Document downloaded from:

http://hdl.handle.net/10251/190010

This paper must be cited as:

Dos Santos, I.; Bosman, G.; Aleixandre Tudo, J.; Du Toit, W. (2022). Direct quantification of red wine phenolics using fluorescence spectrosopy with chemometrics. Talanta. 236:1-13. https://doi.org/10.1016/j.talanta.2021.122857

The final publication is available at

https://doi.org/10.1016/j.talanta.2021.122857

Copyright Elsevier

Additional Information

Direct quantification of red wine phenolics using fluorescence spectroscopy with chemometrics

Isabel dos Santos[1], Gurthwin Bosman[2], Jose Luis Aleixandre-Tudo[1,3,*] and Wessel du Toit[1]

[1]South African Grape and Wine Research Institute (SAGWRI), Department of Viticulture and Oenology, Stellenbosch University, South Africa

[2]Department of Physics, Stellenbosch University, South Africa

[3]Instituto de Ingeniería de Alimentos para el Desarrollo (IIAD), Departamento de Tecnologia de Alimentos, Universidad Politécnica de Valencia, España.

*Corresponding author

joaltu@sun.ac.za

ABSTRACT

Phenolic compounds are secondary metabolites known to play crucial roles in important chemical reactions impacting the mouthfeel, colour and ageing potential of red wine. Their complexity has resulted in a number of advanced analytical methods, which often prevent routine phenolic analysis in winemaking. Fluorescence spectroscopy could be an alternative to current spectrophotometric techniques and its combination with chemometrics was investigated for its suitability in directly quantifying phenolic content of unaltered red wine and fermenting samples. Front-face fluorescence was optimised and used to build predictive models for total phenols, total condensed tannins, total anthocyanins, colour density and polymeric pigments. Machine learning algorithms were used for model development. The most successful models were built for total phenols, total condensed tannins and total anthocyanins with coefficient of determination (R^2cal) and RMSECV of 0.81, 0.89, 0.80 and 5.71, 104.03 mg/L, 60.67 mg/L, respectively. The validation results showed R^2val values of 0.77, 0.8 and 0.77, and RMSEP values of 7.6, 172.37 mg/L and 76.57 mg/L, respectively. A novel approach for the classification of South African red wine cultivars based on unique fluorescent fingerprints was also successful with an overall cross validation score of 0.8. The best classification ability (validation score = 0.93) was shown for the data set containing only fermenting wines for the most widely represented cultivars (>20 samples). This approach may provide a useful tool for authentication and quality control by regulatory bodies.

KEYWORDS

Fluorescence spectra, direct measurements, unaltered samples, phenolic compounds, chemometrics, machine learning

1. INTRODUCTION

Phenolic compounds are a diverse group of secondary metabolites found in grapes and wine that can be classified into two families; flavonoids (flavonols, flavan-3-ols and anthocyanins) and non-flavonoids (phenolic acids and stilbenes) [1,2]. The final phenolic composition of a wine is dependent on numerous factors including viticultural aspects influencing grape berry development and ripening, the grape cultivar and chemical composition at harvest, as well as the winemaking practices implemented throughout fermentation and ageing [1]. Phenolic compounds have been widely studied for their crucial roles in various chemical reactions that greatly impact important wine attributes, such as mouthfeel, colour and ageing potential [2,3,4].

The complexity and diversity of red wine phenolic compounds has resulted in numerous analysis methods being developed in order to simplify complex phenolic chemistry into the most relevant phenolic information. The basic spectrophotometric methods most often used are UV-Vis based and rely on the spectral properties of the aromatic ring present in all phenolic compounds, allowing for differentiation between phenolic groups according to characteristic wavelength peaks [5,6]. Alternatives such as high-performance liquid chromatography (HPLC) are highly sensitive but rarely used outside of research applications while infrared spectroscopies, specifically Fourier transform, have been reported as suitable in phenolic analysis [2,7,8,9]. Several of these existing methods may require expensive equipment and reagents as well as the need for trained personnel, preventing the routine analysis of important phenolic parameters during winemaking outside of phenolic research. Spectroscopy combined with chemometrics is becoming increasingly investigated in both academic and industry domains to meet growing demands for rapid, accurate, cost-effective and user-friendly analysis techniques that may be applied on site as well as developed into process monitoring, optimisation and control systems.

Fluorescence spectroscopy has been widely used in chemistry and biochemistry disciplines due to its success in analysing the structures, functions and reactivities of numerous compounds, thereby allowing it to become an important tool in the authentication and quality control of many food science disciplines [10]. The advantages of fluorescence spectroscopy include being non-destructive, user-friendly, cost effective and highly sensitive when compared to other spectrophotometric methods [10,11,12]. The fluorescent capabilities of the complex wine matrix have been investigated with polyphenols being identified as the largest concentration of naturally occurring fluorophores [11]. Previous research has been conducted to analyse these fluorescent compounds both qualitatively and quantitatively, with Cabrera-Bañegil *et al*. [13,14] able to quantify pure compounds including catechin, epicatechin, vanillic acid, caffeic acid and resveratrol. Classification tasks have, however, been the focus in wine fluorescence research, with wine authentication according to cultivar, appellation and vintage having been successful [11,15]. Understanding the limitations and principles of fluorescence instrumentation is important when conducting analysis, with sample geometry being a major consideration. The conventional right-angled technique traditionally used in fluorescence spectroscopy is used in the analysis of clear or diluted samples. Owing to the complexity of the wine matrix and the chemical interactions taking place within it, as well as the sensitivity of fluorophores to their surrounding environment, a front-face technique developed by Parker [16] overcomes the need for dilution and allows the analysis of unaltered samples while minimising sample absorbance and spectral distortions [11,12,17]. Front-face fluorescence therefore presents itself as a potential alternative for the direct and non-invasive analysis of samples during the winemaking process, directly from the fermentation vessel.

Combining spectroscopy with chemometrics (multivariate statistical analysis) holds several advantages including the decomposition and interpretation of complex data sets in a considerably reduced analysis time, its non-destructive nature, and the simultaneous quantification of several analytes from a single spectral measurement [2,18]. The most commonly used multi-way techniques in fluorescence analysis have included parallel factor analysis (PARAFAC) as well as unfolded and N-way partial least squares (U-PLS and N-PLS) [13,19]. Modern machine learning techniques have previously not been investigated in this research area despite their success in complex data handling and ubiquitous use in current technologies.

The need for real-time, rapid, cost-effective and accurate phenolic analysis methods is steadily increasing and routine implementation may aid in the decision-making of winemakers and producers during red wine production. The potential for automation and on-line systems as well as optical portable devices is possible due to the beneficial combination of spectroscopy and chemometrics [20]. The aim of this study was therefore to investigate the suitability of front-face fluorescence spectroscopy to quantify phenolic content of undiluted red wine samples. The five parameters of interest included total phenols, total condensed tannins, total anthocyanins, colour density and polymeric pigments. Previous wine fluorescence research has, to the best of our knowledge, not investigated the potential of fluorescence spectroscopy to quantify such broad phenolic parameters with a focus on the implications for real-time analysis during the winemaking process. Classification of South African red wine cultivars using fluorescent excitation-emission matrices was also explored for its potential in authentication and quality control.

2. MATERIALS AND METHODS

2.1. REAGENTS

Ammonium sulphate, hydrochloric acid (HCl 1 M), methyl cellulose, sulphur dioxide (SO_2), ethanol (96%) and sodium metabisulfite (2.5 %) were purchased from Sigma-Aldrich Chemie (Steinheim, Germany). (-)-Epicatechin and malvidin-3-glucoside were purchased from Extrasynthese (Genay, France).

2.2. SAMPLES

The collection of 200 fermenting red wine samples took place over the 2019 vintage, following a diverse range of cultivars, vinification practices and terroirs. Both commercial and experimental scale conditions were included, with 91 samples collected from commercial cellars (Stellenbosch University Welgevallen Wine Cellar, Thelema Mountain Vineyards and Kanonkop Wine Estate) and 109 samples collected from the JHN Neethling experimental cellar at the Department of Viticulture and Oenology (Stellenbosch University). Samples were immediately frozen upon collection. During analysis, samples were thawed and immediately centrifuged at 5000 rpm for 2 min in an Eppendorf 5415D centrifuge (Hamburg, Germany). Additionally, 100 red wine samples from the Agricultural Research Council (ARC Infruitec-Nietvoorbij, Stellenbosch) spanning several vintages (2007-2018) and cultivars were collected, stored at room temperature and centrifuged at 5000 rpm for 2 min on the day of analysis. The cultivars represented in the study, each with varying numbers of samples, included Shiraz (90), Pinotage (49), Cabernet Sauvignon (47), Merlot (36), Malbec (19), Petit Verdot (14), Grenache (9), Pinot noir (9), Mourvedre (6), Tempranillo (5), Cinsaut (4), Arinarnoa (4), a blend (Pinotage, Shiraz and Malbec) (4), Marselan (2), Cabernet Franc (1) and Sangiovese (1).

2.3. SPECTROPHOTOMETRIC ANALYSIS

All analyses were conducted with UV-Vis spectroscopy using a Multiskan GO Microplate Spectrophotometer (Thermo Fisher Scientific, Inc., Waltham, MA, USA). The total phenolics index and total anthocyanin contents were quantified using the methodology reported by Iland *et al.* [21]. One hundred µl of sample supernatant was diluted 50 times with 1 M HCl, vortexed and stored for 1 hour in a dark cupboard before the absorbances between 200-700 nm at 2 nm intervals were recorded. The total phenolics index was calculated as the absorbance at 280 nm multiplied with the dilution factor while total anthocyanin content was calculated in mg/L malvidin-3-glucoside using the absorbance at 520 nm. Total condensed tannin concentration was determined using the methyl cellulose precipitable (MCP) tannin assay protocol developed by Sarneckis [22] and later modified by Mercurio *et al.* [23]. In 2 ml microfuge tubes, the treatment involved 50 µl of wine diluted with 600 µl of MCP solution (0.04% w/v), vortexed and left for 2-3 min before 400 µl of ammonium sulphate and 950 µl of distilled water was added. The control tubes contained no MCP solution but rather a total of 1.55 ml distilled water. Both control and treatment stood for 10 min before being centrifuged in an Eppendorf 5415D centrifuge (Hamburg, Germany) at 10 000 rpm for 5 min. The tannin content was then calculated using the difference between control and treatment samples at 280 nm and converted to mg/L using a calibration curve in epicatechin equivalents and a dilution factor of 40. Colour density was determined using the method reported by Glories [24] whereby 50 µl of wine was analysed against a blank of deionised water and the absorbance recorded at 420 nm, 520 nm and 620 nm. The sum of the three wavelengths was used to determine the colour density of the sample. Polymeric pigments were calculated using the modified Somers assay [23]. In 2 ml microfuge tubes, 200 µl of sample supernatant was diluted with 1.8 ml buffer solution (12% v/v ethanol, 0.5 g/L w/v tartaric acid at pH 3.4) containing 2.5 % sodium metabisulfite, and vortexed. The samples were stored for 1 hour and then analysed at 520 nm. The polymeric or SO_2 resistant pigments were then calculated in absorption units (AU) using a dilution factor of 10.

2.4. FLUORESCENCE INSTRUMENTATION

Parameters of a Perkin Elmer LS50B Spectrophotometer were investigated with regards to the intensity, excitation and emission ranges appropriate for wine analysis using diluted samples and conventional fluorescence analysis. A front-face accessory was thereafter investigated to ensure similarly appropriate parameters were obtained, and the optimal angle of incidence identified as that between the excitation beam and the sample perpendicular, was determined as 30 degrees. Inner filter effects were explored and deemed minor within the scope of the study. This calibration from conventional to front-face fluorescence was conducted using a Cabernet Sauvignon wine sample (2018) and validated with a Merlot wine sample (2018) (data not shown).

2.5. FLUORESCENCE SPECTROSCOPY

Front-face fluorescence analysis was conducted on all undiluted samples at room temperature within an air-conditioned area to minimise the effects of instrumental fluctuations. A 700 µl quartz cuvette (2 mm width) (Hellma Analytics, Germany) was used together with a 2 cm in diameter aperture fitted in the emission path in order to provide additional filtering of Rayleigh scattering. The excitation-emission matrix (EEM) per sample was recorded as emission spectra between 245 nm and 500 nm at 0.5 nm intervals for excitation wavelengths between 245 nm and 400 nm at 5 nm intervals. Scanning speed was set at 500 nm/min and the excitation and emission slit widths were set at 3 nm and 5 nm, respectively. The UV Winlab instrument software was used for data acquisition.

2.6. DATA PRE-PROCESSING

A single, complete dataset containing the combined 289 EEMs was created (11 samples were excluded due to unexplained oversaturation during fluorescence analysis). Once combined, spectral interferences were removed from the EEMs as described by Airado-Rodríguez *et al*. [11]. First and second order Rayleigh scatter were removed by excluding the excitation peaks on the identity line ($\lambda_{ex} = \lambda_{em}$) and at ($2\lambda_{ex} = \lambda_{em}$), respectively. The triangular non-chemical region below the identity line ($\lambda_{ex} > \lambda_{em}$) was set to zero. The software used for data and image processing throughout the study include the open-source web-based user interface JupyterLab (Project Jupyter, USA) using the Python 3 language library scikit-learn [25] and Matlab version 9.5 (The Mathworks Inc., MA, USA).

2.7. CHEMOMETRICS

2.7.1. PARALLEL FACTOR ANALYSIS (PARAFAC)

PARAFAC was performed in Matlab using the PLS_Toolbox (The Mathworks Inc., MA, USA) as described in literature [11,14,26]. The pre-treated EEMs of the 289 samples were stacked in a trilinear arrangement of I x J x K vectors (*samples* x *excitation wavelengths* x *emission wavelengths*) resulting in an initial 289 x 32 x 480 three-dimensional array. Spectral artifacts led to a reduction in EEM size from excitation and emission wavelengths between 245-400 nm and 260-500 nm, to 245-340 nm and 265-500 nm, respectively. The final three-way array of 289 x 20 x 470 was obtained. The appropriate number of components was chosen based on the core consistency diagnostic (CORCONDIA) and explained variance for non-negativity constrained models. Split-half analysis was conducted for model validation. Linear regression was then performed in JupyterLab on the resulting score values to determine univariate calibration models.

2.7.2. MACHINE LEARNING

Conventional linear regression in the form of principal component regression and partial least squares regression (PCR and PLSR) were investigated in JupyterLab. The exploration of linear regression included specific region selection based on phenolic fluorescence as found in literature [11], data scaling and outlier removal. Machine learning was investigated as a data modelling alternative and an exploration of the optimal pre-processing parameters focused on variable selection, data scaling, spectral region selection and choice of modelling technique. A machine learning pipeline was built in Python consisting of five consecutive steps namely, a column selector used to select for specific columns within the data and allow for spectral region selection between excitation 245-400 nm and emission 245-500 nm, a savgol transform used to apply a Savitzky-Golay filter for data smoothing [27], a pre-processing selector used to find the optimal scaling technique, principal component analysis (PCA) for data decomposition, and the XGBoost regressor to build a tree-based gradient boosted model [28]. Bayesian optimisation was used as the framework for automatically tuning the hyper-parameters of the pipeline [29,30] and explored over 2 000 iterations and over 160 model configurations per model.

Figure 1 is a graphical representation of the machine learning pipeline procedure. Briefly, the data was automatically and randomly split using the Kennard-Stone algorithm into train and test sub datasets, of which 20 samples were retained for model validation. Following this train and test split, a (Synthetic Minority Over-Sampling Technique for Regression) SMOTER algorithm was applied to the training set data. SMOTER makes use of interpolation of target samples identified as extreme cases or within the minority in order to create synthetic samples that improve upon model training [31]. A 99% threshold was used, identifying cases within the rare extreme and a k=3 value for k- Nearest Neighbours (KNN) was defined as the interpolation parameter to create the synthetic samples. The training data was thereafter passed through each consecutive step of the pipeline per phenolic parameter, with Bayesian optimisation automatically identifying the best hyper-parameters required for optimal prediction accuracy. Evaluation metrics including coefficient of determination (R^2cal and R^2val), root mean square error (RMSE) and mean absolute error (MAE) were reported for 10-fold cross validation, whereby 10 randomly and equally sized sub datasets were partitioned, retaining 2 samples per sub dataset for internal test validation. RMSE was the key metric used by the Bayesian optimisation algorithm in order to improve upon each new hyper-parameter configuration it explored. The pipeline was repeated until an inflection point was reached and automatically recognised as no further improvement in validation via early stopping, and the parameters that resulted in the best cross validated RMSE over all the fits was then used to save a final model configuration. Lastly, the retained 20 sample test dataset was used to evaluate the final model’s performance on unseen data.

In order to optimise the pipeline for each phenolic parameter (total phenols, total condensed tannins, total anthocyanins, colour density and polymeric pigments), four main tests were conducted including running the complete pipeline, the pipeline without synthetic samples, the pipeline with synthetic samples but without region selection and lastly, the pipeline without region selection nor synthetic samples. The optimal pipeline parameters were chosen unique to each phenolic model. Each of the four tests were run several times in order to evaluate the optimal number of components in principal component analysis (PCA). The average train and test scores per number of PCA components were evaluated with a focus on optimal decomposition coupled with model stability. Six components were chosen due to this being consistently optimal for all phenolic models and was thereafter inserted into the pipeline as a fixed hyper-parameter (**Figure 1**). Once the optimal parameters were obtained, further model development involved adjusting the phenolic ranges to eliminate minority sample groups from negatively impacting model accuracy, as well as outlier identification and removal.

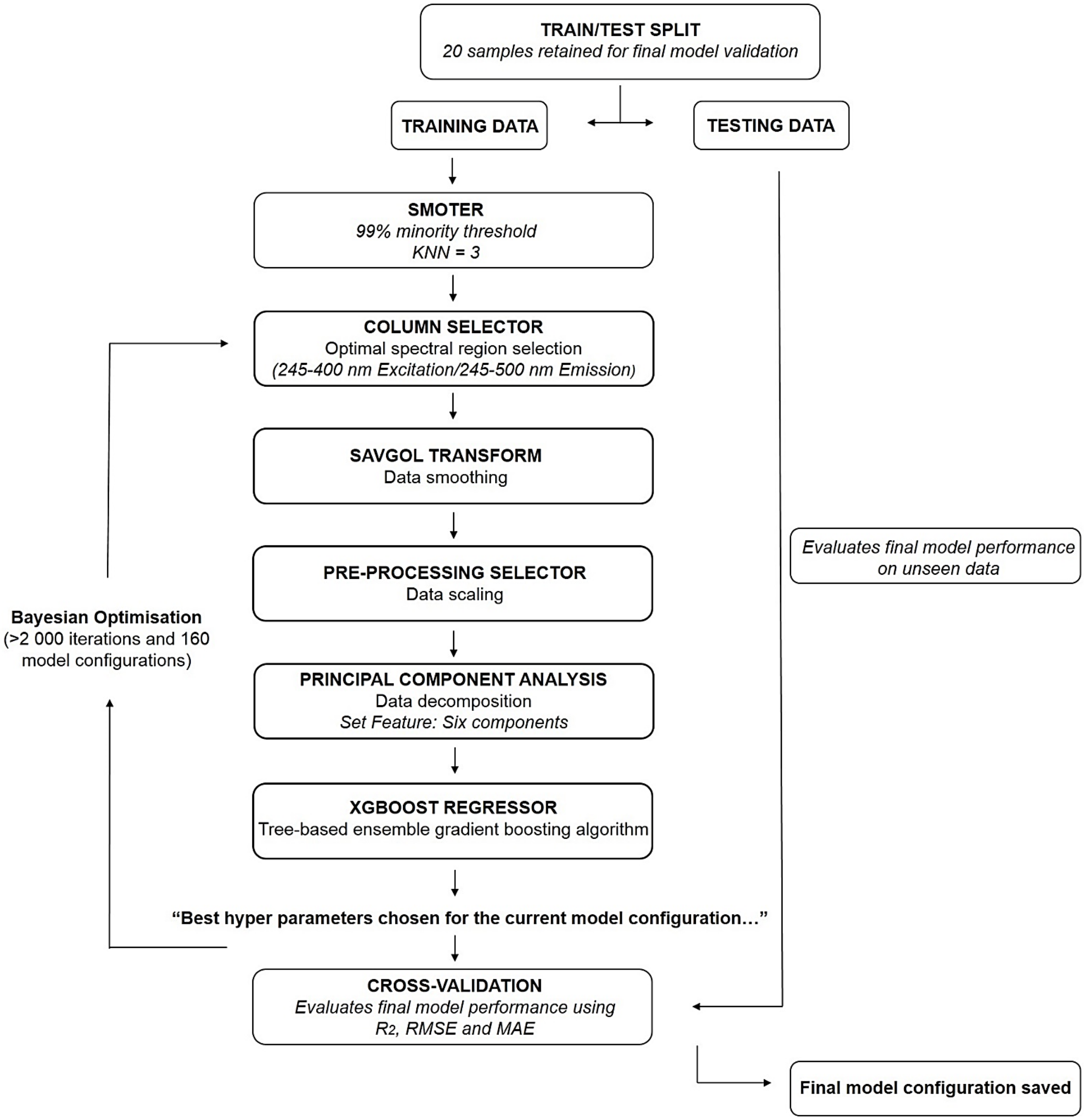
TRAIN/TEST SPLIT
20 samples retained for final model validation
TRAINING DATA
TESTING DATA
SMOTER
99% minority threshold
KNN = 3
COLUMN SELECTOR
Optimal spectral region selection
(245-400 nm Excitation/245-500 nm Emission)
SAVGOL TRANSFORM
Data smoothing
PRE-PROCESSING SELECTOR
Data scaling
PRINCIPAL COMPONENT ANALYSIS
Data decomposition
Set Feature: Six components
XGBOOST REGRESSOR
Tree-based ensemble gradient boosting algorithm
Bayesian Optimisation
(>2 000 iterations and 160 model configurations)
Evaluates final model performance on unseen data
"Best hyper parameters chosen for the current model configuration…"
CROSS-VALIDATION
Evaluates final model performance using
R2, RMSE and MAE
Final model configuration saved

Figure 1. Schematic diagram of the machine learning pipeline.

2.8. CLASSIFICATION

PARAFAC performed in Matlab, and PCA and neighbourhood component analysis (NCA) performed in Python were the techniques used to evaluate the classification and discrimination abilities of fluorescence spectroscopy. PARAFAC scores obtained per component were plotted against each other [11] focusing on the four main cultivar types included in this study (Cabernet Sauvignon, Merlot, Pinotage and Shiraz) as well the sample state of either fermenting must or wine. PCA was conducted in a similar manner to PARAFAC. NCA was conducted using linear discriminant analysis (LDA) as the linear transformation initialisation method and due to the large variation in number of samples per cultivar, classification was conducted on cultivars with more than or equal to 5, 8, 14 and 20 samples, respectively. NCA was repeated with a focus on classifying according to the sample state of either fermenting must or wine as well as on fermenting musts and wine separately. Leave-one-out cross validation was conducted per set of NCA with score values used to determine classification accuracy.

3. RESULTS AND DISCUSSION

3.1. WINE EXCITATION-EMISSION MATRICES (EEMS)

Figure 2 is an example of a pre-processed EEM belonging to a randomly chosen Cabernet Sauvignon sample from this study. Two different spectral regions can be observed as a result of the fluorescent properties of wine previously reported in literature [11,17]. Excitation between the more energetic wavelengths of 250 and 290 nm results in emission between 300 and 430 nm, while excitation at wavelengths longer than 300 nm results in emission between 360 and 450 nm [11,17]. **Figure 3** is an integrated depiction adapted from literature indicating the characteristic excitation and emission wavelengths of important phenolic compounds [11]. The non-flavonoid family including phenolic acids (cinnamic-like and benzoic-like), phenolic aldehydes and stilbene-like compounds extends between the ranges of excitation 260-330 nm and emission 320-440 nm. Gentisic acid possesses a unique fluorescence in that it deviates further right of the EEM compared to the rest of the non-flavonoids. The flavonoid family is split into two unique regions with flavonols extending between excitation 260-268 nm and emission 370-422 nm, and flavan-3-ols occurring within excitation 278-290 nm and emission 310-360 nm. Apart from polyphenols, other naturally occurring fluorescent compounds in fermenting musts and wine, such as vitamins and amino acids, have previously been reported [17,32]. The fluorescent properties of the amino acid tryptophan have been included, as reported [33]. **Figure 3** is merely an approximate representation as the excitation-emission regions illustrated below are reported for compounds in dilution measured using the conventional right-angled technique, and spectral shifts, band fluctuations and quantum yield changes may occur when measured within the unaltered wine matrix [11].

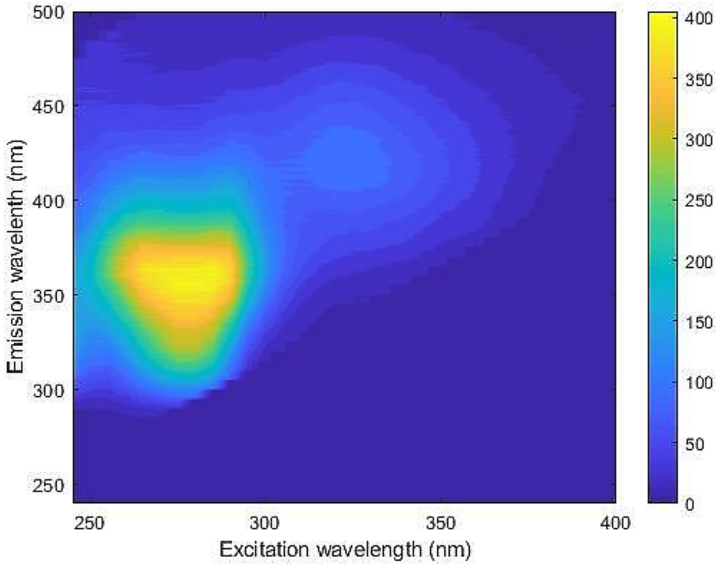

Figure 2. Excitation-emission matrix of a fermenting Cabernet Sauvignon sample included in this study (Sample 1) with the scale bar representing fluorescence intensity.

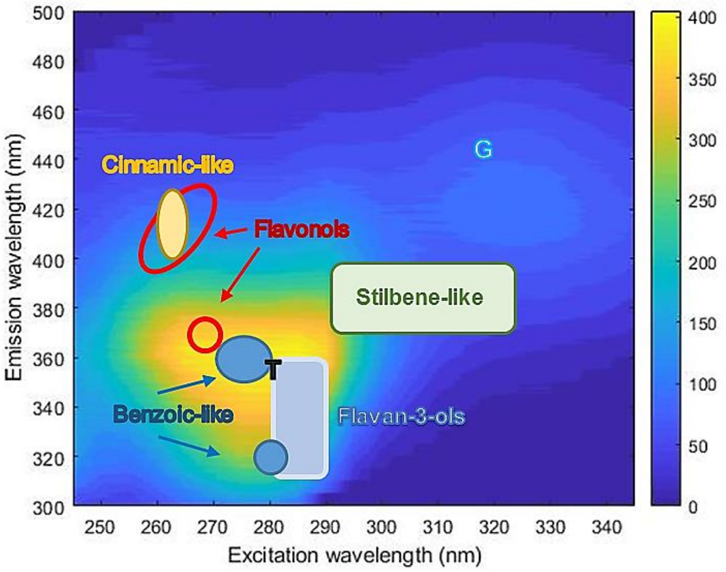

Figure 3. Excitation-emission matrix of a fermenting Cabernet Sauvignon sample included in this study (Sample 1) indicating the fluorescent properties of wine fluorophores adapted from literature [11]. G and T represent gentisic acid and tryptophan, respectively.

	Total Phenols Index	Total Condensed Tannins (mg/L)	Total Anthocyanins (mg/L)	Colour Density (AU)	Polymeric Pigments (AU)
Maximum	126.10	2912.08	1306.44	42.52	8.09
Minimum	5.11	731.44	9.26	1.89	0.24
Average	44.50	1474.22	350.98	14.01	1.80
Standard deviation	18.02	425.74	194.71	6.06	1.13

Table 1. Maximum, minimum, standard deviation and average values per spectrophotometric analysis reference method.

Table 1 illustrates the phenolic variability achieved during sample collection of both fermenting musts and wine samples. All spectrophotometric methods were performed within a coefficient of variation less than 5%, considered acceptable for reproducibility. The final wine phenolic profile is the result of complex chemical interactions influenced by numerous factors such as those influencing the chemical composition of the grape berry as well as the viticultural and oenological practices implemented throughout processing [1]. This naturally high variability obtained illustrates the importance of including an extensive dataset during model development in order to sufficiently challenge and train the model on diverse ranges of phenolic levels. Introducing high sample variability aids in building robust calibration models able to make accurate predictions on future samples.

3.2. PARALLEL FACTOR ANALYSIS (PARAFAC)

PARAFAC is a trilinear decomposition modelling technique resulting in components (score and loading vectors) that are representative of signals from individual fluorophores. The optimal number of components was chosen to be four, based on the core consistency diagnostic (CORCONDIA) and explained variance obtained for non-negativity constrained models [19] as well as corresponding with results from previous red wine PARAFAC analyses in which components were tentatively correlated with phenolic compounds [17,39]. Visual inspection of the loadings was performed to confirm the optimal number of components as well as to remove spectral artifacts interfering with the model stability, resulting in a reduced spectral region of 245-340 nm excitation and 265-500 nm emission with a final three-way array of 289 x 20 x 470 (*samples* x *excitation wavelengths* x *emission wavelengths*). Split-half analysis was conducted to validate the uniqueness and stability of the final model.

Figure 4 shows the final model scores obtained per sample for each PARAFAC component, as well as the excitation and emission loadings per PARAFAC component. Score values are estimates of the relative concentrations of the responsible fluorophore and can be used to build univariate calibration models or determine relationships contained within the fluorescence information for potential clustering [17,19]. Components 1 to 4 have been tentatively assigned to their responsible fluorophores in literature by correlating the resulting PARAFAC component excitation and emission peaks with HPLC measurements and bibliographic information [11,39]. Component 1 is characterised by an excitation peak around 260 nm with an emission shoulder at 370 nm and peak at 390 nm, and has been suggested as representing phenolic aldehydes, benzoic-like acids, myricetin and trans-resveratrol [11] and caffeic acid [39]. Component 2 is characterised by an excitation peak around 280 nm and emission peak around 320 nm. This second component has been consistently matched with monomeric flavan-3-ols, catechin and epicatechin, with high correlations achieved for catechin ($R^2 = 0.9221$) and epicatechin (R2 = 0.8761) [17] as well as the sum of both ($R^2 = 0.8468$) [13]. Component 3 consists of an excitation peak between 320-330 nm and an emission peak around 420 nm, while component 4 is characterised by an excitation shoulder at 270 nm and peak at 280 nm with an emission peak at 370 nm. Schueuermann *et al.* [39] proposed cinnamic-like acids, caffeic and p-coumaric, responsible for component 3 while p-coumaric acid, gentisic acid and stilbene-like non-flavonoids were proposed by Airado-Rodŕiguez *et al.* [17]. Component 4 has been suggested as benzoic-like acids as well as tryptophan [11,39]. The complexity of the wine matrix results in PARAFAC components most likely corresponding to several different fluorophores or those within the same chemical group rather than individual compounds. No correlations were found between the obtained score values and the reference data per phenolic parameter (Suplemmentary information S1). Despite the potential for component 2 to be well correlated with total condensed tannins, the best R^2 value obtained after linear regression was 0.21. In the context of this study, PARAFAC was unsuccessful in building calibrations for such broad phenolic parameters such as total condensed tannins versus the successful correlations achieved for pure compounds of catechin or epicatechin [13,17]. The structural similarity of the phenolic classes and difficulty in separating them into their singular structures based on their PARAFAC components may be hindering the predictive ability of regression modelling. Conducting PARAFAC on fermenting musts and wine separately did not improve upon results.

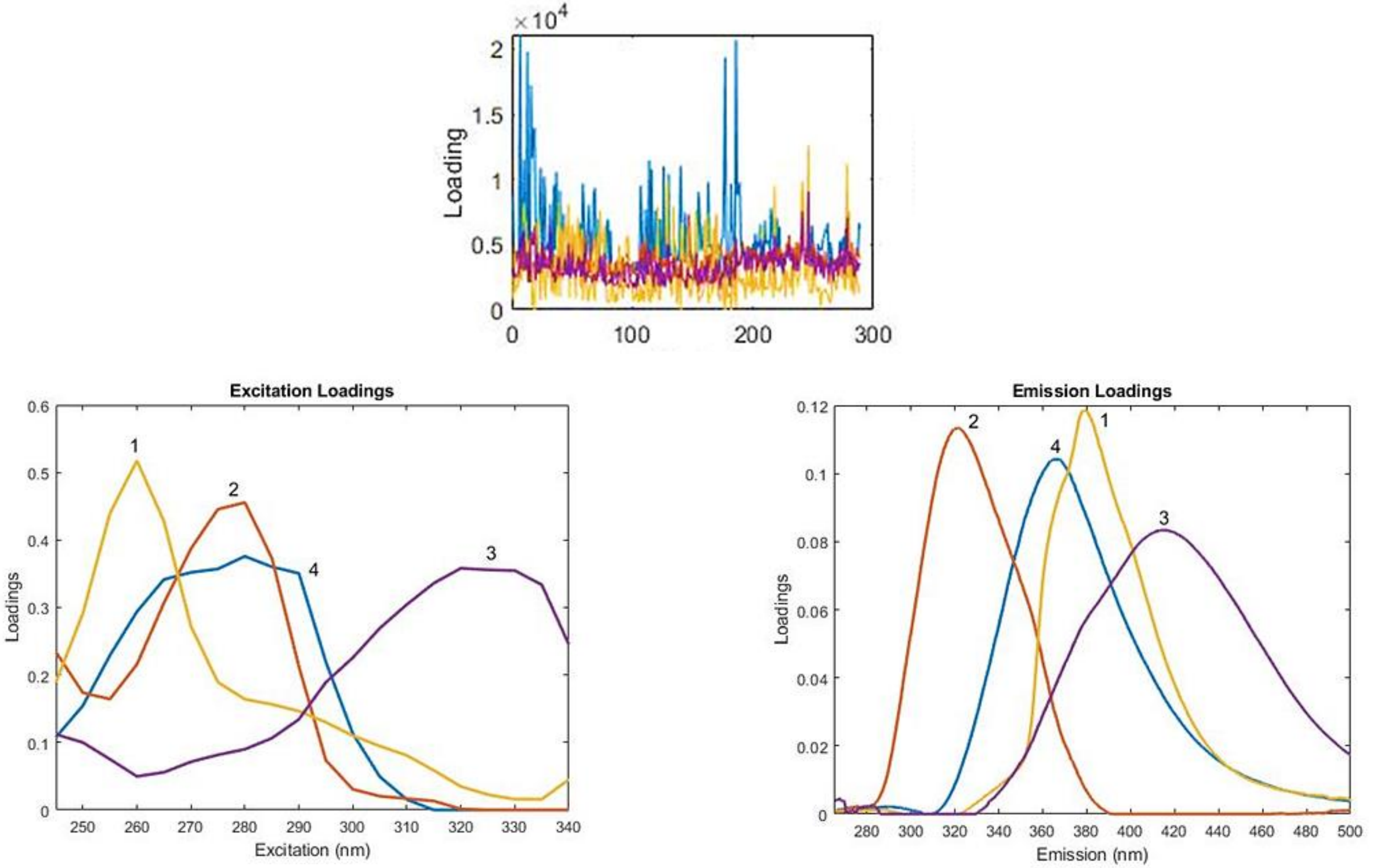

Figure 4. Score values per sample (Mode 1), excitation loadings and emission loadings for the four component, non-negativity constrained PARAFAC model. Component 1(yellow), 2(red), 3(purple) and 4 (blue).

3.3. MACHINE LEARNING

Conventional linear regression in the form of principal component regression (PCR) and partial least squares regression (PLSR) was performed on the fluorescence excitation-emission spectra and reference data. These methods proved unsuccessful despite exploring fluorescent region selection, phenolic range manipulation and outlier removal, with poor calibration and validation scores (data not shown). This suggested a complex dataset (three-dimensional fluorescence data configuration of intensity *x* emission *x* excitation against phenolic reference data) requiring more intensive data handling and the exploration of machine learning algorithms. The decision behind using a boosting modelling technique, such as XGBoost, involved the beneficial linear collection of numerous sequentially modelled regression trees rather than a single model of best fit as with simpler regression methods [28]. Each successive tree optimises on the residuals of the previous tree's predictions and thereby minimises the loss of predictive ability from previously sub-optimal models [40,41]. Gradient boosting is a highly effective technique for classification and regression problems and a favoured option throughout the data science community. This can be seen in the preferred choice of machine learning algorithms used on Kaggle, the largest data science community platform and machine learning competitive scene [42].

Briefly, a five-step machine learning pipeline was built consisting of fluorescent region selection, data smoothing and scaling, data decomposition with PCA and lastly, the XGBoost regressor (**Figure 1**). The minority over-sampling technique in the form of a SMOTER algorithm applied to the training sub dataset following the train/test split, proved useful in creating a more balanced training model for a widely variable input dataset of fermenting musts and wines. Six principal components showed the most optimal model stability and highest prediction accuracy for all phenolic parameters and was thereafter inserted as a set feature for further model development. Generally, calibration models should be cautiously considered with regards to overly optimistic results. Internal validation in the form of 10-fold cross validation as well as the evaluation of the final model on a retained validation dataset was therefore performed in order to reduce these risks. Each phenolic parameter was individually explored to determine the most optimal pipeline resulting in the highest prediction accuracy and model stability. **Table 2** shows the prediction accuracy metrics and characteristics of the best models per phenolic parameter. Once the most optimal pipeline parameters were determined, the pipeline was re-run several times to allow for outlier removal and refinement.

The best total phenols model depicted in **Figure 5** (R^2 = 0.81; RMSEV = 7.16; MAEV = 5.39) made use of region selection between 260-360 nm excitation and 370-400 nm emission which overlaps the flavonols and stilbene-like regions as represented in **Figure 3**. Poor prediction accuracy and unstable models were found when trying to incorporate the entire phenolic region as referenced in literature. Due to a minority of samples with high phenolic values, samples above 80 index units were removed as the model struggled to predict above this threshold.

The best total condensed tannins model (**Figure 6**) made use of region selection between 285-340 nm excitation and 290-350 nm emission, overlapping with the flavan-3-ols region depicted in **Figure 3**. Samples with tannin levels above 2300 mg/L were removed as the model struggled to predict above this minority group of samples. An R^2 of 0.89, RMSEV of 172.37 and MAEV of 129.14 were obtained. The best total anthocyanins model (**Figure 7**) required removing samples with levels above 800 mg/L and made use of region selection between 280-300 nm excitation and 330-380 nm emission which correlates well with the fluorescence of malvidin-3-glucoside. Prediction scores of R^2 = 0.8, RMSEV = 76.57 and MAEV = 61.57 were obtained. Poorer but stable models were built for colour density (**Figure 8**) and polymeric pigments (**Figure 9**), the metrics of which are reported in **Table 2**. No ideal region could be selected for both models and little improvement was observed with outlier removal and range manipulation. Due to a minority of samples in the higher ranges, samples above 25 absorption units and above 4 absorption units were removed for colour density and polymeric pigments, respectively. The inability to develop a promising regression model for colour density may be due to the characteristics of colour density as a metric. Red wine colour experiences numerous transitions over time as a result of chemical reactions between anthocyanins and other phenolic compounds [5]. The widely used Glories method [24] is an estimation of total colour by using the sum of absorbances at three wavelengths, namely 420 nm (yellow colouration), 520 nm (red colouration) and 620 nm (blue colouration). The excitation-emission matrix chosen for this study therefore may not have encompassed all responsible compounds, provided they possess fluorescent abilities, and a summation of fluorescent measurements at these absorbances should be considered for future modelling. The poorer prediction accuracy metrics obtained for the polymeric pigments model may be due to the chosen excitation-emission matrix not encompassing the fluorescent regions of such pigments, as has been identified by the novel fluorescence approach developed using a fluorescence ratio of F700/F560 [34]. However, the unbalanced dataset of 190 fermenting musts and 110 wines may be affecting model calibration due to a minority group of samples with higher polymeric pigment levels (only 40 wine samples with levels above 3 absorption units).

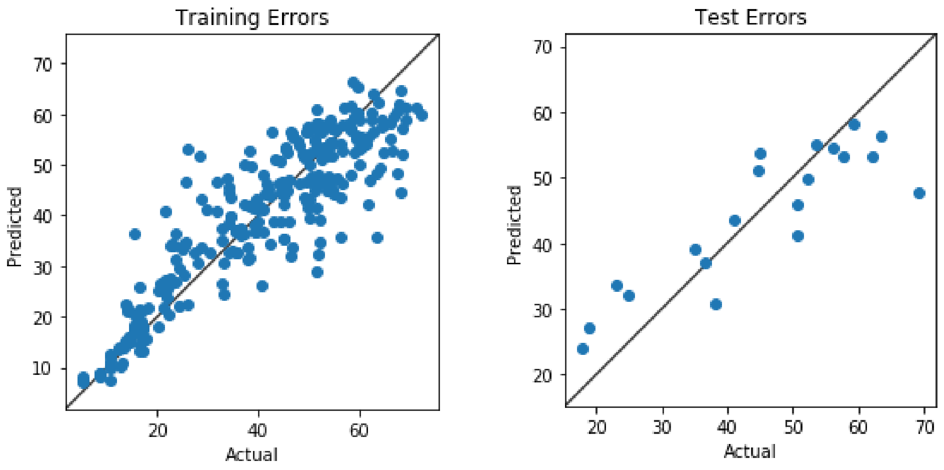

Figure 5. Total phenols regression plots, calibration model (left) and validation set (right).

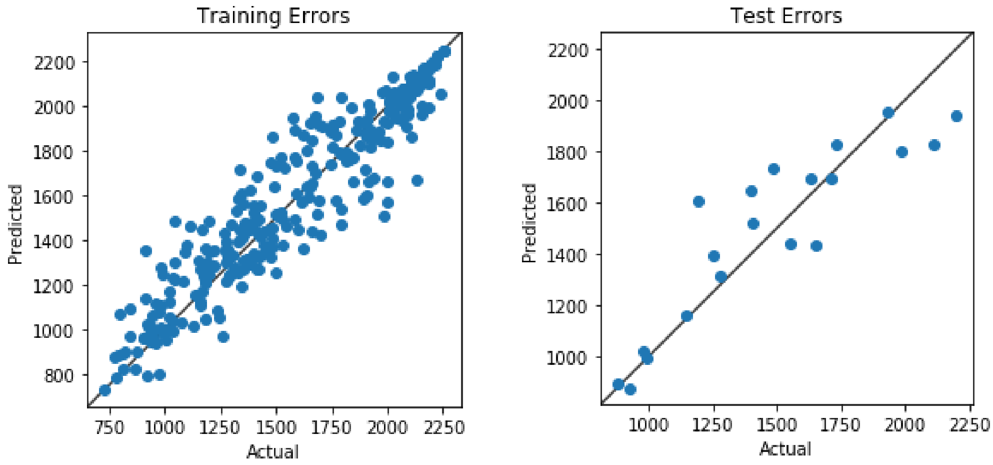

Figure 6. Total condensed tannins (mg/L) regression plots, calibration model (left) and validation set (right).

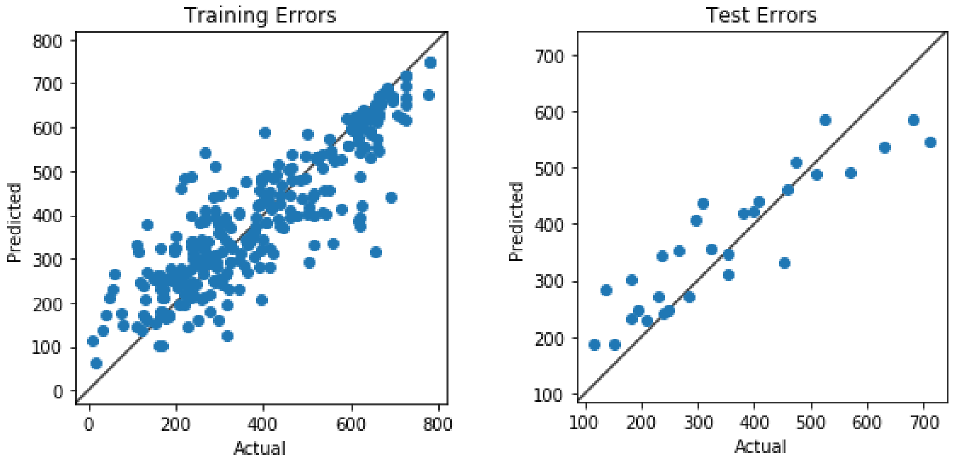

Figure 7. Total anthocyanins (mg/L) regression plots, calibration model (left) and validation set (right).

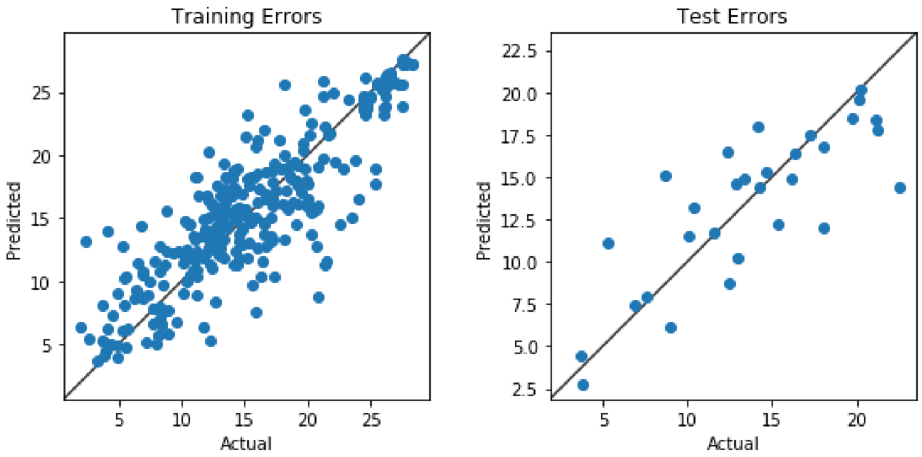

Figure 8. Colour density (AU) regression plots, calibration model (left) and validation set (right).

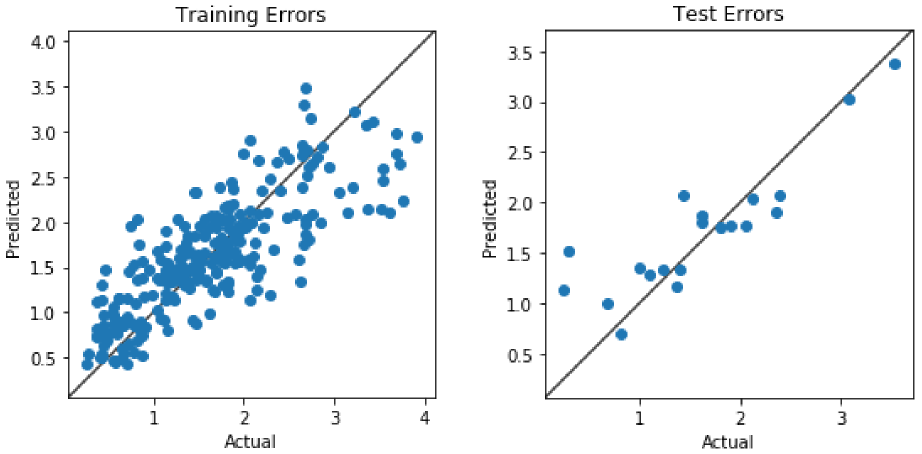

Figure 9. Polymeric pigments (AU) regression plots, calibration model (left) and validation set (right).

Cultivar based models were explored per phenolic parameter for the four main cultivars, Cabernet Sauvignon, Shiraz, Merlot and Pinotage. The only model with promising results was built for Cabernet Sauvignon and total condensed tannins with average R^2 train and test scores of 0.78 and 0.81, respectively. This may be a result of high tannin levels characteristic of the cultivar as well as an equally balanced dataset of fermenting musts and wine. Only 45 samples were used in the model and therefore only show promise as to the potential of building a cultivar-based model.

Due to differences in fluorescence between fermenting musts and wine suggested in PCA (**Figure 10**), age-based models were explored and the prediction accuracy metrics reported in **Table 3**. Overall, models built using only fermenting musts for total phenols, total condensed tannins and polymeric pigments performed slightly better than those built with only finished wines. This could be a result of too few wine samples with too much variability creating large gaps unable to be adequately trained on despite implementing the SMOTER algorithm. The models built using finished wine samples also appear to be more unstable, specifically with regards to large differences in coefficient of correlation (R^2) between calibration and validation, as seen with total phenols and total condensed tannins (**Table 3**). The fermenting-based models for total condensed tannins and polymeric pigments in **Table 3** possess slightly better prediction accuracy metrics than the models built using all samples and show potential for quantifying other fermenting samples more accurately. Interestingly, the wine-based models built for total anthocyanins and colour density seemed to perform slightly better when looking at RMSE and MAE, however the differences in R^2 should indicate further validation is required. Differences in performance when modelling on fermenting musts and wine separately when compared to the best phenolic models reported in **Table 2** may be a result of the random sampling technique used within the machine learning pipeline or the unique behaviour of specialised models built for a specific sub dataset. Overall, the best phenolic parameter models built using all samples may be more promising in terms of generalisability and the ability to predict any sample, regardless of the stage in red wine production, as opposed to more specialised models built for a specific task, such as fermenting or wine-based models, which may become over-fitted and perform poorly on unseen data.

Several considerations are important for optimal model development and the acceptance of the subsequently obtained models. Including more samples per cultivar as well as a more balanced dataset of fermenting musts and wine may help in model development. Model considerations include over-fitting and over-validating. Cross validation is incorporated to reduce these risks, however, unidentified noise or influences from the fluorescence spectrophotometer may be fitted on during calibration. Additionally, the retained validation set may potentially be from the same cultivar, the same day of analysis or the same level of fermentation and therefore over confidently validate the model.

Table 2. Prediction accuracy metrics (R^2, RMSE and MAE) and pipeline parameters for the best calibration model per phenolic parameter.

	R^2cal	R^2val	RMSEC	RMSEV	MAEV	Excitation/Emission Region (nm)
Total Phenols	0.81	0.77	5.71	7.16	5.39	260-360/370-400
Total Condensed Tannins (mg/L)	0.89	0.80	104.03	172.37	129.14	285-340/290-350
Total Anthocyanins (mg/L)	0.80	0.77	60.67	76.57	61.57	280-300/330-380
Colour Density (AU)	0.68	0.64	2.46	3.10	2.28	245-400/245-500
Polymeric Pigments (AU)	0.64	0.66	0.63	0.49	0.39	245-400/245-500

R^2cal: coefficient of determination in calibration; R^2val: coefficient of determination in validation; RMSEC: root mean square error of calibration; RMSEV: root mean square error of validation; MAEV: mean absolute error of validation.

Table 3. Prediction accuracy metrics (R^2, RMSE and MAE) for fermenting musts and finished wine calibration models per phenolic parameter.

	R^2cal	R^2val	RMSEC	RMSEV	MAEV
			Total Phenols		
Fermenting	0.70	0.66	6.56	7.45	5.74
Wine	0.74	0.37	3.81	7.77	6.17
			Total Condensed Tannins (mg/L)		
Fermenting	0.82	0.78	95.81	128.24	103.20
Wine	0.69	0.34	122.85	241.13	190.09
			Total Anthocyanins (mg/L)		
Fermenting	0.72	0.77	75.22	89.89	72.18
Wine	0.71	0.55	36.51	60.06	51.28
			Colour Density (AU)		
Fermenting	0.78	0.53	2.65	4.20	3.34
Wine	0.72	0.61	2.03	2.38	2.25
			Polymeric Pigments (AU)		
Fermenting	0.62	0.57	0.27	0.33	0.22
Wine	0.60	0.79	0.49	0.42	0.35

R^2cal: coefficient of determination in calibration; R^2val: coefficient of determination in validation; RMSEC: root mean square error of calibration; RMSEV: root mean square error of validation; MAEV: mean absolute error of validation.

3.4. CLASSIFICATION

Unique fluorescent fingerprints of wine have been identified for their potential to classify samples based on cultivar type, wine style or appellation [11,15,43]. The three methods explored in this study for the classification of cultivar type and sample state (fermenting must or wine) included PARAFAC, PCA and NCA. PARAFAC scores were unsuccessful in distinguishing between cultivar or sample state. PCA did not clearly distinguish between cultivars but showed clear distinction between fermenting musts and wine (**Figure 10**). NCA was explored due to its success in achieving better classification results compared to other dimensionality reduction techniques, such as PCA and linear discriminant analysis (LDA), because of its explicit encouragement of local separation between classes [44]. Due to large variation in the number of samples per cultivar, classification was conducted on cultivars with more than or equal to 5, 8, 14 and 20 samples, respectively. Leave-one-out cross validation was conducted per set of NCA analysis with scores reported in **Table 4**.

The two best cultivar classification scores were achieved for 9 different cultivars (> 5 samples) (**Figure 11**) and the four main cultivars (>20 samples) included in this study (**Figure 12**). When distinguishing between fermenting musts and wine, the highest cross validation score of 0.82 was achieved for the four main cultivars (>20 samples) (**Figure 13**). Due to the difference in fluorescence suggested in the stretched appearance of the cultivar classes (**Figure 12**) and confirmed with PCA, NCA was conducted on fermenting musts and wines separately. Overall, the cultivar classification ability was stronger for fermenting musts compared to wine (**Table 4**). **Figures 14** and **15** show the best clustering and classification achieved by analysing only fermenting musts. This improved classification for fermenting musts compared to wines highlights the uniqueness of cultivar types before undergoing processing. The final phenolic composition of a wine is a complex chemical matrix influenced by several factors including viticultural practices, different terroirs and various winemaking techniques implemented throughout fermentation and ageing, and therefore clarifies the poorer results for classifying wines purely based on cultivar [1,17]. Additionally, the initial composition of grape must may possess higher levels of fluorescent compounds such as vitamins and amino acids before being metabolised by yeast cells during fermentation, while the phenolic composition changes occurring throughout fermentation may also suggest greater fluorescence of monomeric compounds compared to the polymerised compounds found later in wine. Spectral considerations include a reduced fluorescence intensity from darker samples, the result of which is obtained following increased anthocyanin extraction during fermentation [17,32]. Interestingly, the Pinotage, Malbec and Shiraz blend (PMS) in **Figure 14** is situated relatively central to each of the corresponding pure cultivars included in the fermenting blend and suggests the potential of fluorescence spectroscopy in determining the constituents of blends which may be helpful in authentication and quality control by industry bodies.

Figure 15 is an integrated depiction of the highest cross validated cultivar classification for the four main cultivars (>20 samples) combined with three-dimensional EEMs of each cultivar. Each sample depicted was chosen based on their phenolic levels to illustrate the unique fluorescent fingerprint per cultivar despite possessing similar phenolic levels (**Table 5**). Although showing a similar general three-dimensional fluorescent shape, each cultivar has their own characteristic peak within the EEM and level of fluorescence intensity, with Pinotage having the lowest of the four. Pinotage also exhibits tighter clustering in **Figures 11** to **15** compared to other cultivars. This may be a result of a particularly unique phenolic composition compared to other cultivars [45]. When investigating the fluorescent intensities of Pinotage samples, more stable fluorescent levels between fermenting musts and wines were observed compared to other cultivars which experienced more extreme variations in fluorescent intensities, the cause of which has not been clearly identified and requires further investigation.

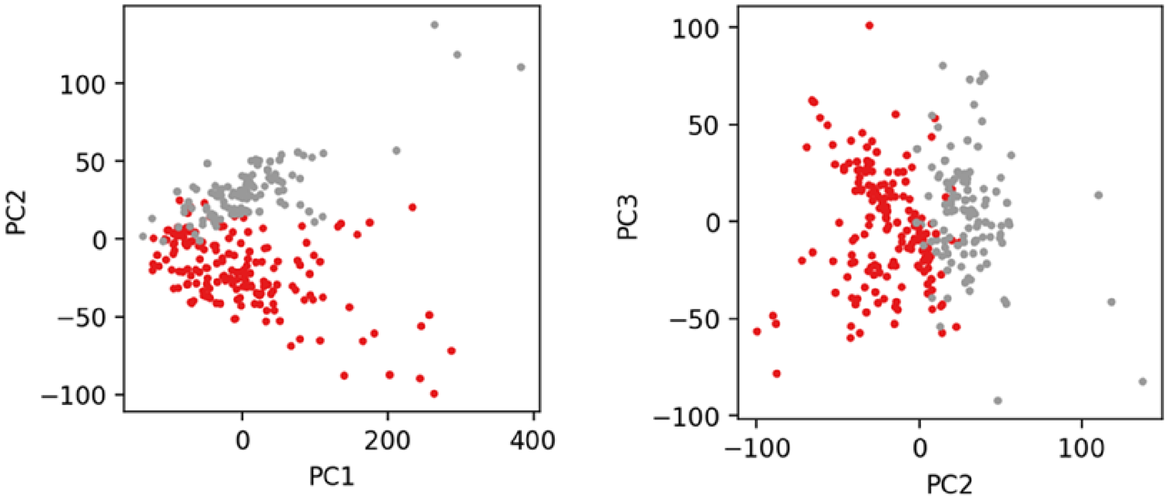

Figure 10. Principal Component Analysis (PCA) plot showing fermenting musts (red) and finished wines (grey).

Table 4. Leave-one-out cross validation scores per neighbourhood component analysis (NCA) conducted for cultivar classification, sample state classification, fermenting musts and wine classification.

Number of samples per cultivar	Cross Validation Score
Cross validation scores for cultivar classification using all samples	
≥ 5	0.84
≥ 8	0.80
≥ 14	0.72
≥ 20	0.86
Cross validation scores for sample state classification (fermenting musts and wine)	
≥ 5	0.79
≥ 8	0.78
≥ 14	0.77
≥ 20	0.82
Cross validation scores for cultivar classification of fermenting musts only	
≥ 5	0.87
≥ 20	0.93
Cross validation scores for cultivar classification of wine only	
≥ 5	0.76
≥ 20	0.79

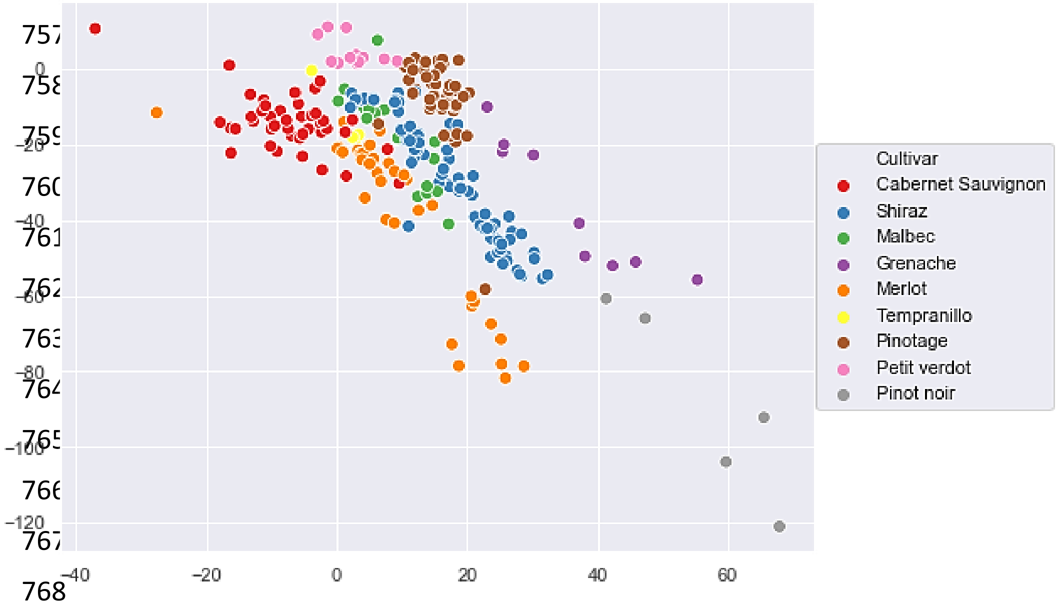

Figure 11. Cultivar classification using NCA for cultivars with 5 or more samples (fermenting musts and wine) with a cross validation score of 0.84.

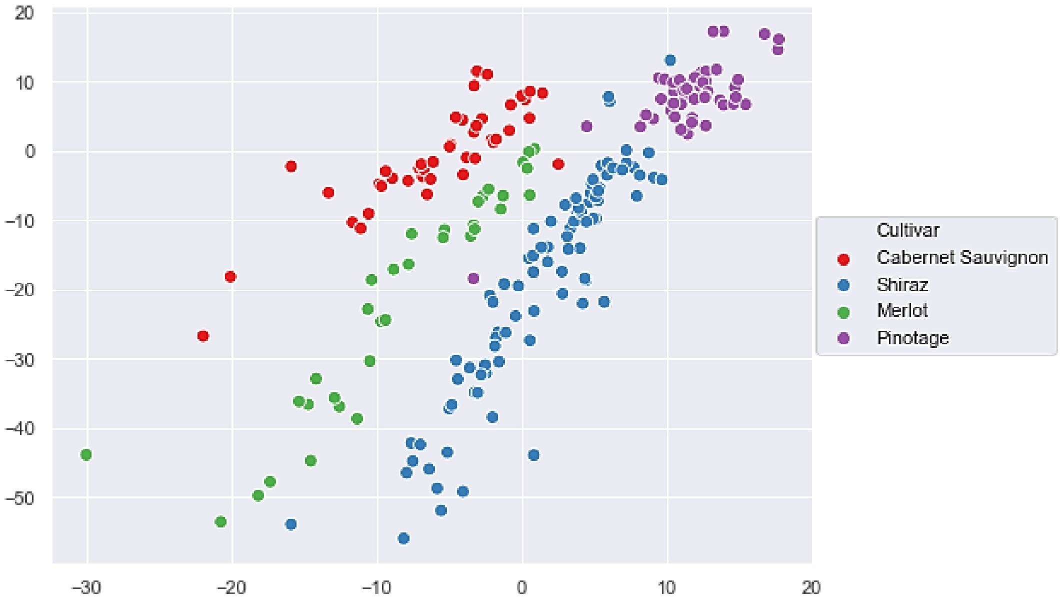

Figure 12. Cultivar classification using NCA for cultivars with 20 or more samples (fermenting musts and wine) with a cross validation score of 0.86.

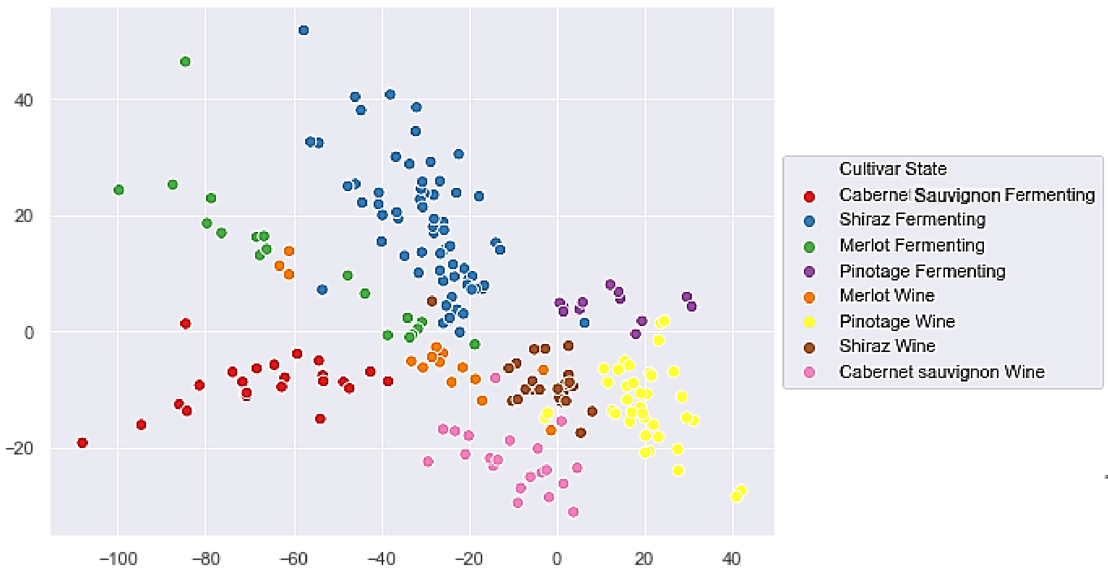

Figure 13. Cultivar classification using NCA for the four main cultivars (≥ 20 samples) distinguishing between fermenting musts and wine with a cross validation score of 0.82.

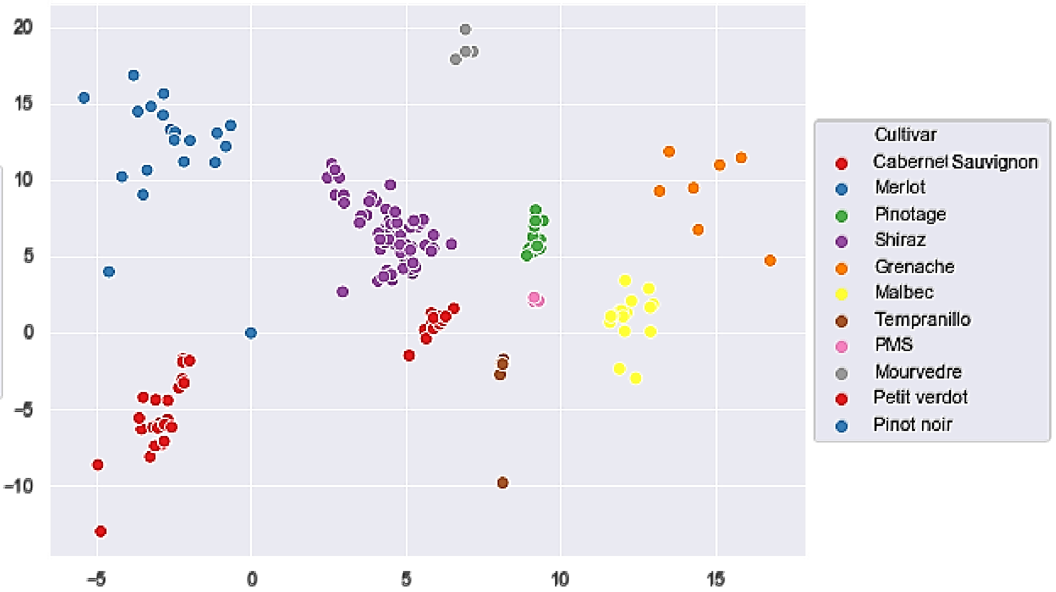

Figure 14. Cultivar classification using NCA for cultivars with 5 or more samples on only fermenting musts with a cross validation score of 0.87.

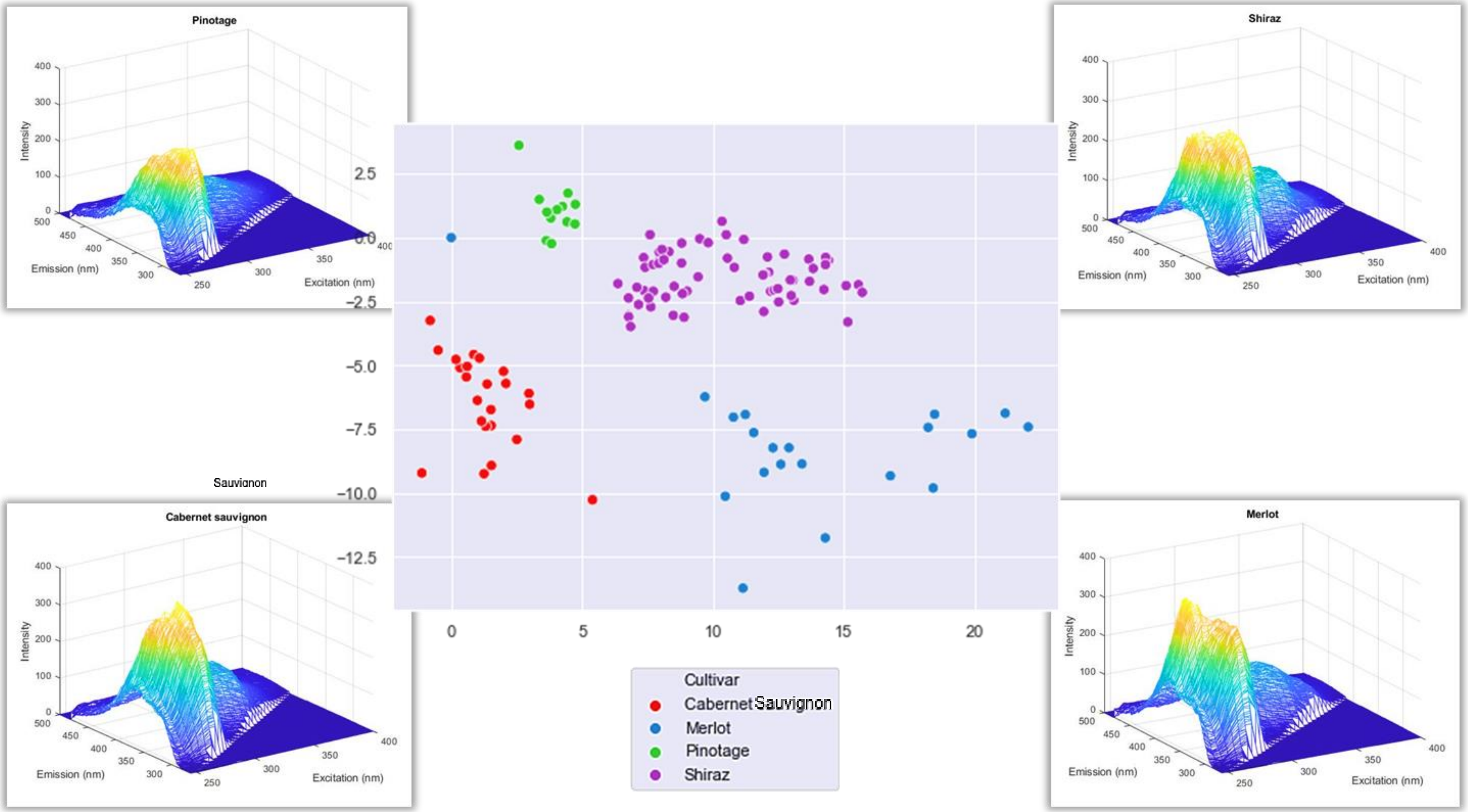

Figure 15. Cultivar classification using NCA for the four main cultivars (≥ 20 samples) on only fermenting musts with a cross validation score of 0.93. Three-dimensional excitation-emission matrices of phenolically similar samples corresponding to each cultivar.

Table 5. Spectrophotometric analysis measurements showing the phenolic similarity between wines made from different cultivars namely, Merlot, Shiraz, Cabernet Sauvignon and Pinotage (samples 293, 209, 292 and 227).

	Total Phenols	Total Condensed Tannins (mg/L)	Total Anthocyanins (mg/L)	Colour Density (AU)	Polymeric Pigments (AU)
Merlot	59.95	1902.66	304.93	11.02	2.01
Shiraz	59.50	1974.06	324.30	16.50	2.25
Pinotage	59.15	1908.30	313.43	10.71	2.03
Cabernet Sauvignon	60.10	1901.09	231.44	16.67	3.14
Average	59.53	1928.34	314.22	12.74	2.10
Standard deviation	0.33	32.41	7.93	2.66	0.18

4. CONCLUSION

Monitoring phenolic extraction throughout fermentation and ageing may aid in decision-making during red wine production. This study showed the potential of front-face fluorescence spectroscopy coupled with chemometrics to quantify important phenolic parameters in fermenting musts and wine. PCR, PLSR and PARAFAC were explored but produced poor results and highlighted the need for more complex data handling techniques. Calibration models built using a gradient boosting technique, XGboost, were successful for the quantification of total phenols, total condensed tannins and total anthocyanins. The errors and coefficients of determination obtained in this study are in line with those previously reported for other spectroscopy applications such as UV-Visible or IR further validating the suitability of fluorescence spectroscopy for this application [2,46,47,48,49]. However, the incorporation of more samples within minority sample groups as well as obtaining a more balanced dataset of different cultivar types, fermenting musts and wines may improve upon model development and therefore the reported results. Additionally, the wide field of chemometrics allows for the use of other statistical analysis methods not explored in this study which may yield better results. The identification of fluorescent regions for each of the phenolic parameters optimises fluorescence analysis for a reduced analysis time and the development of accurate predictive models using front-face fluorescence spectroscopy may allow for their incorporation into future optical portable devices or automated systems, able to analyse samples directly from their fermentation vessels or barrels. This approach could serve as an alternative to IR portable devices that work similarly capturing the reflected light of the wine samples and proved to also quantify phenolic content successfully. Moreover, the fact that fluorescence signals rely on the excitation of the fluorophores with less expensive and well-developed UV-Visible technology makes this technology also cost-wise interesting. Additionally, this study provides a novel approach using NCA for the classification of South African red wine cultivars as well as proposing the potential for analysing and possibly determining the constituents of red wine blends, both of which may be useful in authentication and quality control.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Winetech South Africa for funding and support under the grant number (JT-NP07)

5. REFERENCES

[1] Garrido, J. and Borges, F., 2013. Wine and grape polyphenols - A chemical perspective. Food Research International, 54(2), 1844–1858.

[2] Aleixandre-Tudo, J. L., Nieuwoudt, H., Olivieri, A., Aleixandre, J. L. and du Toit, W., 2018. Phenolic profiling of grapes, fermenting samples and wines using UV-Visible spectroscopy with chemometrics. Food Control, 85, 11–22.

[3] Vidal, S., Francis, L., Guyot, S., Marnet, N., Kwiatkowski, M., Gawel, R., Cheynier, V., and Waters, E., 2003. The mouth-feel properties of grape and apple proanthocyanidins in a wine-like medium. Journal of the Science of Food and Agriculture, 83(6), 564–573.

[4] Monagas, M., Bartolomé, B. and Gómez-Cordovés, C., 2005. Updated knowledge about the presence of phenolic compounds in wine. Critical reviews in food science and nutrition, 45(2), 85–118.

[5] Harbertson, J. F. and Spayd, S., 2006. Measuring phenolics in the winery. American Journal of Enology and Viticulture, 57(3), 280-288.

[6] Aleixandre-Tudo, J. L., Buica, A., Nieuwoudt, H., Aleixandre, J. L. and du Toit, W., 2017. Spectrophotometric analysis of phenolic compounds in grapes and wines. Journal of Agricultural and Food Chemistry, 65(20), 4009–4026.

[7] Romera-fernández, M. Berrueta, L. A., Garmón-lobato, S., Gallo, B., Vicente, F. and Moreda, J. M., 2012. Talanta Feasibility study of FT-MIR spectroscopy and PLS-R for the fast determination of anthocyanins in wine. Talanta, 88, 303–310.

[8] Dambergs, R. G., Mercurio, M. D., Kassara, S., Cozzolino, D. and Smith, P. A., 2012. Rapid measurement of methyl cellulose precipitable tannins using ultraviolet spectroscopy with chemometrics: Application to red wine and inter-laboratory calibration transfer. Applied Spectroscopy, 66(6), 656-664.

[9] Daniel, C., 2015. The role of visible and infrared spectroscopy combined with chemometrics to measure phenolic compounds in grape and wine samples. Molecules, 20(1), 726–737.

[10] Strasburg, G. M. and Ludescher, R. D., 1995. Theory and applications of fluorescence spectroscopy in food research. Trends in Food Science and Technology, 6(3), 69-75.

[11] Airado-Rodríguez, D., Durán-Merás, I., Galeano-Díaz, T. and Wold, J., 2011. Front-face fluorescence spectroscopy: A new tool for control in the wine industry. Journal of Food Composition and Analysis, 24(2), 257–264.

[12] Karoui, R. and Blecker, C., 2011. Fluorescence spectroscopy measurement for quality assessment of food systems — a review. Food and Bioprocess Technology, 4(3), 364–386.

[13] Cabrera-Bañegil, M., Hurtado-Sánchez, M., Galeano-Díaz, T.and Durán-Merás, I., 2017. Front-face fluorescence spectroscopy combined with second-order multivariate algorithms for the quantification of polyphenols in red wine samples. Food Chemistry, 220, 168–176.

[14] Cabrera-Bañegil, M., Valdés-Sánchez, E., Moreno, D., Airado-Rodríguez, D. and Durán-Merás, I., 2019. Front-face fluorescence excitation-emission matrices in combination with three-way chemometrics for the discrimination and prediction of phenolic response to vineyard agronomic practices. Food Chemistry, 270, 162–172.

[15] Letort, A., Laguet, A., Lebecque, A. and Serra, J. N., 2006. Investigation of variety, typicality and vintage of French and German wines using front-face fluorescence spectroscopy. Analytica Chimica Acta, 563, 292–299

[16] Parker, C.A., 1968. Apparatus and experimental methods. In: Parker, C.A. (Ed.), Photoluminescence of Solutions with Applications to Photochemistry and Analytical Chemistry, 128–302.

[17] Airado-Rodŕiguez, D., Durán-Merás, I., Galeano-Díaz, T. and Wold, J., 2009. Usefulness of fluorescence excitation-emission matrices in combination with parafac, as fingerprints of red wines. Journal of Agricultural and Food Chemistry, 57(5), 1711–1720.

[18] Gishen, M., Dambergs, R. and Cozzolino, D., 2005. Grape and wine analysis - enhancing the power of spectroscopy with chemometrics. Australian Journal of Grape and Wine Research, 11(3), 296-305.

[19] Andersen, C. M. and Bro, R., 2003. Practical aspects of PARAFAC modeling of fluorescence excitation-emission data. Journal of Chemometrics, 17(4), 200–215.

[20] Giovenzana, V., Beghi, R., Mena, A., Civelli, R., Guidetti, R., Best, S. and Leòn Gutiérrez, L.F., 2013. Quick quality evaluation of Chilean grapes by a portable VIS/NIR device. Acta Horticulturae, 978, 93-100.

[21] Iland, P., Ewart, A., Sitters, J., Markides, A., and Bruer, N., 2000. Techniques for chemical analysis and quality monitoring during winemaking. 1st ed., 1-111. Campbelltown, South Australia: Patrick Iland Wine Promotions.

[22] Sarneckis, C. J., Dambergs, R. G., Jones, P., Mercurio, M., Herderich, M. J. and Smith, P. A., 2006. Quantification of condensed tannins by precipitation with methyl cellulose: development and validation of an optimized tool for grape and wine analysis. Australian Journal of Grape and Wine Research, 12(1), 39−49.

[23] Mercurio, M. D., Dambergs, R. G., Herderich, M. J. and Smith, P. A., 2007. High throughput analysis of red wine and grape phenolics − adaptation and validation of methyl cellulose precipitable tannin assay and modified somers color assay to a rapid 96 well plate format. Journal of Agricultural and Food Chemistry, 55(12), 4651−4657.

[24] Glories, Y., 1984. La couleur des vins rouges, 2eme partie. Connaissance de la Vigne et du Vin, 18, 253−271.

[25] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnav, E., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[26] Bro, R., 1997. PARAFAC. Tutorial and applications. Chemometrics and intelligent laboratory systems, 38(2), 149-172.

[27] Savitzky, A. and Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry, 36(8), 1627-1639.

[28] Chen, T. and Guestrin, C., 2016. XGBoost : A scalable tree boosting system: in proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785–794.

[29] Swersky, K., Snoek, J. and Adams, R.P., 2013. Multi-task bayesian optimization. Advances in neural information processing systems, 26, 2004-2012

[30] Pelikan, M., Goldberg, D.E. and Cantú-Paz, E., 1999. BOA: The Bayesian optimization algorithm. In: Proceedings of the genetic and evolutionary computation conference GECCO-99, 1, 525-532.

[31] Torgo, L., Ribeiro, R.P., Pfahringer, B. and Branco, P., 2013. SMOTE for regression. In: Portuguese conference on artificial intelligence, 378-389.

[32] Hoenicke, K., Simat, T.J., Steinhart, H., Kohler, H.J. and Schwab, A., 2001. Determination of free and conjugated indole-3-acetic acid, tryptophan and tryptophan metabolites in grape must and wine. Journal of Agricultural and Food Chemistry 49, 5494–5501.

[33] Christensen, J., Nørgaard, L., Bro, R. and Engelsen, S.B., 2006. Multivariate autofluorescence of intact food systems. Chemical reviews, 106(6), 1979-1994.

[34] Agati, G., Matteini, P., Oliveira, J., de Freitas, V. and Mateus, N., 2013. Fluorescence approach for measuring anthocyanins and derived pigments in red wine. Journal of Agricultural and Food Chemistry, 61(42), 10156-10162.

[35] Giusti, M. and Wrolstad, R., 2001. Characterization and measurement of anthocyanins by UV-visible spectroscopy. Current protocols in food analytical chemistry, (1), F1-2.

[36] Baluja, J., Diago, M.P., Goovaerts, P. and Tardaguila, J., 2012. Assessment of the spatial variability of anthocyanins in grapes using a fluorescence sensor: relationships with vine vigour and yield. Precision Agriculture, 13(4), 457-472.

[37] Pinelli, P., Romani, A., Fierini, E. and Agati, G., 2018. Prediction models for assessing anthocyanins in grape berries by fluorescence sensors: Dependence on cultivar, site and growing season. Food chemistry, 244, 213-223.

[38] Le Moigne, M., Dufour, E., Bertrand, D., Maury, C., Seraphin, D. and Jourjon, F., 2007. Front face fluorescence spectroscopy and visible spectroscopy coupled with chemometrics have the potential to characterise ripening of Cabernet Franc grapes. Analytica chimica acta, 621(1), 8-18.

[39] Schueuermann, C., Silcock, P. and Bremer, P., 2018. Front-face fluorescence spectroscopy in combination with parallel factor analysis for profiling of clonal and vineyard site differences in commercially produced Pinot Noir grape juices and wines. Journal of Food Composition and Analysis, 66, 30–38

[40] Elith, J., Leathwick, J.R. and Hastie, T., 2008. A working guide to boosted regression trees. Journal of Animal Ecology, 77(4), 802-813.

[41] Brillante, L., Gaiotti, F., Lovat, L., Vincenzi, S., Giacosa, S., Torchio, F., Segade, S.R., Rolle, L. and Tomasi, D., 2015. Investigating the use of gradient boosting machine, random forest and their ensemble to predict skin flavonoid content from berry physical–mechanical characteristics in wine grapes. Computers and Electronics in Agriculture, 117, 186-193.

[42] Nielsen, D., 2016. Tree boosting with xgboost - why does xgboost win" every" machine learning competition? Master's thesis, Norwegian University of Science and Technology.

[43] Coelho, C., Aron, A., Roullier-Gall, C., Gonsior, M., Schmitt-Kopplin, P. and Gougeon, R., 2015. Fluorescence fingerprinting of bottled white wines can reveal memories related to sulfur dioxide treatments of the must. Analytical chemistry, 87(16), 8132–8137.

[44] Goldberger, J., Hinton, G.E., Roweis, S. and Salakhutdinov, R.R., 2004. Neighbourhood components analysis. Advances in neural information processing systems, 17, 513-520.

[45] Rossouw, M., 2003. The Phenolic Composition of South African Pinotage, Shiraz and Cabernet Sauvignon Wines. South African Journal of Enology and Viticulture, 25(2), 94-104.

[46] Fragoso, S., Aceña, L., Guasch, J., Mestres, M. and Busto, O., 2011. Quantification of phenolic compounds during red winemaking using FT-MIR spectroscopy and PLS-regression. Journal of Agricultural and Food Chemistry, 59, 10795–10802.

[47] Aleixandre-Tudo, J.L., Nieuwoudt, H., Aleixandre, J.L. and du Toit, W., 2018. Chemometric compositional analysis of phenolic compounds in fermenting samples and wines using different infrared spectroscopy techniques. Talanta, *176*, 526-536.

[48] Dambergs, R.G., Mercurio, MD., Kassara, S., Cozzolino, D. and Smith, P.A., 2012. Rapid measurement of methyl cellulose precipitable tannins using ultraviolet spectroscopy with chemometrics: Application to red wine and inter-laboratory calibration transfer. Applied Spectroscopy, *66*(6), 656-664.

[49] Beaver, C.W. and Harbertson, J.F., 2016. Comparison of multivariate regression methods for the analysis of phenolics in wine made from two vitis vinifera cultivars. American Journal of Enology and Viticulture, *67*(1), 56-64.