phenotypes in Cannabis sativa. Nature 245 (5421), 147–148. https://doi.org/10.1038/245147a0.

Ståhle, L., Wold, S., 1987. Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. J. Chemom. 1 (3), 185–196. https://doi.org/10.1002/cem.1180010306.

Student,, 1908. The probable error of a mean. Biometrika 6 (1), 1–25. https://doi.org/10.1093/biomet/6.1.1.

Su, K., et al., 2022. NIR spectroscopy for rapid measurement of moisture and cannabinoid contents of industrial hemp (Cannabis sativa). Ind. Crop. Prod. 184 https://doi.org/10.1016/j.indcrop.2022.115007.

Valinger, D., et al., 2021. Development of ANN models based on combined UV-vis-NIR spectra for rapid quantification of physical and chemical properties of industrial hemp extracts. Phytochem. Anal 32 (3), 326–338. https://doi.org/10.1002/pca.2979.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics 1 (6), 80–83.

Yao, S., Ball, C., Miyagusuku-Cruzado, G., Giusti, M.M., Aykas, D.P., Rodriguez-Saona, L. E., 2022. A novel handheld FT-NIR spectroscopic approach for real-time screening of major cannabinoids content in hemp. Talanta 247. https://doi.org/10.1016/j.talanta.2022.123559.

Zou, S., Kumar, U., 2018. Cannabinoid receptors and the endocannabinoid system: signaling and function in the central nervous system. Int. J. Mol. Sci. 19 (3), 833. https://doi.org/10.3390/ijms19030833.