

Article

Rapid In Situ Near-Infrared Assessment of Tetrahydrocannabinolic Acid in Cannabis Inflorescences before Harvest Using Machine Learning

Jonathan Tran ^{1,2,*}, Simone Vassiliadis ², Aaron C. Elkins ², Noel O. O. Cogan ^{1,2}  and Simone J. Rochfort ^{1,2} 

¹ School of Applied Systems Biology, La Trobe University, Bundoora, VIC 3083, Australia; noel.cogan@agriculture.vic.gov.au (N.O.O.C.); simone.rochfort@agriculture.vic.gov.au (S.J.R.)

² Agriculture Victoria Research, AgriBio Centre, AgriBio, Melbourne, VIC 3083, Australia; simone.vassiliadis@agriculture.vic.gov.au (S.V.); aaron.elkins@agriculture.vic.gov.au (A.C.E.)

* Correspondence: jonathan.tran@agriculture.vic.gov.au

Abstract: Cannabis is cultivated for therapeutic and recreational purposes where delta-9 tetrahydrocannabinol (THC) is a main target for its therapeutic effects. As the global cannabis industry and research into cannabinoids expands, more efficient and cost-effective analysis methods for determining cannabinoid concentrations will be beneficial to increase efficiencies and maximize productivity. The utilization of machine learning tools to develop near-infrared (NIR) spectroscopy-based prediction models, which have been validated from accurate and sensitive chemical analysis, such as gas chromatography (GC) or liquid chromatography mass spectroscopy (LCMS), is essential. Previous research on cannabinoid prediction models targeted decarboxylated cannabinoids, such as THC, rather than the naturally occurring precursor, tetrahydrocannabinolic acid (THCA), and utilize finely ground cannabis inflorescence. The current study focuses on building prediction models for THCA concentrations in whole cannabis inflorescences prior to harvest, by employing non-destructive screening techniques so cultivators may rapidly characterize high-performing cultivars for chemotype in real time, thus facilitating targeted optimization of crossbreeding efforts. Using NIR spectroscopy and LCMS to create prediction models we can differentiate between high-THCA and even ratio classes with 100% prediction accuracy. We have also developed prediction models for THCA concentration with a $R^2 = 0.78$ with a prediction error average of 13%. This study demonstrates the viability of a portable handheld NIR device to predict THCA concentrations on whole cannabis samples before harvest, allowing the evaluation of cannabinoid profiles to be made earlier, therefore increasing high-throughput and rapid capabilities.

Keywords: prediction models; cannabinoids; regression; classification; portability; THCA; PLS; SVM; XGB



Citation: Tran, J.; Vassiliadis, S.; Elkins, A.C.; Cogan, N.O.O.; Rochfort, S.J. Rapid In Situ Near-Infrared Assessment of Tetrahydrocannabinolic Acid in Cannabis Inflorescences before Harvest Using Machine Learning. *Sensors* **2024**, *24*, 5081. <https://doi.org/10.3390/s24165081>

Academic Editor: Ivan Andonovic

Received: 20 June 2024

Revised: 23 July 2024

Accepted: 1 August 2024

Published: 6 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Cannabis sativa L. is known for its medicinal and recreational uses, despite being used for thousands of years, its prohibition in many countries has hindered scientific research [1,2]. However, in recent years, the legalization of cannabis in some states and countries has led to a surge in studies exploring the potential benefits and risks of its use [3]. Advancements in cannabis research have prompted larger cultivation trials, aimed at exploring diverse phenotypic and chemovar variations, with the primary goal being to enhance production efficiencies and develop high-yielding cannabinoid cultivars [4]. Simultaneously, efforts are being made to simplify manual processing during cultivation and harvesting [4].

The two main cannabinoids of interest are cannabidiol (CBD) and delta-9-tetrahydrocannabinol (THC). CBD is used for the treatment of epilepsy, particularly in children [5], and as an analgesic and anti-anxiolytic [6]. THC has been investigated as a

treatment for multiple sclerosis, spasticity, chronic pain, and posttraumatic stress disorder (PTSD) bringing THC to the forefront of medicinal cannabis's focus [7–9]. THC and CBD are converted from tetrahydrocannabinolic acid (THCA) and cannabidiolic acid (CBDA), respectively, through a process called decarboxylation, which is typically achieved by heating cannabis inflorescences. These precursor compounds are naturally formed in the cannabis plants and found in abundance, where THC and CBD are present in relatively minor amounts, so it is important to target these THCA and CBDA compounds *in planta*, over their THC and CBD counterparts [10].

Traditional methods of cannabis analysis are performed using gas chromatography (GC) and liquid chromatography (LC), coupled with a mass spectroscopy (MS) as a detector [11], other methods use flame-ionization detection (FID) coupled with GC or an ultra-violet diode array detector coupled with LC [11,12]. Mass spectroscopy provides highly accurate and selective results; however, all forms and variations of analysis are expensive and slow compared to other techniques such as near-infrared (NIR) spectroscopy. NIR spectroscopy is a rapid and non-destructive technique capable of detecting molecular vibrations through organic bonds and discern them, however, it cannot fully elucidate chemical compounds and is non-selective for individual cannabinoids [13–17]. These spectra can be used to make cannabinoid content predictions or distinguish between cultivars once paired with high-quality quantitation data and multivariate statistical software [16,18].

Methods for using near-infrared spectroscopy for the analysis of cannabinoids in finely ground inflorescence material is well-researched and when combined with high-quality MS data can produce strong prediction models to predict cannabinoid content [13–17,19,20]. Sample preparation requirements can be reduced by leaving inflorescences whole or by coarsely grinding the tissues, as opposed to finely grinding the tissues which is necessary for LCMS and GCMS, thereby accelerating the sampling processes where high-throughput screening is needed. However, this is typically associated with loss in prediction model accuracy, as samples are less homogenous. This was illustrated by Deidda et al. (2021), where THC prediction models generated from finely ground inflorescences (R^2 0.93) outperformed coarsely ground material (R^2 0.74) [19]. Prior work compared coarsely ground inflorescences and finely ground inflorescences using a FT-NIR Bruker MPA II, where prediction models for 14 cannabinoids were produced. Partial least squares regression (PLS-R) models for CBDA, cannabinol (CBN), cannabinolic acid (CBNA) and cannabichromenic acid (CBCA) using finely ground inflorescences ($R^2 = 0.99, 0.71, 0.79$ and 0.60 , respectively) were comparable to those obtained using coarsely ground inflorescences ($R^2 = 0.98, 0.69, 0.79$ and 0.59 , respectively), while all other cannabinoid prediction models performed poorly (lower R^2 values) when coarsely ground inflorescences were used to make predictions [16].

Multiple machine learning modelling techniques should be employed when developing prediction methods, as certain techniques may produce a more robust outcome than others. PLS models are a set of versatile statistical methods which can be used for regression (PLS-R) and classification (partial least squares discriminant analysis—PLS-DA). PLS methods require a data training set and accurate phenotypes (in this case, NIR spectra and quantitative LCMS data). The methods prioritize data variance that is correlated with the trait of interest. PLS has been used in a wide variety of areas such as genomics, chemometrics, environment studies and neuroinformatics [21]. Support vector machine (SVM) is a supervised machine learning algorithm works by maximizing the difference between finding the optimal hyperplane that best separates the data using support vectors. SVM has been used in cheminformatics, drug discovery and many other disciplines [22,23]. Extreme gradient boosting (XGB) is an ensemble machine learning algorithm. Ensemble learning is where the model is built upon multiple 'weak' learning models to create a more robust and accurate model which is better suited to solve problems than single models. XGB has been used in diagnostic classification of cancers, optimization of manufacturing and credit risk assessment models for financial institutions [24–26].

This research aims to provide prediction models for THCA in intact inflorescences of *C. sativa* grown under glasshouse conditions. NIR data were acquired using a portable

handheld device (VIAVI MicroNIR) on whole, intact inflorescences to allow in situ, in planta scanning. Chemometrics included PLS-R, support vector machine regression (SVM-R) and extreme gradient boosting regression (XGB-R). This represents the first research to use a portable, handheld NIR device to distinguish between high THCA and even ratio chemovars in a glasshouse environment. High THCA was defined as classes that had a concentration ratio of CBDA to THCA lower than 1:1; and even ratio had a concentration ratio approximating 1:1 (CBDA:THCA ratio). This growing trial included 264 cannabis plants which is the largest sample size used for cannabinoid prediction models on whole cannabis inflorescences. This trial also has 88 unique genotypes to ensure a diverse and variable cannabinoid profile amongst the entire dataset to reduce population bias. The benefit of predicting cannabinoids using whole inflorescence is the elimination of needing to process the inflorescence into a finely ground powder, thus reducing cost and time, and increasing high-throughput capability.

2. Materials and Methods

2.1. Growth & Conditions

Cannabis plants were legally grown at a secured undisclosed location within a glasshouse by Cann Group (Cann Group Limited, Australia) under permit MC029-S01-C02 from the Office of Drug Control, Department of Health and Aged Care Australia. Plants grown during the vegetative state were exposed to a day-and-night cycle of 16:8 h. Plants during the flowering stages were exposed to a day-and-night cycle of 12:12 h. The temperature was maintained at 24–30 °C with 50–80% relative humidity. There were 88 unique genotypes cloned in triplicate to produce 264 cannabis plants. The chemovars present within population were high-producing THCA (n = 231) cultivars and even-ratio (n = 33) cultivars. Data collection took place over two days just prior to harvest. The data was collected at 10 am to 11 am each day at a temperature 25 °C with 60% relative humidity.

2.2. Instrumentation & Parameters

A VIAVI MicroNIR Onsite-W (VIAVI Solutions Inc., Scottsdale, AZ, USA) was utilized for scanning and data were collected within the 950–1650 nm (10,526–6060 cm⁻¹) range, with a resolution of 6.2 nm. The handheld portable MicroNIR uses two integrated vacuum tungsten lamps with a 128-pixel InGaAs photodiode array detector. The dispersing element is a VIAVI linear variable filter (LVF) (VIAVI Solutions Inc., Scottsdale, AZ, USA) and the windowed measurement collar uses an antireflective sapphire window. The signal-to-noise ratio is 25,000 and integration time of 10 ms. The portability and field deployability allows it to be used in the field or within growing facilities. The detector was thoroughly cleaned using 80% methanol (Thermo Fisher Scientific Inc., Waltham, MA, USA) (methanol:water) and a KimTech KimWipe tissue (Kimberly-Clark, Irving, TX, USA) between all scans until there was no residue left. All scans were performed on cannabis inflorescences before harvest, on the top apical inflorescence of each cannabis plant, in triplicate, scanning a different section of the top apical inflorescence each time and the resulting data averaged. The device was configured in diffuse reflectance mode with an integration time of 12 milliseconds and 100 scan counts. Data acquisition was performed using MicroNIR Pro v3.0 software (VIAVI Solutions Inc., Scottsdale, AZ, USA). In total, 264 cannabis samples were scanned (88 genotypes × 3 replicates per genotype).

2.3. Sample Preparation for LCMS Analysis

The top apical inflorescence of each cannabis plant was taken, dried for 72 h, and weighed to assess moisture loss. Each sample was then vacuum sealed and sent to Agriculture Victoria Research (Bundoora, Melbourne, Victoria, Australia) for LCMS analysis and stored at room temperature prior to sample preparation and analysis. All samples were freeze-dried for 48 h using a VirTis General Purpose Freeze Dryer (Scientific Products, Warminster, PA, USA), transferred into 15 mL grinding jars, placed in liquid nitrogen for 1 min and ground to fine powder using a SPEX SamplePrep 2010 Geno/Grinder (SPEX

SamplePrep, Metuchen, NJ, USA) for 1 min at 1500 rpm. 10 mg of the fine powder was then subsampled and quantitatively analyzed by LCMS for cannabinoid content, method described by Elkins et al. (2019) [12].

2.4. Statistical Analysis

VIAVI MicroNIR Onsite-W spectra data were exported from MicroNIR Pro v3.0 and imported into Excel (Microsoft Corporation, Redmond, WA, USA), where the data from the three scans were averaged. The averaged data was imported into MATLAB 2022a (Mathworks, Natick, MA, USA) with PLS-Toolbox 9.0 (Eigenvector Research Inc., Manson, WA, USA) for analysis. The full spectra range of the MicroNIR from 950–1650 nm ($10,526\text{--}6060\text{ cm}^{-1}$) was utilized (Figure 1). For statistical modelling, pre-processing used a combination of detrend, 2nd order derivate, standard normal variate (SNV) and mean centering to obtain the strongest model based on R^2 prediction value (Figure 2). Principal component analysis (PCA) was performed to check data quality and explore trends within the sample population. PLS-DA models were used to categorize chemovars as high THCA and even ratio, and PLS-R and SVM-R models were used to create predictive models for THCA. PLS, SVM and XGB models used LCMS quantitative data as dependent variables and NIR data from the VIAVI MicroNIR Onsite-W as independent variables. Venetian blinds cross-validation was used within 10 splits and blind thickness set to 1.

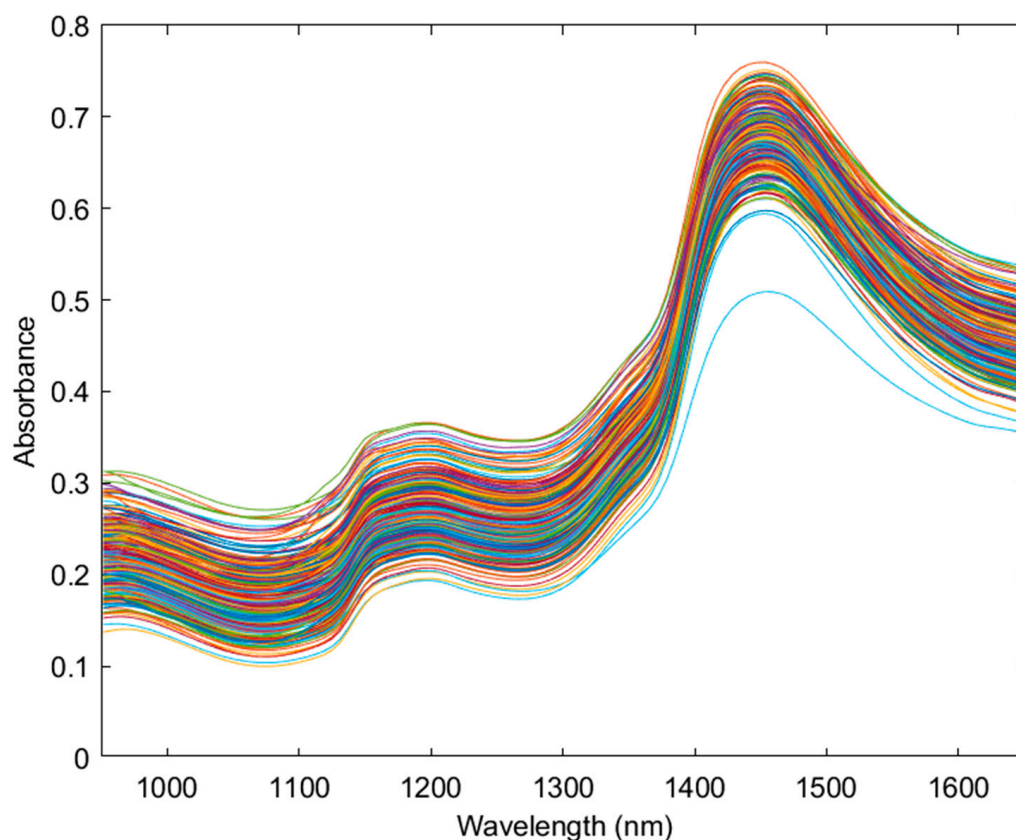


Figure 1. The raw near-infrared spectral data of the cannabis dataset ($n = 264$) from the 950 nm to 1650 nm range. Each color represents a unique scanned sample.

The dataset would be separated into calibration and validation sets where 75% ($n = 198$) of the data was used to train the model and 25% ($n = 66$) of the data was used to test the model. The Kennard-Stone algorithm was used to ensure the data selected for the calibration set was uniform and representative of the whole data set while the validation set contains samples that are interior and exterior to the calibration set. The algorithm achieves this by randomly seeking two sample with the data set with the largest distance measure

using either Euclidean or Mahalanobis distance [27]. This eliminates extrapolation of the calibration model when applied to the validation test. The same calibration and validation dataset split was used for all models. To test for statistical significance, permutation tests ($n = 200$) were performed for all models (Wilcoxon, Sign Test and Rand t -test).

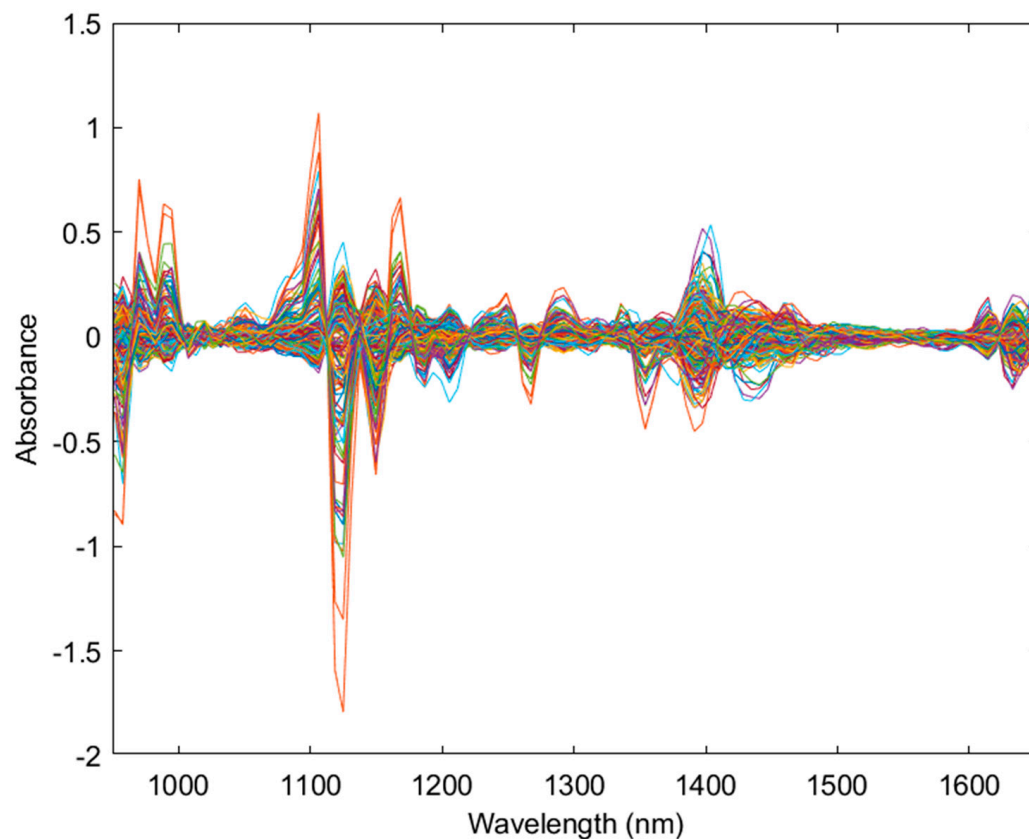


Figure 2. The pre-processed (detrend, standard normal variate, 2nd derivative and mean center) near-infrared spectral data of the cannabis dataset ($n = 264$) from the 950 nm to 1650 nm range. Each color represents a unique scanned sample.

3. Results

3.1. Principal Component Analysis of NIR Data

PCA (Figure 3) was performed on the dataset to identify trends and assess data quality. The data is spread across principal component (PC) 1 and 2, with PC1 accounting for 85.72% of the variation within the dataset, but no clusters separating into even ratio and high THCA chemovars were identified.

Previously, Tran et al. [16] performed a PCA on their cannabis datasets and reported major clusters in previous cannabis datasets which are categorized into high THCA, high CBDA and even ratio chemovars. Clustering indicates that discriminant analysis and regression models will likely be exceptional for the related chemovar. In the current paper, the data does not cluster into chemovars (high THCA and even ratio). An explanation for the absence of clustering may be due to the lack of homogeneity of whole cannabis inflorescence compared to finely ground cannabis inflorescence, where the latter would have more exposed surface area and uniform cannabinoid distribution when performing scans on the MicroNIR.

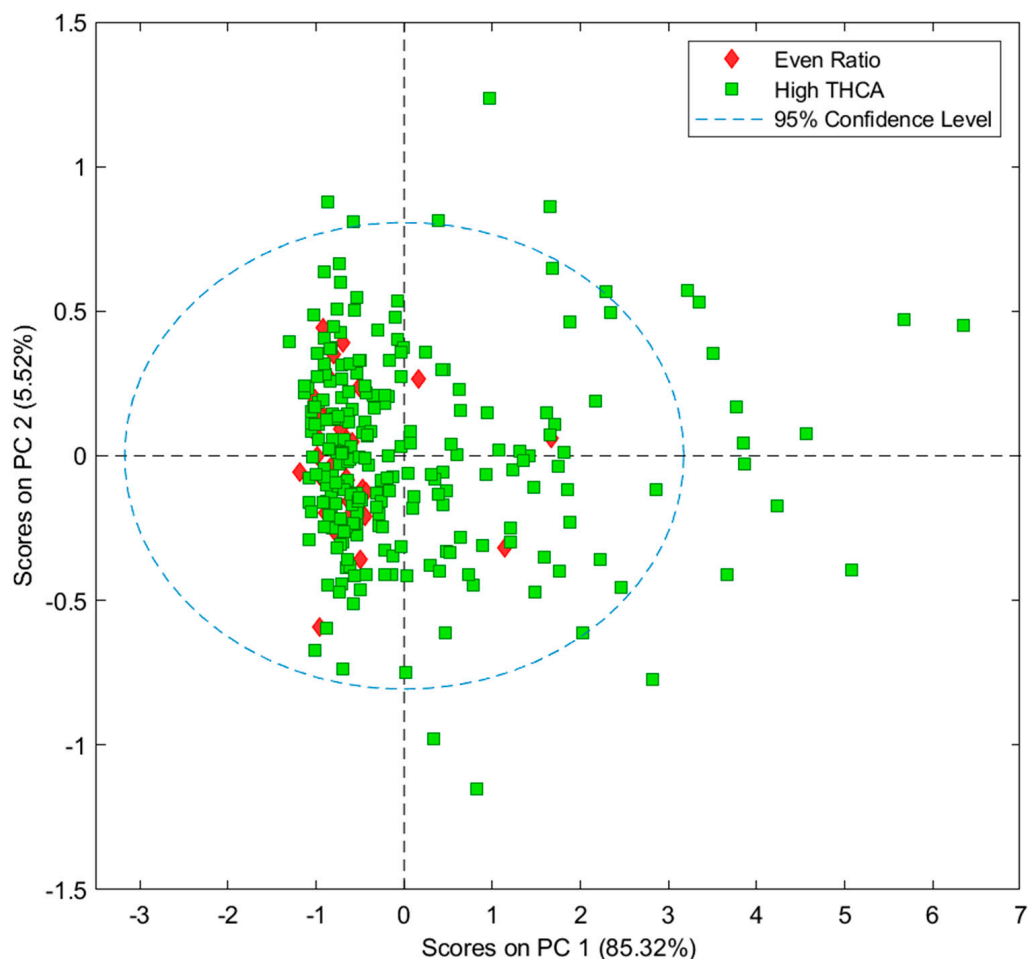


Figure 3. Scores of two principal components of the entire cannabis sample set (averaged data, $n = 264$). High THCA = green squares ($n = 231$), even ratio = red diamonds ($n = 33$).

3.2. Partial Least Squares Discriminant Analysis (PLS-DA) Modelling

PLS-DA models were able to predict cannabis samples as either high THCA or even-ratio classes with 100% accuracy (Class error Pred., 0%) (Table 1). High THCA was defined as a class that had a concentration ratio of CBDA to THCA lower than 1:1; and even ratio had a concentration ratio of more than 1:1 (CBDA:THCA ratio). The sample with highest CBDA:THCA ratio present was 1.723:1 (even ratio) and the lowest was 0.002:1 (high THCA), this data is available in Table S1. Sensitivity indicates how well the model can correctly classify samples as either high THCA or even ratio, whilst specificity measures how well the model can predict samples that are not high THCA or even ratio samples. The sensitivity and specificity values were equal to 1.00 for high-THCA and even ratio predictors, meaning a perfect discrimination between the two classes. The PLS-DA score plot of the validation dataset (Figure 4) showed clustering along the Latent Variable (LV) 2 with high THCA chemovars clustering together and the remaining even ratio chemovars distinctly separating towards the lower end of LV 2. This separation between the two chemovars is reflected in the classification predictions (Figure 5) where the validation results show that the PLS-DA model can discriminate between high THCA and even-ratio with 100% accuracy [28,29].

Table 1. Calibration (n = 198) and validation (n = 66) values of partial least squares discriminant analysis (PLS-DA) models on the ratio of cannabinoid compounds analyzed by the VIAVI MicroNIR Onsite-W instrument.

	High THCA ¹	Even Ratio ²
Sensitivity (Cal) ³	1.00	1.00
Specificity (Cal)	1.00	1.00
Sensitivity (CV) ³	1.00	1.00
Specificity (CV)	1.00	1.00
Sensitivity (Pred) ³	1.00	1.00
Specificity (Pred)	1.00	1.00
Class. Err ⁴ (Cal)	0.00	0.00
Class. Err (CV)	0.00	0.00
Class. Err (Pred)	0.00	0.00
RMSEC ⁵	0.15	0.15
RMSECV ⁶	0.15	0.15
RMSEP ⁷	0.12	0.12
Bias	0.00	0.00
CV Bias	0.00	0.00
Pred Bias	0.01	−0.01
R^2_{Cal} ⁸	0.83	0.83
R^2_{CV}	0.82	0.82
R^2_{Pred} ⁹	0.79	0.79

¹ THCA: tetrahydrocannabinolic acid; high THCA < 1:1. ² Even ratio > 1:1 (1:1 = CBDA:THCA ratio). ³ Cal: calibration; CV: cross validation; Pred: prediction. ⁴ Class. Err.: classification error. ⁵ RMSEC: root-mean square error of calibration. ⁶ RMSECV: root-mean-square error of cross-validation. ⁷ RMSEP: root-mean-square error of prediction. ⁸ R^2 : coefficient of determination. ⁹ R^2_{Pred} : coefficient of regression of measured data vs. predicted data; Permutation testing (n = 200) using Wilcoxon, Sign test and Rand *t*-test returned *p*-value < 0.05 for cross-validation (Figure S3a).

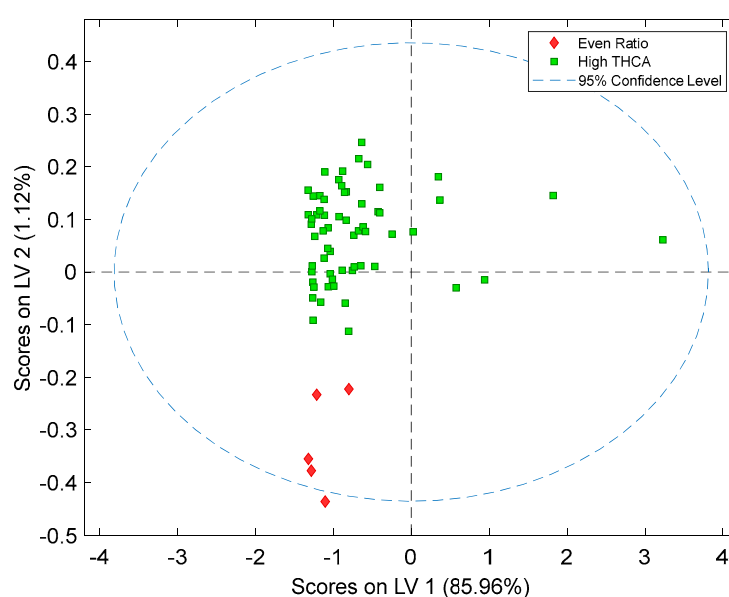


Figure 4. The PLS-DA score plot of two latent variables (LV 1 and LV 2) from the validation dataset (n = 66). Two clusters across LV2 were assigned as high-THCA = green squares (n = 61) and even ratio = red diamonds (n = 5).

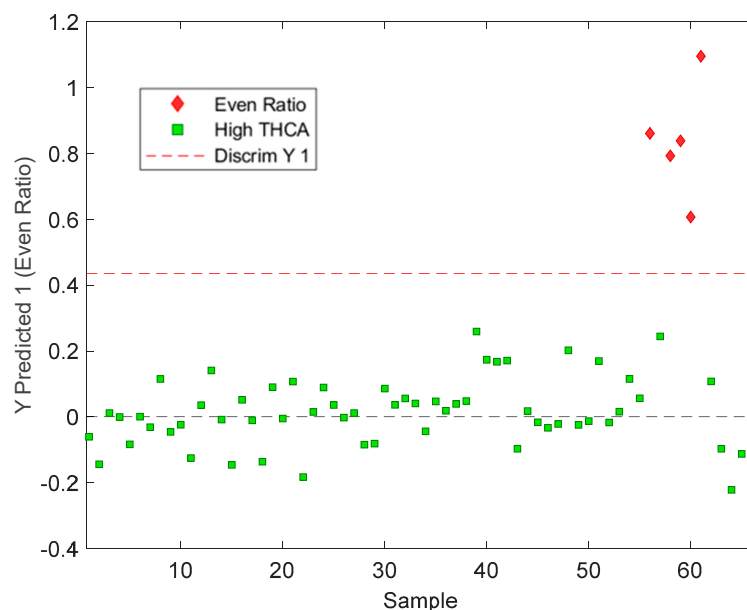


Figure 5. The PLS-DA classification of high-THCA = green squares ($n = 61$) and even ratio red diamonds ($n = 5$) classes of the cannabis validation dataset ($n = 66$).

Birenboim et al. [18] performed classification predictions with NIR but utilized homogenized cannabis powder instead of whole cannabis inflorescence. The ability to be able to use NIR to scan directly onto whole cannabis inflorescence before harvest whilst obtaining 100% accuracy in predictions is of great added value to the cannabis industry as there will be less focus on processing of every cannabis inflorescence to evaluate their chemovar profile using more expensive processes such as LCMS or GCMS.

The variable importance in projection (VIP) scores plot (Figure 6) displays the significance of each variable where scores of greater than one can be considered important. The VIP scores can be explained with respect to various organic functional groups; 950 nm (Ar-OH), 1150 nm (CH_2 & CH_3 groups), 1200 nm (CH groups), 1400 nm (CH and R-OH groups) and 1650 nm (Ar-CH bonds). The two compounds that factor heavily into these predictions are THCA and CBDA, which have Ar-OH, Ar-CH, CH and R-OH groups all which correspond to wavelengths on the NIR spectrum. THCA lacks one Ar-OH group whereas CBDA has two.

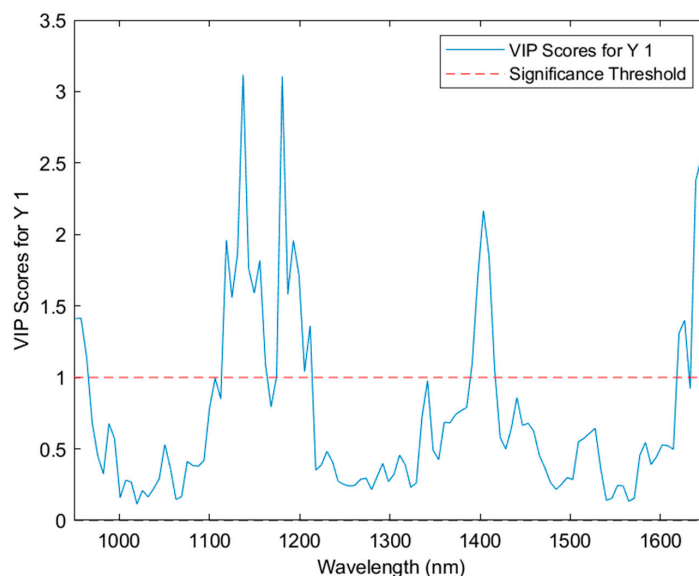


Figure 6. The VIP scores of the PLS-DA model across the near-infrared spectrum from 950 to 1650 nm.

3.3. Partial Least Squares Regression (PLS-R), Support Vector Machine Regression (SVM-R) and XGBoost Regression (XGB-R) Modelling

Different pre-processing parameters were applied to optimize the training models of each modelling algorithm for THCA concentrations (Table 2). XGB-R produced the highest R^2 calibration values (0.93, 1.00 and 1.00), followed by SVM-R (0.68, 0.78, 0.68) and PLS-R (0.62, 0.57, 0.62), however, when optimizing validation models, PLS-R (0.78, 0.75, 0.73) and SVM-R (0.77, 0.76, 0.74) models performed slightly better than the XGB-R (0.74, 0.59, 0.74) models.

Despite the XGB-R models exceptional performance with high R^2 calibration values, this was not reflected in validation tests, reducing R^2 prediction values to 0.74, 0.59 and 0.74. R^2 prediction values of 0.74 is suitable for screening, however, a R^2 prediction value of 0.59 is significant reduction from the R^2 calibration value of 1.00, suggesting an overfitting of the training model. The calibration step is very important in selecting the best performing modelling algorithm and preprocessing parameters, but it is vital that validation tests are performed against the training model to ensure the model is not overfitted to ensure suitability for real-world applications.

From this, a PLS-R modelling algorithm (Figure 6), with an R^2 prediction value of 0.78 and RPD (residual prediction deviation) of 2.08 was selected as the best prediction model and was suitable for screening THCA concentrations in whole cannabis inflorescence [30] whilst having a low prediction bias of -0.68 . It is to be noted that a positive value for prediction bias indicates an overfitting of a model, whilst a negative value for prediction bias indicates the underfitting of a model. Therefore, values nearing zero are ideal and indicate a lower bias. Ultimately, PLS-R produced the strongest model and the second lowest prediction bias.

THCA concentration across the cannabis population ranged from 37.29 to 229.87 mg/g as shown in Table S1. The residual prediction deviation (RPD) is typically measured as a goodness of fit of a prediction model, however as Minasny et al. (2013) [31] mentions, RPD is the same measure as R^2 , instead, a more suitable measure would be the ratio of standard error laboratory (SEL) versus the standard error prediction (SEP) [16,32]. The SEL and SEP ratio reported in supplementary files (Table S2) shows poor precision, but this is expected due to the heterogeneity of the cannabis samples analyzed via the NIR. The prediction error average was 13% for the validation set (Table S3) and defines the limits of this method, but it will still be a suitable method for general screening of whole cannabis inflorescence for THCA concentrations, especially for the cannabis industry where high THCA chemovars are the focus for both recreational and medicinal markets. Previous research which developed cannabinoid prediction models using dried finely ground cannabis inflorescence material have lower errors and stronger prediction models compared to whole inflorescence, likely due to a more homogenous sample with greater surface area compared to inflorescences in situ [13–17,19]. Another factor is the presence of water in the in-situ samples. As each scan was performed on a cannabis inflorescence before harvest, there was no drying step to remove moisture.

Water exhibits absorption bands in the NIR range at 1400–1450 nm. These bands correspond to hydroxyl groups (O-H) in water molecules which overlap with signals from hydrocarbon (C-H) and alcohol (R-OH) groups within the spectra and potentially cause inaccurate predictions. This provides challenges in creating robust prediction models, but the current study demonstrates that, in practice, this is a minor issue. Trimming the NIR spectra and removing the 1400–1450 nm range is alternative if eliminating the water interference is priority, however this resulted in significantly lower R^2 values (Table S4, Figure S1), likely due to the fact the omitted range still contain valuable peak data that corresponds to organic bonds found in cannabinoid compounds. The purpose of scanning inflorescences which are in situ, pre-harvest, intact and non-dried is intentional, as these conditions emulate the practices performed by industry cultivators.

Deidda et al. [19] (2021) used a MicroNIR (950–1650 nm) where three prediction models for THC concentration were developed for whole cannabis (0.93), coarsely ground (0.76)

and finely ground (0.77) inflorescences. In the same study a, NIR-S-G1 (900–1700 nm), also a handheld NIR device, was used but showed contradictory results where the R^2 prediction values were reversed: whole cannabis (0.73), coarsely ground inflorescences (0.74) and finely ground (0.93). Typically, the more homogenous a sample is, the more surface area is exposed for the NIR detector to scan organic compounds giving stronger R^2 prediction values. When a sample is heterogenous, a worse R^2 prediction value would be expected as there would be more sample variance due to the lack of uniformity in the cannabinoids as they are distributed across a whole cannabis inflorescence. This was shown in previous work by Tran et al. [16] (2023) when comparing coarsely ground and finely ground inflorescences, where finely ground produced stronger PLS-R models using a VIAVI MicroNIR Onsite-W and FT-NIR Bruker MPA II. In addition, Deidda et al. employed genetic algorithm (GA) for variable selection which can be prone to overfitting. This occurs when a dataset contains over 200 variables (each wavelength number), which may overestimate a model's performance and which may produce results not generalized enough to real world samples [33,34]. This current manuscript did not include GA results for reasons above, but it is worth noting that the R^2 performance was comparable with that of the PLS-R results in Table 2. The dataset was reduced to 36 samples to replicate the sample size of Deidda's study, but it resulted in poorer R^2 values (0.59) and significant prediction bias (8.03) (Table S5, Figure S2).

In previous literature, cannabinoid prediction models via NIR have shown limitations which calls for more robust models. This current manuscript improves on the model by containing 264 cannabis samples with 88 unique genotypes, making it the largest sample size used for predicting THCA concentrations from whole cannabis inflorescences. We have also conducted our analysis on in situ whole inflorescences before harvest, meaning the inflorescences are still attached to the plant with no sample handling processes. Finally, the current study performs no pre-treatment, where the cannabis inflorescences have not undergone any heating or decarboxylation step, leaving much of the THCA in its acid form rather than its neutral form, THC. This study furthers the cannabis industry by demonstrating that a handheld NIR system is suitable for the rapid determination of THCA concentration on in situ inflorescences, removing the need for extended drying periods and complex analyses. See Figures 7–9.

Table 2. The partial least squares regression (PLS-R), support vector machine regression (SVM-R) and extreme gradient boosting regression (XGB-R) models for the determination of THCA concentration.

Model	Figure Key	Region (nm) ¹	Scatter Correction ²	Derivative ^{2a}	N ³	RMSEC ⁴	R ² _{Cal} ⁵	RMSECV ⁶	R ² _{CV} ⁷	RMSEP ⁸	Pred Bias ⁹	R ² _{Pred} ¹⁰	RPD ¹¹
PLS-R	(a)	950–1650	DT, SNV and MC	2, 2, 5	264	26.34	0.62	28.87	0.54	21.49	−0.68	0.78	2.08
	(b)	950–1650	DT, SNV and MC	2, 2, 7	264	28.00	0.57	30.49	0.50	23.49	−2.61	0.75	1.91
	(c)	950–1650	DT, SNV and MC	2, 2, 3	264	26.41	0.62	30.93	0.48	23.34	0.77	0.73	1.93
SVM-R	(a)	950–1650	DT, SNV and MC	2, 2, 5	264	24.87	0.68	28.48	0.56	22.95	−1.47	0.77	1.96
	(b)	950–1650	DT, SNV and MC	2, 2, 7	264	23.87	0.70	29.40	0.53	22.49	−3.38	0.76	2.00
	(c)	950–1650	DT, SNV and MC	2, 2, 3	264	25.11	0.68	30.34	0.51	24.87	−1.80	0.74	1.81
XGB-R	(a)	950–1650	DT, SNV and MC	2, 2, 5	264	12.27	0.93	31.10	0.48	23.31	−3.46	0.74	1.88
	(b)	950–1650	DT, SNV and MC	2, 2, 3	264	0.02	1.00	34.61	0.37	28.77	0.44	0.59	1.56
	(c)	950–1650	DT and MC	2, 2, 5	264	0.25	1.00	34.09	0.38	23.02	−1.64	0.74	1.95

¹ nm: wavelength in nanometers. ² Preprocessing parameters. DT: detrend; SNV: standard normal variate; MC: mean centering. ^{2a} Derivative pre-treatment: the first digit is the polynomial order; the second digit is the derivative order, and the third digit is the data point gap which the derivative is calculated. ³ N: number of unique samples (Dataset was split using the Kennard-Stone algorithm into a calibration (n = 198) and validation set (n = 66) and used for all models. ⁴ RMSEC: root mean standard error of calibration. ⁵ R²_{Cal}: coefficient of determination of calibration. ⁶ RMSECV: root mean standard error of cross validation. ⁷ R²_{CV}: coefficient of determination of cross validation. ⁸ RMSEP: root mean standard error of prediction. ⁹ Pred Bias: calculated prediction bias. ¹⁰ R²_{Pred}: coefficient of regression of measured data vs. predicted data; Permutation testing (n = 200) using Wilcoxon, Sign test and Rand *t*-test returned *p*-value < 0.05 for cross-validation results (Figure S3). ¹¹ RPD: residual prediction deviation.

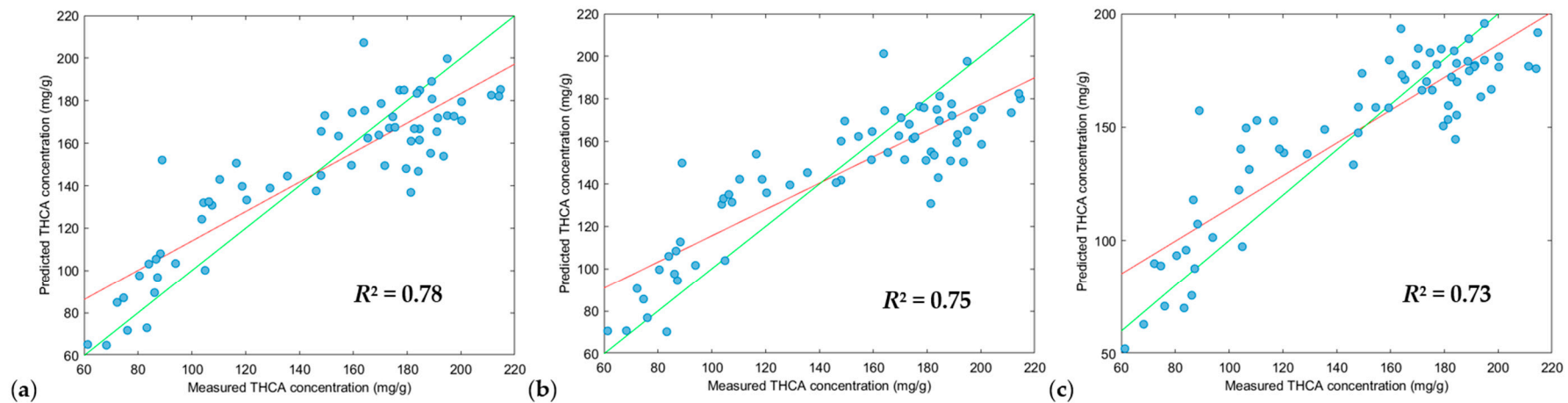


Figure 7. Plot of measured values versus the predicted values of THCA of the validation dataset ($n = 66$) using the PLS-R tool. Green line: line of best fit from reference data, red line: line of best fit from predicted data. Scatter correction and derivative settings: (a) DT, SNV and MC (2, 2, 5); (b) DT, SNV and MC (2, 2, 7); (c) DT, SNV and MC (2, 2, 3). Preprocessing parameters. DT: detrend; SNV: standard normal variate; MC: mean centering.

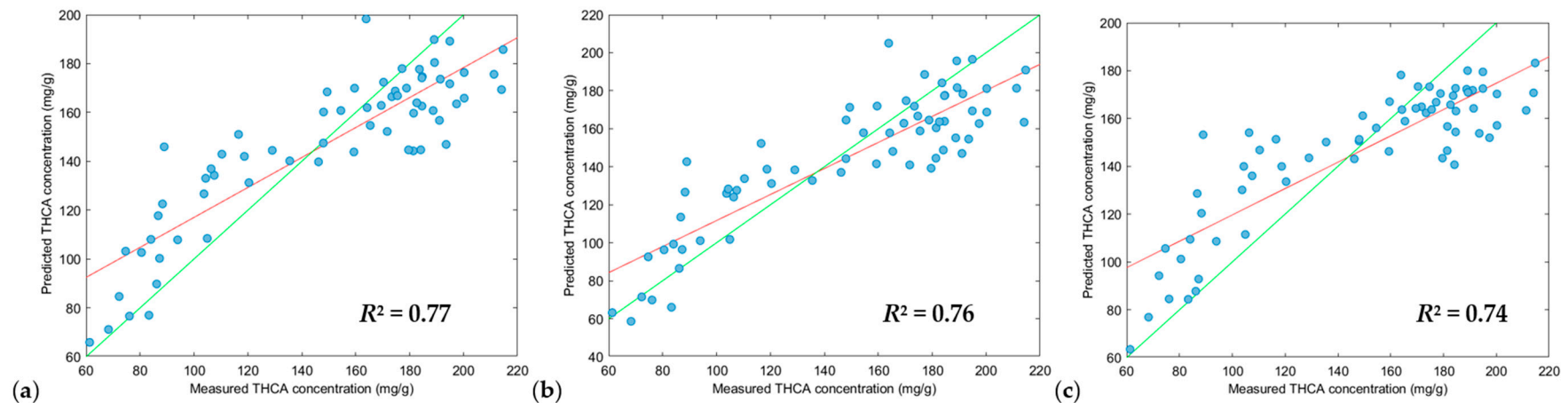


Figure 8. Plot of measured values versus the predicted values of THCA of the validation dataset ($n = 66$) using the SVM-R tool. Green line: line of best fit from reference data, red line: line of best fit from predicted data. Scatter correction and derivative settings: (a) DT, SNV and MC (2, 2, 5); (b) DT, SNV and MC (2, 2, 7); (c) DT, SNV and MC (2, 2, 3). Preprocessing parameters. DT: detrend; SNV: standard normal variate; MC: mean centering.

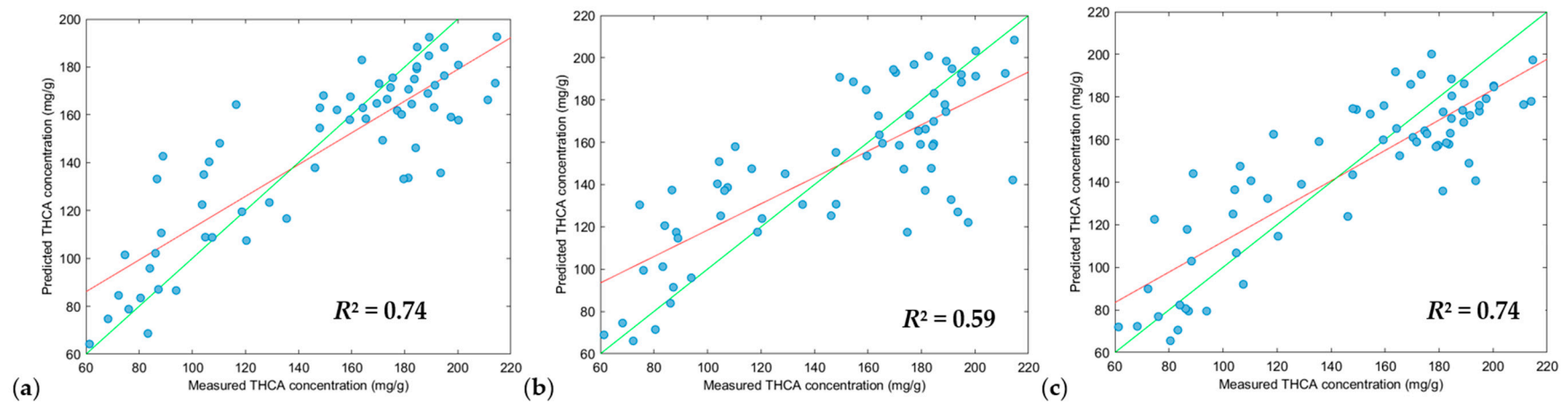


Figure 9. Plot of measured values versus the predicted values of THCA of the validation dataset ($n = 66$) using the XGB-R tool. Green line: line of best fit from reference data, red line: line of best fit from predicted data. Scatter correction and derivative settings: (a) DT, SNV and MC (2, 2, 5); (b) DT, SNV and MC (2, 2, 3); (c) DT and MC (2, 2, 5). Preprocessing parameters. DT: detrend; SNV: standard normal variate; MC: mean centering.

4. Conclusions

The focus of this study was to expand upon the predictive capabilities using NIR spectroscopy to produce prediction models for THCA concentrations and predict between high THCA and even-ratio chemovars on whole cannabis inflorescence. The VIAVI MicroNIR Onsite-W as selected for its portability, rapid acquisition, and low cost. PLS-DA models were 100% accurate for defining cannabis samples as either high THCA or even ratio chemovar. PLS-R, SVM-R and XGB-R models were used to predict THCA concentrations, where PLS-R produced the best R^2 prediction with low prediction bias. This method could be extended to predict other cannabinoids, provided there are sufficient sample numbers, and a range of cannabinoid concentrations present within the training set. The practicality of scanning whole cannabis inflorescences has improved the high-throughput capability by eliminating the need for grinding or processing the tissue down to fine powders resulting in reduced cost and time, remembering to consider that recalibrations should be performed based on temperature, humidity, and light changes. Scanning in real-time, allows for researchers, auditors, and cultivators to derive results onsite without needing to take samples to an analytical laboratory. This greatly benefits cultivators, as they can rapidly chemotype their plants and transfer their high-performing THCA plants into manufacturing glasshouses. In addition, real time, in situ screening may allow optimization of harvest times based on cannabinoid concentration, though this would require further research to demonstrate.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/s24165081/s1>, Table S1. The liquid chromatography mass spectra data of CBDA and THCA concentrations, ratios and chemovar classification within the cannabis dataset, Table S2. Standard error laboratory versus the standard error prediction for THCA, Table S3. Error of prediction values for the validation dataset used to test the PLS ($R^2 = 0.78$) prediction model. Table S4. The partial least squares regression (PLS-R) models for the determination of THCA concentration where the water peak region (1400–1450 nm) from the NIR spectra has been removed. Table S5. The partial least squares regression (PLS-R) models for the determination of THCA concentration where the sample size ($n = 36$) has been reduced. Figure S1. Plot of measured values versus the predicted values of THCA of the validation dataset using the PLS-R tool where the water peak region (1400–1450 nm) from the NIR spectra has been removed. Figure S2. Plot of measured values versus the predicted values of THCA of the validation dataset using the PLS-R tool where sample size ($n = 36$) is reduced. Figure S3. The permutation results of each discriminant and regression model.

Author Contributions: Conceptualization, J.T., S.V., A.C.E. and S.J.R.; methodology, J.T.; formal analysis, J.T.; investigation, J.T.; data curation, J.T. and A.C.E.; writing—original draft preparation, J.T.; writing—review and editing, J.T., S.V., A.C.E., S.J.R. and N.O.O.C.; visualization, J.T.; supervision, project administration, funding acquisition, S.V., A.C.E., S.J.R. and N.O.O.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Agriculture Victoria Research and Agriculture Victoria Services.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Raw data are available upon request.

Acknowledgments: Molecular Phenomics, Molecular Genetics and Cultivation teams at Agriculture Victoria Research. Fillipa Brugleria and Stefanie Kabitz at Cann Group Limited.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cooper, Z.D.; Abrams, D.I.; Gust, S.; Salicrup, A.; Throckmorton, D.C. Challenges for Clinical Cannabis and Cannabinoid Research in the United States. *J. Natl. Cancer Inst. Monogr.* **2021**, *2021*, 114–122. [[CrossRef](#)] [[PubMed](#)]
2. National Academies of Sciences, Engineering, and Medicine; Health and Medicine Division; Board on Population Health and Public Health Practice; Committee on the Health Effects of Marijuana: An Evidence Review and Research Agenda. Challenges and Barriers in Conducting Cannabis Research. In *The Current State of Evidence and Recommendations for Research*; National Academies Press: Washington, DC, USA, 2017.
3. Farrelly, K.N.; Wardell, J.D.; Marsden, E.; Scarfe, M.L.; Najdzionek, P.; Turna, J.; MacKillop, J. The Impact of Recreational Cannabis Legalization on Cannabis Use and Associated Outcomes: A Systematic Review. *Subst. Abuse.* **2023**, *17*, 11782218231172054. [[CrossRef](#)]
4. Ingvarsdén, C.R.; Brinch-Pedersen, H. Challenges and potentials of new breeding techniques in *Cannabis sativa*. *Front. Plant Sci.* **2023**, *14*, 1154332. [[CrossRef](#)]
5. Stockings, E.; Zagic, D.; Campbell, G.; Weier, M.; Hall, W.D.; Nielsen, S.; Herkes, G.K.; Farrell, M.; Degenhardt, L. Evidence for cannabis and cannabinoids for epilepsy: A systematic review of controlled and observational evidence. *J. Neurol. Neurosurg. Psychiatry.* **2018**, *89*, 741–753. [[CrossRef](#)]
6. Black, N.; Stockings, E.; Campbell, G.; Tran, L.T.; Zagic, D.; Hall, W.D.; Farrell, M.; Degenhardt, L. Cannabinoids for the treatment of mental disorders and symptoms of mental disorders: A systematic review and meta-analysis. *Lancet Psychiatry* **2019**, *6*, 995–1010. [[CrossRef](#)] [[PubMed](#)]
7. Therapeutic Goods Administration. *Guidance for the Use of Medicinal Cannabis in the Treatment of Multiple Sclerosis in Australia*; Department of Health, Australia Government: Canberra, Australia, 2017; pp. 9–14.
8. Pacitto, R.; Peters, C.; Iadipalo, A.; Rabinak, C.A. Cannabinoid modulation of brain activation during volitional regulation of negative affect in trauma-exposed adults. *Neuropharmacology* **2022**, *218*, 109222. [[CrossRef](#)]
9. Maayah, Z.H.; Takahara, S.; Ferdaoussi, M.; Dyck, J.R.B. The anti-inflammatory and analgesic effects of formulated full-spectrum cannabis extract in the treatment of neuropathic pain associated with multiple sclerosis. *Inflamm. Res.* **2020**, *69*, 549–558. [[CrossRef](#)]
10. Tahir, M.N.; Shahbazi, F.; Rondeau-Gagne, S.; Trant, J.F. The biosynthesis of the cannabinoids. *J. Cannabis Res.* **2021**, *3*, 7. [[CrossRef](#)]
11. Pourseyed Lazarjani, M.; Torres, S.; Hooker, T.; Fowlie, C.; Young, O.; Seyfoddin, A. Methods for quantification of cannabinoids: A narrative review. *J. Cannabis Res.* **2020**, *2*, 35. [[CrossRef](#)] [[PubMed](#)]
12. Elkins, A.C.; Deseo, M.A.; Rochfort, S.; Ezernieks, V.; Spangenberg, G. Development of a validated method for the qualitative and quantitative analysis of cannabinoids in plant biomass and medicinal cannabis resin extracts obtained by super-critical fluid extraction. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **2019**, *1109*, 76–83. [[CrossRef](#)]
13. Birenboim, M.; Kengisbuch, D.; Chalupowicz, D.; Maurer, D.; Barel, S.; Chen, Y.; Fallik, E.; Paz-Kagan, T.; Shimshoni, J.A. Use of near-infrared spectroscopy for the classification of medicinal cannabis cultivars and the prediction of their cannabinoid and terpene contents. *Phytochemistry* **2022**, *204*, 113445. [[CrossRef](#)] [[PubMed](#)]
14. Su, K.; Maghirang, E.; Tan, J.W.; Yoon, J.Y.; Armstrong, P.; Kachroo, P.; Hildebrand, D. NIR spectroscopy for rapid measurement of moisture and cannabinoid contents of industrial hemp (*Cannabis sativa*). *Ind. Crops Prod.* **2022**, *184*, 115007. [[CrossRef](#)]
15. Yao, S.; Ball, C.; Miyagusuku-Cruzado, G.; Giusti, M.M.; Aykas, D.P.; Rodriguez-Saona, L.E. A novel handheld FT-NIR spectroscopic approach for real-time screening of major cannabinoids content in hemp. *Talanta* **2022**, *247*, 123559. [[CrossRef](#)] [[PubMed](#)]
16. Tran, J.; Vassiliadis, S.; Elkins, A.C.; Cogan, N.O.I.; Rochfort, S.J. Developing Prediction Models Using Near-Infrared Spectroscopy to Quantify Cannabinoid Content in Cannabis Sativa. *Sensors* **2023**, *23*, 2607. [[CrossRef](#)] [[PubMed](#)]
17. Sanchez-Carnero Callado, C.; Nunez-Sanchez, N.; Casano, S.; Ferreiro-Vera, C. The potential of near infrared spectroscopy to estimate the content of cannabinoids in *Cannabis sativa* L.: A comparative study. *Talanta* **2018**, *190*, 147–157. [[CrossRef](#)]
18. Birenboim, M.; Chalupowicz, D.; Maurer, D.; Barel, S.; Chen, Y.; Fallik, E.; Paz-Kagan, T.; Rapaport, T.; Sadeh, A.; Kengisbuch, D.; et al. Multivariate classification of cannabis chemovars based on their terpene and cannabinoid profiles. *Phytochemistry* **2022**, *200*, 113215. [[CrossRef](#)] [[PubMed](#)]
19. Deidda, R.; Coppey, F.; Damergi, D.; Schelling, C.; Coic, L.; Veuthey, J.L.; Sacre, P.Y.; De Bleye, C.; Hubert, P.; Esseiva, P.; et al. New perspective for the in-field analysis of cannabis samples using handheld near-infrared spectroscopy: A case study focusing on the determination of Delta(9)-tetrahydrocannabinol. *J. Pharm. Biomed. Anal.* **2021**, *202*, 114150. [[CrossRef](#)] [[PubMed](#)]
20. Jarén, C.; Zambrana, P.C.; Pérez-Roncal, C.; López-Maestresalas, A.; Ábrego, A.; Arazuri, S. Potential of NIRS Technology for the Determination of Cannabinoid Content in Industrial Hemp (*Cannabis sativa* L.). *Agronomy* **2022**, *12*, 938. [[CrossRef](#)]
21. Mehmood, T.; Ahmed, B. The diversity in the applications of partial least squares: An overview. *J. Chemom.* **2015**, *30*, 4–17. [[CrossRef](#)]
22. Rodríguez-Perez, R.; Bajorath, J. Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery. *J. Comput. Aided. Mol. Des.* **2022**, *36*, 355–362. [[CrossRef](#)]
23. Cervantes, J.; Garcia-Lamont, F.; Rodríguez-Mazahua, L.; López, A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing* **2020**, *408*, 189–215. [[CrossRef](#)]
24. Lee, S.; Park, J.; Kim, N.; Lee, T.; Quagliato, L. Extreme gradient boosting-inspired process optimization algorithm for manufacturing engineering applications. *Mater. Des.* **2023**, *226*, 111625. [[CrossRef](#)]

25. Ma, B.; Meng, F.; Yan, G.; Yan, H.; Chai, B.; Song, F. Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. *Comput. Biol. Med.* **2020**, *121*, 103761. [[CrossRef](#)] [[PubMed](#)]
26. Chang, Y.-C.; Chang, K.-H.; Wu, G.-J. Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Appl. Soft Comput.* **2018**, *73*, 914–920. [[CrossRef](#)]
27. Gallagher, N.B.; O’Sullivan, D. Selection of Representative Learning and Test Sets Using the Onion Method. Available online: https://eigenvector.com/wp-content/uploads/2020/01/Onion_SampleSelection.pdf (accessed on 9 September 2022).
28. Szymanska, E.; Saccenti, E.; Smilde, A.K.; Westerhuis, J.A. Double-check: Validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics* **2012**, *8*, 3–16. [[CrossRef](#)]
29. Grasel, F.S.; Ferrão, M.F. A rapid and non-invasive method for the classification of natural tannin extracts by near-infrared spectroscopy and PLS-DA. *Anal. Methods* **2016**, *8*, 644–649. [[CrossRef](#)]
30. Williams, P.; Manley, M.; Antoniszyn, J. *Near-Infrared Technology: Getting the Best out of Light*, 1st ed.; African Sun Media: Stellenbosch, South Africa, 2019; pp. 69–102.
31. Minansy, B.; McBratney, A.B. Why you don’t need to use RPD. *Pedometron* **2013**, *33*, 14–15.
32. Garcia Martin, J.F. Potential of Near-Infrared Spectroscopy for the Determination of Olive Oil Quality. *Sensors* **2022**, *22*, 2831. [[CrossRef](#)] [[PubMed](#)]
33. Leardi, R.; Nørgaard, L. Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions. *J. Chemom.* **2004**, *18*, 486–497. [[CrossRef](#)]
34. Kawamura, K.; Tsujimoto, Y.; Nishigaki, T.; Andriamananjara, A.; Rabenarivo, M.; Asai, H.; Rakotoson, T.; Razafimbelo, T. Laboratory Visible and Near-Infrared Spectroscopy with Genetic Algorithm-Based Partial Least Squares Regression for Assessing the Soil Phosphorus Content of Upland and Lowland Rice Fields in Madagascar. *Remote Sens.* **2019**, *11*, 506. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.