Computers and Electronics in Agriculture 212 (2023) 108098

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

In situ grape ripeness estimation via hyperspectral imaging and deep autoencoders☆

Nikolaos L. Tsakiridis [a,*], Nikiforos Samarinas [a], Stylianos Kokkas [a], Eleni Kalopesa [a], Nikolaos V. Tziolas [c], George C. Zalidis [b]

[a] SpectraLab group, Laboratory of Remote Sensing, Spectroscopy, and GIS, School of Agriculture, Aristotle University of Thessaloniki, Thermi, 57001, Greece
[b] Interbalkan Environment Center, 18 Loutron Str., Lagadas, 57200, Greece
[c] Southwest Florida Research and Education Center, Department of Soil and Water Sciences, Institute of Food and Ecosystem Sciences, University of Florida, 2685 State Rd 29N, Immokalee, 34142, FL, United States of America

ARTICLE INFO

Keywords:
Dissolved solids
VNIR
Vis–NIR
Grape cultivars
Deep learning

ABSTRACT

The estimation of the grapes' maturity in the field using non-destructive techniques is of high interest for the high-valued vinified grapes, particularly towards the development of fully automated agrobots that perform selective harvesting operations. Whereas infrared spectroscopy has been employed using point spectrometers in the laboratory and in the field, imaging spectrometers have mainly been tested in controlled laboratory conditions due to issues with varying illumination. In this paper, the application of the autoencoder framework is proposed, which is employed to transform the raw recorded spectra, regardless of illumination conditions, into standardized reflectance spectra; thus addressing the inherent difficulties which hamper the direct application of hyperspectral imaging in the field. To validate the methodology, the sugar content (°Brix) of four grape varieties, namely Chardonnay, Malagouzia, Sauvignon-Blanc, and Syrah, is estimated. Two different autoencoder architectures are examined: deep fully-connected (DAE) and deep convolutional autoencoders (DCAE), while the estimation of sugar content takes place using as input both from the encoded (latent) space and from the autoencoders' output, i.e., the transformed standardized spectra. The use of multiple spectral pre-treatments is further examined to enhance the accuracy of prediction. Despite that DAE and DCAE showcase comparable similarity metrics, DCAE statistically outperforms DAE when using both the encoded space and the autoencoders' output, attesting to the suitability of the convolutional autoencoder framework. On the other hand, there is no statistical significant difference when employing multiple input pre-treatments. The accuracy of estimation (mean RMSE 1.83 °Brix, R^2 0.70, RPIQ 2.43) is comparable to other studies that directly work with standardized reflectance spectra in laboratory conditions.

1. Introduction

The origins of wine production predate recorded history (McGovern, 2013) and throughout antiquity wine was considered a gift from the gods with the elite of the society reserving the best wines for themselves (Bisson et al., 2002). The global market for wine was estimated at 345.9 Billion USD for 2020, and it is projected to reach a size of 456.1 Billion USD by 2027 (Global Industry Analysts, 2022). In 2021 grapes were the fifth most widely produced fruit worldwide with approximately 79.5 million tonnes (FAO, 2021). In 2022, the world wine production (excluding juices and musts) is estimated at 259.9 million hectoliters, slightly below its 20-year average but relatively stable for four consecutive years (International Organisation of Vine and Wine, 2022b). The vinified production in the European Union in 2021 was 153.7 million hectoliters (approximately 59% of the world production) from an estimated 22.7 million tonnes of total harvested grapes for wine, with Italy, Spain and France together accounting for 47% of the world wine production in 2021 (International Organisation of Vine and Wine, 2022a; Eurostat, 2022). Evidently, the production of grapes for wine has a considerable worldwide economic impact and is one of the highest valued crops.

☆ The research leading to these results has received funding from the European Community's Framework Programme Horizon 2020 under grant agreement No 871704, project BACCHUS. The authors would also like to thank Ktima Gerovassiliou for providing the grape samples. The dataset and implementation of this work are available upon request to the corresponding author.

* Corresponding author.
E-mail addresses: tsakirin@ece.auth.gr (N.L. Tsakiridis), smnikiforos@topo.auth.gr (N. Samarinas), kokkassk@auth.gr (S. Kokkas), kalopesa@agro.auth.gr (E. Kalopesa), ntziolas@ufl.edu (N.V. Tziolas), zalidis@agro.auth.gr (G.C. Zalidis).

https://doi.org/10.1016/j.compag.2023.108098
Received 26 April 2023; Received in revised form 27 June 2023; Accepted 20 July 2023
Available online 7 August 2023

Considering that grapes are non-climacteric fruits (i.e., they do not ripen any further if harvested) (Prasanna et al., 2007), their maturity degree is a decisive factor that determines wine chemical composition and sensory traits (Niimi et al., 2017). Monitoring of grape quality is traditionally carried out directly in the field using destructive techniques, in order to ascertain the appropriate harvest time. This evaluation takes place offline via classical physical and chemical methods, using a limited number of samples selected by the experts. This process is however time-consuming, laborious, costly, invasive and oftentimes subjective. Taking into account that the trend in modern agricultural practices involves the use of robots and automated solutions throughout the production process, including harvesting (Power et al., 2019; Fountas et al., 2020), it is important to design an automated solution that can determine the maturity level of grapes for appropriate decision making.

Infrared spectroscopy has been employed to estimate phenolic composition, quality indicators and authenticity in grapes and wines demonstrating its efficacy as a useful tool to replace the traditional approaches (Ferrer-Gallego et al., 2022). A review by Power et al. (2019) studied the evolution on the measurement of grape composition from the laboratory to the vineyard via new technologies (i.e., fiber optics, LED, hyperspectral imaging), highlighting the cost and time saving by using spectroscopy techniques in the field. However, despite their desirable characteristics, hyperspectral images in the near infrared obtained from either unmanned aerial vehicles flying on lower altitudes or on ground hyperspectral are more difficult to process because it is more complex to compute the reflectance at each pixel. The reason is that the illumination of the acquired hyperspectral data (even within a single image) varies depending on the angle of the surface depicted (Zhang et al., 2022), due to light scattering, shadowing of plant components and occlusions, and a complicated interaction between scattering and shadowing (Mishra et al., 2020). The sort of plants and their intricate geometry determine how much of an impact these factors have. Moreover, when obtaining multiple images, due to fluctuating cloud cover density it is possible that there are considerable changes of illumination within a narrow time frame. Illumination compensation has been attempted using custom ground vehicles that cover the canopy and use artificial light sources (Wendel and Underwood, 2017) or by custom set-ups that block completely the ambient light (Polder et al., 2019).

Various approaches in the literature employing hyperspectral cameras have focused on their application only in highly controlled laboratory conditions and mostly using line-scanning technologies which dictate the need for a linear translation platform (Baiano et al., 2012; Chen et al., 2015; Gabrielli et al., 2021; Xu et al., 2022). Interestingly, in some works (Silva et al., 2018; Gomes et al., 2021) the authors opted to use a single line scan (i.e., not a hyperspectral cube) to extract the reflectance of the grapes in the laboratory (Benelli et al., 2021) used a hyperspectral camera in situ to estimate the grape ripeness in one cultivar, but the camera employed line-scanning to create the respective hyperspectral cubes which dictates the need for a custom cart to mount the camera and the data may be influenced by ambient factors (e.g., wind, bumps in the field, etc.). Moreover, they employed classifications algorithms to classify the grapes as ripe or non-ripe, without quantifying their °Brix content. Line-scanning imaging for the classification of grape varieties was also utilized by Gutiérrez et al. (2018). Rodríguez-Pulido et al. (2022) used a hyperspectral sensor (400–1000 nm) with push-broom principle to predict chemical compounds (sugar content, phenols, anthocyanins) in whole wine grape bunches of red varieties. They applied mathematical prediction models on laboratory conditions and transferred them in field environment for quantitative chemical analysis, without however accounting for the difficult ambient conditions (e.g., shadows). Furthermore, it should be also noted that a review by Silva and Melo-Pinto (2021) examined different dimensionality reduction methods for prediction of sugar content from hyperspectral images of wine grape berries, concluding that Principal Component Analysis (PCA) prevails over more elaborate methods.

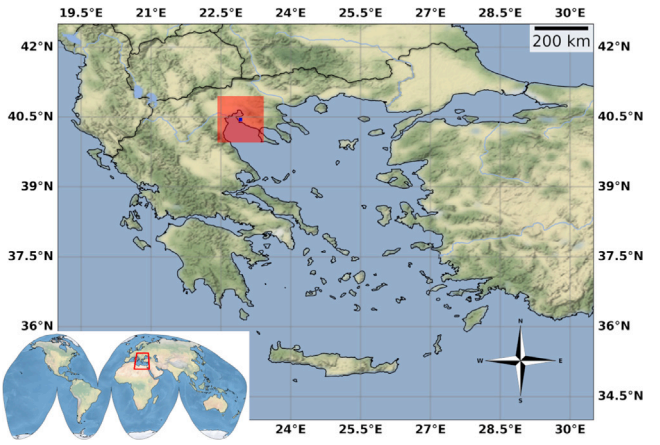


Fig. 1. Location of the Gerovassiliou estate (blue marker in red box) in Northern Greece.

Autoencoders (Bank et al., 2020) are a class of self-supervised neural networks, and, in their simplest form, are a lossy algorithm mapping an input to compressed representation (called latent space) using an encoder, and then back to itself via a decoder. It can be considered as a generalization of Principal Component Analysis, where instead of constructing a low dimensional hyperplane for the compressed feature space, it is able to learn a non-linear manifold (Hinton and Salakhutdinov, 2006). Denoising autoencoders (Vincent et al., 2008) are a robust framework that is used for error correction, where the input is disrupted by a noise with the autoencoder expected to reconstruct the original (undisrupted) input. Convolutional autoencoders (Masci et al., 2011) extend the autoencoder framework by employing the operation of convolution; the encoder uses convolution and max pooling layers, whereas the decoder employed deconvolution and upscaling layers. Autoencoders have various applications; for example, they have been used for cyber security intrusion detection systems (Berman et al., 2019; Ferrag et al., 2020), metagenome binning and assembly (Nissen et al., 2021), and for epileptic seizures detection in electroencephalography signals (Shoeibi et al., 2021). Variations of the autoencoder framework have also been applied in hyperspectral imagery (Signoroni et al., 2019), in cases such as spectral unmixing (Su et al., 2019; Palsson et al., 2021), anomaly detection (Zhang and Cheng, 2019; Lv et al., 2023), dimensionality reduction (Nalepa et al., 2020; Mantripragada et al., 2022), feature selection (Wang et al., 2021) and land cover and crop type classifications (Kussul et al., 2017). Deep autoencoders have also been employed in infrared spectroscopy, both in point spectroscopy and hyperspectral imaging, for performing feature extraction to enhance the final model estimations (Liu et al., 2017; Zhang et al., 2020, 2021; Liu et al., 2022).

In this paper, we examine the application of deep learning techniques on in situ hyperspectral VNIR data for the estimation of the grapes' sugar content. The main novelty of this approach lies in the usage of the autoencoder framework to address the impact of the in situ illumination conditions on the spectral signal. Thus, this paper focuses on the application of infrared hyperspectral imaging in real field conditions and hence examining the potential of mounting such devices on robotic platforms for real-time harvesting.

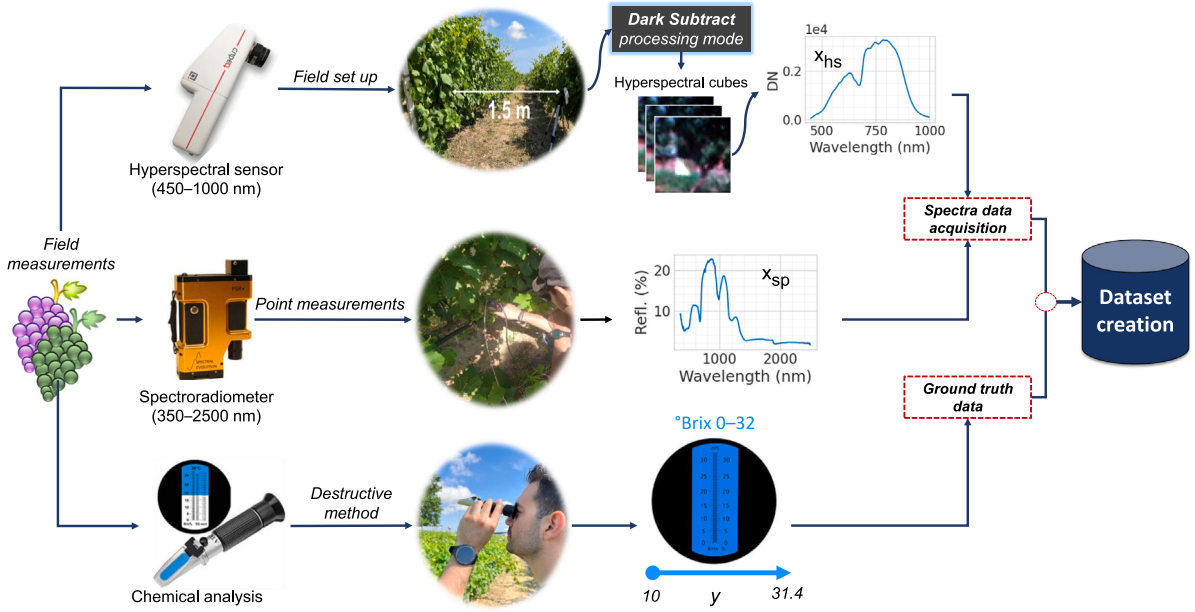


Fig. 2. A flowchart illustrating the data collection process for a single sample of the Sauvignon-Blanc variety.

2. Materials and methods

2.1. Study area

The field data were collected in Ktima Gerovassiliou, which is situated in the Epanomi region, located 25 km south-east of Thessaloniki within the administrative borders of the Municipality of Thermaikos and the Prefecture of Thessaloniki in Northern Greece (Fig. 1). Epanomi is a peninsula suitable for viticulture and has been used for grape cultivation for many decades. The climate of the region is classified as Csb under the Köppen climate classification system. The estate itself covers an area of around 72 ha and many different grape varieties are cultivated there. We selected four different grape varieties, namely Chardonnay, Malagouzia, Sauvignon-Blanc, and Syrah. Malagouzia (also spelled as Malagousia) is an aromatic white variety grown primarily in Central Greece and Greek Macedonia, best known for its citrus and peach characteristics. It was selected as a representative of local grape varieties. Chardonnay and Sauvignon-Blanc are also white varieties, while Syrah is a red variety. These three are amongst the top 10 most widely cultivated varieties throughout the world (International Organisation of Vine and Wine, 2017).

2.2. Data collection

The process of data collection is summarized in Fig. 2. Field visits took place in the cultivating period (i.e., July–August) of the 2021 vintage. All data (including spectral measurements) was recorded in-situ with real-world environmental conditions. In order to construct a robust dataset encapsulating the within-field variability that may be found in situ, it is important not only to include grapes spanning the entire maturity period (from veraison to harvesting), but also from different areas of the vineyard, corresponding to different conditions. As far as the monitored maturity period is concerned, according to the BBCH scale (Lorenz et al., 1995) describing the phenological growth stages of the grapevine, we monitored stages 81 (beginning of ripening where berries begin to brighten in color) through 89 (berries ripe for harvest). For each of the examined grape variety, we selected at least two different regions in the vineyard and, for each region, we took care to monitor vine trees showcasing different vigour.

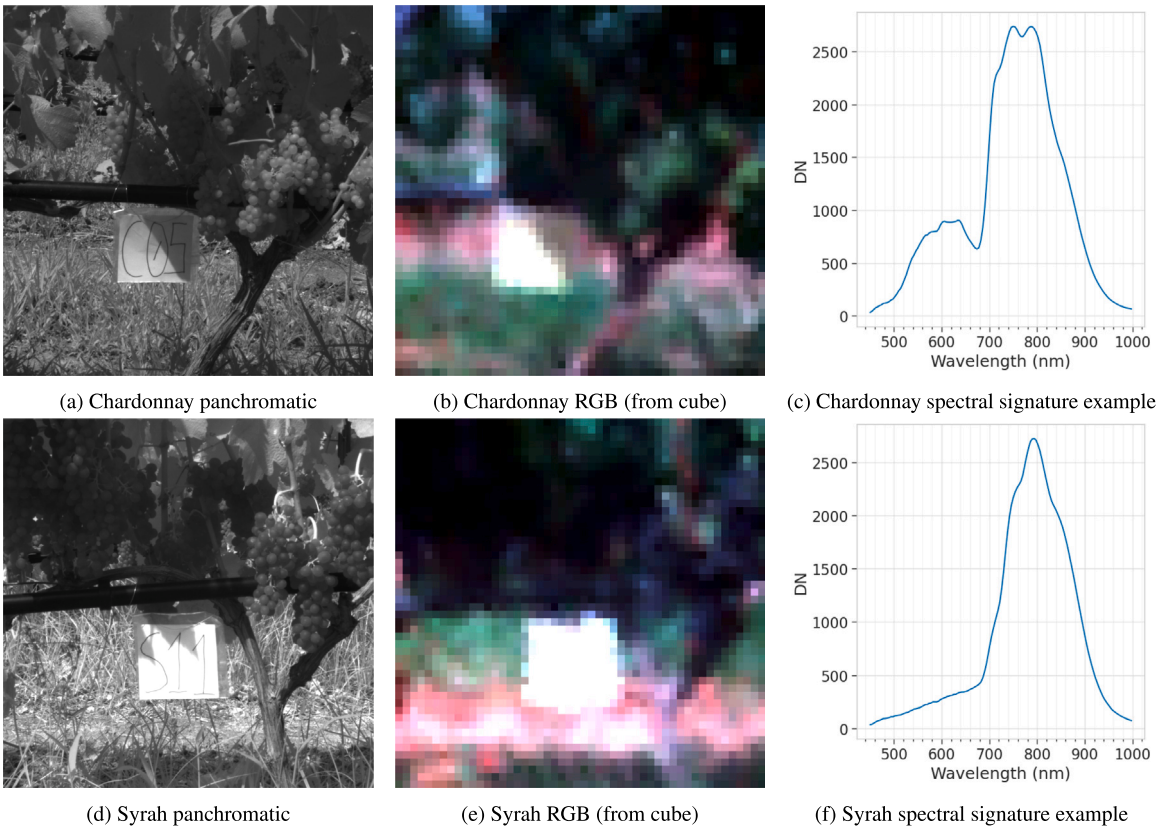
To ensure that on each successive field visit the same vine trees were recorded, as well as future reference through photos could be readibly established, four rows corresponding to the four cultivars were selected and labels were placed accordingly on the trees. For each grape variety we selected 20 trees, with the exception of the local grape cultivar of Malagouzia where 30 trees were selected (i.e., a total of 90 trees). On each field visit, the following was recorded from each of the selected trees:

1. First, at least one hyperspectral VNIR image was recorded using a snapshot hyperspectral camera capturing the bunches of grapes; the camera was mounted on a tripod and directed manually towards the grapes.
2. Then, a single berry was randomly selected from one of the bunches that was recorded with the camera, and its point VNIR–SWIR diffuse reflectance with a contact probe was recorded;
3. Finally, the sugar content from the selected berry was determined by crushing it, collecting its juice, and using a portable refractometer.

In the following subsections, we analyze each of the aforementioned steps.

2.2.1. Recording of hyperspectral VNIR datacubes

Hyperspectral cubes were recorded in situ via the Cubert FireflEYE V185 snapshot hyperspectral sensor, which was mounted on a tripod. The sensor works in 450–1000 nm with a spectral sampling of 4 nm and full width at half maximum of 8 nm at 532 nm. Thus, it captures 3D cubes with a resolution of 50 × 50 pixels in the 2D spatial dimensions and 128 channels in the spectral dimension. Based on the sensor's general specifications (i.e., field of view, lens angle etc.) the optimal distance from the grapes was calculated to 1.5 m. Undoubtedly, in situ measurements using hyperspectral cameras involve several obstacles, mainly in terms of sensor calibration, due to climate condition variations and lighting sensitivity which are addressed in this paper with the novel methodology of autoencoders. The data acquisition took place between the morning and noon (i.e., 09:00 to 14:00 local time) and the camera's exposure time was set manually between 12 and 20 ms depending on the ambient conditions and as determined through test images (e.g., no overexposure of any pixels). Measurements taken closer to noon are subject to higher ambient temperatures and the accumulated heating effect from recording data at prolonged times; we thus took efforts to keep the equipment in the shade at all times, and



(a) Chardonnay panchromatic (b) Chardonnay RGB (from cube) (c) Chardonnay spectral signature example

(d) Syrah panchromatic (e) Syrah RGB (from cube) (f) Syrah spectral signature example

Fig. 3. Visualization of the hyperspectral cubes recorded for the Chardonnay and Syrah varieties; (a) and (d) depict the panchromatic (grayscale) 1000 × 1000 pixel image corresponding to the hyperspectral cubes of (b) and (e) respectively, whereas (c) and (f) illustrate some example signatures extracted from the hyperspectral cubes.

use both passive and active cooling on all instruments to improve heat dissipation and cooling. In this context, the use of the dark subtract processing mode was nevertheless utilized to remove the dark current noise. This mode is activated after calibrating the sensor in the black area by placing a completely black and high-quality material (provided by the manufacturer) on the lens and by closing the camera's shutter. The dark measurements were repeated after 10 measurements to ensure the stability of the recordings. The output of this mode is a spectral signature expressed as a digital number (DN) as shown in Figs. 3(c) and 3(f).

2.2.2. Using the spectrometer to record reference VNIR–SWIR spectra

The PSR+3500 (Spectral Evolution Inc., Lawrence, MA, USA) spectrometer was also used to record in situ the reference VNIR–SWIR spectrum (350–2500 nm) of the grapes. It is a portable high accurate contact probe spectrometer which came into contact with the berries in the bunches to record their spectrum. Direct reflectance mode was utilized after the necessary calibration process by using a white reference plate made of Spectralon® material. Every 10 successful measurements the calibration procedure was repeated in order to ensure the robustness and accuracy of the sensor. It should be noted that despite recording the entire 350–2500 nm range, only the 450–1000 nm range was retained to match the spectral range of the hyperspectral camera.

Because two separate spectral measurements were conducted, one for the VNIR datacubes and one for the reference VNIR–SWIR spectra, it is important to perform the matching between the two. A careful note was made of the berry whose reflectance was measured, to allow it to be matched with its corresponding pixel in the hyperspectral cube. This process was implemented manually by an expert who performed the matching operation using open source geospatial tools.

2.2.3. Estimation of sugar content

The grape sugar content estimation was performed immediately after the hyperspectral camera and the spectroradiometer data acquisition. A portable refractometer (RHB-32ATC- Laxco Inc., Bothell, WD, USA) was used to record the soluble solid content expressed in degrees Brix (°Brix) which is a measurement of the relative density of dissolved sucrose in unfermented grape juice, in grams per 100 milliliters. The refractometer has a range of 0–32 °Brix, an accuracy of 0.20 °Brix and a resolution of ±0.20°Brix. The berry of each bunch that was previously used for the point spectra information with the contact probe spectroradiometer, was selected for the °Brix measurement with a destructive effect on the fruit. The berry was carefully cut and squeezed to use its juicy for analysis while specific measurement protocols were followed such as cleaning the prism of the refractometer with deionized water after each measurement to ensure the quality and their reliability of the measurement for each berry.

2.3. Spectral pre-treatments

Spectral data can be complex, noisy, and difficult to interpret, and pre-treatment methods are applied to enhance the quality of the data and to remove sources of variability that can impede the effectiveness of chemometric analysis. Such sources are baseline offsets, noise, and various interferences. Pre-treatments can also help to highlight features of interest in the spectra and to remove irrelevant information that may obscure the underlying signal.

Standard normal variate (SNV) is a widely employed spectral preprocessing method in chemometrics, aiming to correct the baseline offset and to reduce the scattering in the spectra. It involves a transformation of the spectra to create a new spectrum of zero mean and unit variance, i.e., standardized to a standard normal distribution. The transformation is performed by subtracting the mean value of each

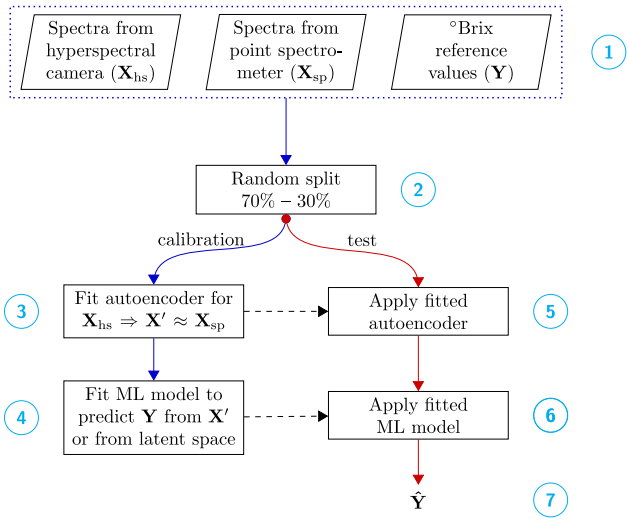


Fig. 4. Overall approach undertaken for each grape cultivar.

spectrum and then dividing the result by its standard deviation It is particularly useful when analyzing spectra from different sources or instruments, as it can help to remove systematic differences in the data that may arise from varying experimental conditions or instrumental drift.

First and second derivatives using Savitzky–Golay filters are another widely used spectral pre-processing technique in chemometrics. The Savitzky–Golay filter is a smoothing algorithm that is applied to a moving window of data points in the spectrum. The algorithm calculates the slope or curvature of the data within the window, and then replaces the central data point with the calculated value; the result is a smoothed, differentiated spectrum. By taking the first or second derivative of the spectrum, it is possible to highlight subtle changes in the spectral signal that may be difficult to discern in the original data. This can be particularly useful in applications where the goal is to identify and quantify spectral features, such as in the analysis of complex mixtures of compounds.

2.4. Analysis using artificial intelligence

The subsections that follow entail the analysis steps taken to estimate the sugar content from the raw hyperspectral images. It is important to note that the analysis was conducted separately for each grape cultivar. The overall approach is presented in Fig. 4.

2.4.1. Dataset split

To train the autoencoder and the machine learning models, the dataset was randomly split into calibration (70%) and tests set (30%), as indicated in step 2 of Fig. 4. Following this, the calibration set was further divided into a training (70%) and a validation set (30%) used to train the autoencoders (step 3). The first is used for parameter fitting (i.e., model weights) while the latter to avoid overfitting. After transforming $\mathbf{X}_{hs}$ the machine learning models were fitted. In this case, the calibration set was randomly divided into 5-folds which were utilized for hyperparameter tuning of the learning algorithms, where the optimal hyperparameter set is selected as the one providing the lowest prediction error in the held-out sets. The final model per each learning algorithm was established using the whole calibration set and the optimal hyperparameter set.

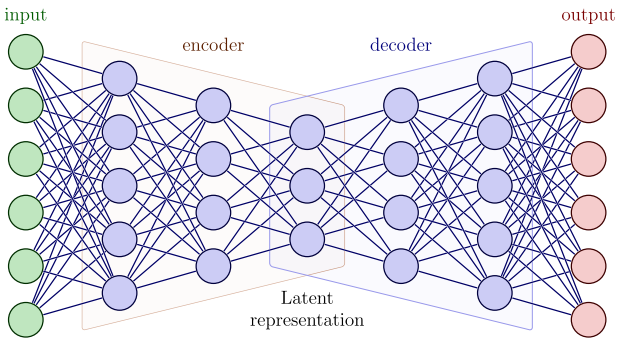


Fig. 5. The general autoencoder framework; the input data are compressed using an encoder to a latent representation, which is accordingly decompressed via the decoder to reconstruct the output.

2.4.2. Autoencoders

Fig. 5 presents the simple autoencoder framework. An autoencoder consists of two main parts: an encoder, which maps the input data to a lower-dimensional representation (also known as the encoding or latent representation), and a decoder, which maps the encoding back to the original input data. In its simple form, as indicated in the figure, it is comprised of fully connected (or dense) layers. In the training phase, the reconstruction error between the original input and the output is minimized. This forces the network to learn a meaningful representation of the input data in the encoding space, while preserving the important features and structure in the data.

Denoising autoencoders are a variant mainly used to remove noise from an input signal. The network is trained by corrupting the input data using random noise and the output is expected to match the noise-free input. The goal is to encourage the network to learn a robust representation of the input that is robust to the presence of noise and can effectively "denoise" the input. They are also used as a stochastic version of the simple autoencoder in cases where there are more hidden layers than inputs, because then there is a risk that the learning algorithm will only learn the identity function. The difference in the training process is that because the input is the noisy version of the original data, the reconstruction error is computed between the clean (noise-free) data and the output of the network, with the network parameters updated accordingly to minimize this error.

Convolutional autoencoders incorporate the convolutional layers and their associate layers, like pooling, transpose convolution, and batch normalization into the encoder–decoder architecture. The convolutional layers provide many advantages over the typical fully-connected layers, including position invariance and fewer parameters. Pooling layers may also be replaced by convolutions with stride, providing comparable or better results without loss of information (Springenberg et al., 2014).

2.4.3. From raw signal to reflectance using autoencoders

To address the issue of different illumination conditions on each captured scene, this study proposes the use of autoencoders (step 3 of Fig. 4). This transformation framework is presented in Fig. 6. Various architectures were considered using both fully connected and convolutional layers, with the aim to optimize the aforementioned evaluation metrics. The following autoencoders architectures are considered in this paper:

1. Deep denoising autoencoder, using fully connected layers, with the input being the SNV transformed raw signatures from the hyperspectral camera and the output the corresponding reflectance signature from the point spectrometer. This network consists of three hidden layers in both the encoder and the decoder. The first layer in the encoder has 240 neurons and a linear

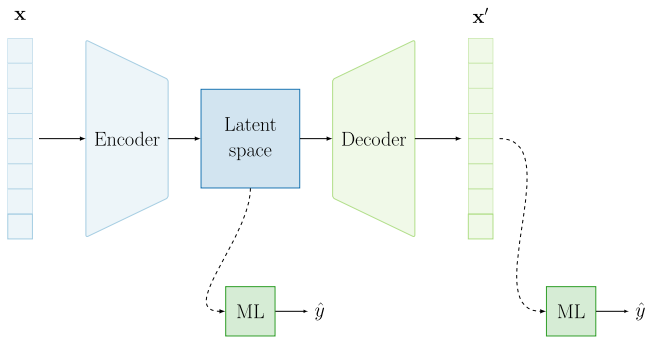


Fig. 6. The proposed de-noising framework; the input spectrum **x** collected from the hyperspectral camera in raw mode is transformed to the output reflectance spectrum **x**′ via the autoencoder. The autoencoder is trained using as reference output data the reflectance recorded from the handheld spectrometer. The machine learning (ML) models to predict the sugar content ($\hat{y}$) may be accordingly developed either from the latent representations or from the de-noised output **x**′.

activation, the second layer 120 neurons with a tanh activation, while the third layer employs 60 neurons and a tanh activation function. The latent representation consists of 30 neurons and uses a linear activation function. Finally, the decoder uses the same architecture as the encoder in reverse. The architecture is presented in Table 1. This network employs 138,038 trainable parameters.

2. Deep multi-input denoising autoencoder, where instead of the raw input, the following pre-treatments are further considered, namely the SNV transformation, the first-derivative and SNV, and the second-derivative and SNV. All these spectral sources are concatenated to form a single continuous vector. The same architecture as above is used, with the distinction that the input and output layers have three times the length of the respective layers used above. This network employs 261,174 trainable parameters.
3. Deep convolutional denoising autoencoder, using convolutional layers and as input the SNV transformed raw signatures from the hyperspectral camera. The encoder consists of four layers, using successively 256, 128, 128, and 64 filters with a kernel size that is progressively reduced from 21 to 5, while the deconvolution process uses the reverse process, as indicated in Table 2. By employing a stride step the input is successively halved and reaches from a length of 128 to 32 in the latent space. The total number of trainable parameters is 1,712,514.
4. Deep convolutional multi-input denoising autoencoder, using the aforementioned pre-treatments as multiple input channels, and the network architecture of Table 2. The number of trainable parameters is increased due to the use of the multi-channel input to 1,725,828.

It is worth noting that the autoencoders use a mix of linear and tanh activation functions because it allows the network to learn a wider range of functions. The advantage is that linear activations in both input and output enable the model to extrapolate to ranges outside those defined in the input, while the tanh activations are useful in the hidden layers of the encoder and decoder because they help normalize the data and prevent the network from becoming too sensitive to small changes in the input. Regarding the specific configuration of neurons for the encoders and decoders, we chose a progressive reduction and expansion scheme. In the proposed architectures, the encoder progressively reduces the dimensionality of the input data through layers with decreasing neuron counts, ultimately reaching the desired latent space of 30 neurons for the fully connected layers and 32 neurons for the convolutional encoder. Similarly, the decoders mirror this progression in reverse order to reconstruct the original input, allowing the model

Table 1
The network architecture of the DAE using a single pre-treatment as input; for the multi-input case (i.e., when 3 spectral pre-treatments are employed), the input and output neurons are 128 × 3 = 384.

Layer	Neurons	Activation
Input	128	
Encoder 1	240	linear
Encoder 2	120	tanh
Encoder 3	60	tanh
Latent	30	linear
Decoder 1	60	linear
Encoder 2	120	tanh
Decoder 3	240	tanh
Output	128	linear

Table 2
The network architecture of the DCAE, where P denotes the number of pre-treatments used; for the multi-input case (i.e., when 3 spectral pre-treatments are employed), the input and output neurons are 128 × 3 = 384.

Layer	Filters	Kernel size	Strides	Activation
Input	Neurons: 128, channels: P			
Encoder 1	256	21	1	linear
Encoder 2	128	15	2	linear
Encoder 3	128	9	2	tanh
Encoder 4	64	7	2	tanh
Latent	1	5	1	linear
Decoder 1	64	7	2	linear
Decoder 2	128	9	2	tanh
Decoder 3	128	15	2	tanh
Decoder 4	256	21	1	linear
Output	P	3	1	linear

to capture hierarchical and abstract representations in the latent space while enabling effective reconstruction of the input data. Moreover, the choice of kernel size for the convolutions was chosen with the motivation of enabling the first layers of the models to learn higher-level features (larger kernel size) while the deeper layers to identify more detailed features (smaller kernel size). The autoencoders using fully connected layers are henceforth denoted as DAE (Deep AutoEncoders), while those employing the convolution operations as DCAE (Deep Convolutional AutoEncoders).

The models were trained for 300 epochs using the AMSGrad (Reddi et al., 2019) stochastic optimization method that optimized a custom loss function. AMSGrad seeks to fix a convergence issue with the popular Adam-based optimizers, using a running maximum of the squared gradients instead of an exponential moving average to update the parameters. While at first a natural choice for the loss functions is the mean squared error between $\mathbf{X}'$ and $\mathbf{X}_{sp}$ nevertheless this does not guarantee that the spectral features are retained. For example, if a small sinusoidal noise is applied on the target reference spectrum then the mean squared error between the noisy values and the original values will be low, however numerous new spectral absorption bands are introduced which hinders the ability of ML models to predict correctly the soil properties and assign importance to specific bands. For this reason, the custom loss function is defined as the function that simultaneously minimizes the mean squared error and maximizes the cosine similarity between the target spectrum **x** and the transformed spectrum **x**′ each comprised of M spectral bands:

$$\mathcal{L}(\mathbf{x},\mathbf{x}') = \sum_{i=1}^{M}(x_i - x_i')^2 \cdot \left(1 - \frac{\mathbf{x}\cdot\mathbf{x}'}{|\mathbf{x}\parallel\mathbf{x}'|}\right) \quad (1)$$

2.4.4. Machine learning algorithms for regression

As indicated in Figs. 4 and 6, to predict the sugar content, two approaches were tested for each of the autoencoder architectures considered: (1) using the latent space as predictors, and (2) using the denoised spectra as predictors. An aspect that should be clarified in the second case listed above, is that when multiple pre-treatments are used

in the input signal, the machine learning algorithms examine an equal number of scenarios to the number of pre-treatments considering that the denoised spectra are also in the form of the same pre-treatments. The learning algorithms detailed below were applied in each case:

1. XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library (Chen and Guestrin, 2016). The algorithm can be used for both regression and classification tasks and has been designed to work with large and complicated datasets. Whereas GBDTs iteratively train an ensemble of shallow decision trees, with XGBoost trees are built in parallel following a level-wise strategy, scanning across gradient values and using these partial sums to evaluate the quality of splits at every possible split in the training set. XGBoost has demonstrated its efficacy across a wide variety of applications, with a key to a success being the proper tuning of its various hyperparameters. In this work, we tuned the following hyperparameters in their respective search spaces: (i) the maximum depth of a tree, $\{3, 4, \ldots, 8\}$, (ii) η or the learning rate, $\{0.02, 0.05, 0.1, 0.2, 0.3\}$, (iii) the number of trees, $\{10, 50, 100, 200, 500\}$, (iv) the subsample ratio of columns when constructing each tree, $\{0.3, 0.5, 0.8\}$, (v) the L1 regularization term on weights, $\{0, 5, 10\}$.
2. The method of Random Forests, an important ensemble learning method, widely used in both classification and nonparametric regression (Breiman, 2001). It builds multiple decision trees and combines their predictions to produce a more accurate and stable prediction. The algorithm works by randomly selecting a subset of features and training a decision tree on those features. This process is repeated multiple times to create a forest of trees, each with a different subset of features. The final prediction is made by averaging the predictions of all the trees in the forest. The algorithm aims to reduce overfitting and improve the generalization ability of the model. In terms of its hyperparameter optimization process, the number of trees were selected from within $\{50, 100, 150, 200\}$ while the number of features to consider when looking for the best split was searched in $\{M, \sqrt{M}, log_2(M)\}$ with M being the total number of features.
3. Cubist, an algorithm for regression problems that combines decision trees with linear models (Kuhn and Johnson, 2013). The algorithm starts by building a decision tree, and then identifies the leaf nodes that are most predictive of the target variable. A linear model is then fit to each of these leaf nodes, where the predictors are the features and the response is the target variable. The final prediction for a given input is made by weighting the predictions from each of the linear models, with the weights determined by the path from the root node of the decision tree to the corresponding leaf node. The algorithm aims to achieve the stability of linear models with the non-linear flexibility of decision trees. Cubist further utilizes: (i) a boosting-like mechanism termed committees which generate iteratively more rule-based models trying to compensate for the prediction errors of the previous iterations, and (ii) composite models which repair the errors of the rule-based models through instance-based learning. Its hyperparameters include the number of committees, whose search space was $\{1, 5, 10, 20\}$ and the number of neighbors used to repair the errors of prediction searched within $\{0, 1, 5, 9\}$, where 0 indicates that no error-correction was employed.
4. Support Vector Regression (SVR), a type of regression analysis in which the prediction is done through a linear equation in a high-dimensional feature space, called kernel space, in which the data are mapped (Cortes and Vapnik, 1995; Awad and Khanna, 2015). SVR creates a flexible tube with a small radius symmetrically surrounding the estimated function, such that the absolute values of errors less than a certain threshold are ignored both above and below the estimate. In this way, points above or below the function that are inside the tube are not penalized, but points outside the tube are. A radial basis kernel function was used, while its two hyperparameters that were optimized are ϵ from $\{0.025, 0.05, \ldots, 0.5\}$ and C from $\{2^{-3}, 2^{-2}, \ldots, 2^{9}\}$.

2.4.5. Evaluation metrics

To evaluate the efficacy of the proposed autoencoders to transform the raw hyperspectral measurements to the calibrated reference reflectance spectra, three distance metrics were employed to assess the difference between the reference spectra and the autoencoders' output:

1. Pearson's correlation coefficient (ρ), a measure of the linear association between the two, whose values range between −1 and 1 and indicate a strong negative and a strong positive linear relationship, respectively;
2. Lin's concordance correlation coefficient (ρ_c) (Lin, 1989), a scalar value between −1 and 1 that indicates the strength and direction of the relationship which unlike ρ takes into account both the linear and non-linear relationships; and
3. The distance according to spectral angle mapper (SAM) (Kruse et al., 1993), based on the idea of finding the angle between the two spectra in feature space; the smaller the angle, the greater the similarity between them. This is equivalent to the cosine similarity.

These three metrics are calculated as follows to quantify the distance between the target spectrum $\mathbf{x}$ and the transformed spectrum $\mathbf{x}'$ each comprised of M spectral bands:

$$\rho(\mathbf{x}, \mathbf{x}') = \frac{\sum_{i=1}^{M}(x_i - \overline{x})(x_i' - \overline{x}')}{\sqrt{\sum_{i=1}^{M}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{M}(x_i' - \overline{x}')^2}} \tag{2}$$

$$\rho_c(\mathbf{x}, \mathbf{x}') = \frac{\frac{2}{M}\sum_{i=1}^{M}(x_i - \overline{x})(x_i' - \overline{x}')}{S_x^2 + S_{x'}^2 + (\overline{x} - \overline{x}')^2} \tag{3}$$

$$\mathrm{SAM}(\mathbf{x}, \mathbf{x}') = \arccos \frac{\sum_{i=1}^{M} x_i x_i'}{\sqrt{\sum_{i=1}^{M} x_i^2}\sqrt{\sum_{i=1}^{M} x_i'^2}} \tag{4}$$

where $\overline{x}, S_x^2$ and $\overline{x}', S_{x'}^2$ are the mean values and variances of the reference and transformed spectra, respectively.

On the other hand, the following metrics were utilized to validate the machine learning models that predict the °Brix content on the independent test set:

1. the root mean squared error (RMSE);
2. the coefficient of determination R^2; and
3. the ratio of performance to interquartile range (RPIQ).

RMSE is the standard deviation of the residuals (prediction errors) and is calculated thusly:

$$\mathrm{RMSE}(\mathbf{y}, \hat{\mathbf{y}}) = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)}{N}} \tag{5}$$

where $\mathbf{y}, \hat{\mathbf{y}}$ are the vector of ground truth data and the model predictions, respectively, of the N patterns.

R^2 is determined via:

$$R^2(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} \tag{6}$$

with $\overline{y}$ representing the mean ground truth value.

RPIQ (Bellon-Maurel et al., 2010) is defined as the interquartile range of the ground truth data divided by the RMSE of the prediction, and because it does not make any assumptions about the distribution of the observed values is considered a more robust accuracy metric:

$$\mathrm{RPIQ}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{Q_3 - Q_1}{\mathrm{RMSE}(\mathbf{y}, \hat{\mathbf{y}})} \tag{7}$$

where Q_1 and Q_3 are the first and third quartiles of $\mathbf{y}$, respectively.

Table 3
Descriptive statistics of the °Brix of the sampled berries per each grape variety, where std is the standard deviation, min and max are then minima and maxima, while Q_1, Q_2, and Q_3 are the 1st, 2nd, and 3rd quartiles, respectively.

Variety	N	mean	std	min	Q_1	Q_2	Q_3	max
Chardonnay	54	21.9	3.3	12.0	20.2	22.1	24.0	30.0
Malagouzia	71	21.3	2.7	16.9	19.7	21.2	22.8	30.0
Sauvignon-Blanc	53	23.3	4.4	10.0	21.5	24.9	25.9	31.4
Syrah	55	20.0	3.4	14.2	17.0	20.1	22.6	28.3

It should be noted that because the models employed herein are non-linear, the coefficient of determination loses its interpretation of proportion of variance explained, thus its magnitude alone may not be sufficient to assess the models' goodness of fit (Kvalseth, 1985; Willett and Singer, 1988). Therefore, R^2 should always be complemented with the RMSE metric and an understanding of the dataset's output distribution (Table 3) for a proper interpretation of model performance.

2.4.6. Statistical tests

The Wilcoxon signed-rank test is a non-parametric statistical test hypothesis test employed to ascertain whether two populations using paired samples differ (Hollander et al., 2015). One of its merits is that it makes no assumptions about the data's underlying distributions, and instead, it relies on the ranks of the data. The test works by comparing the differences between the two samples, and then calculating the ranks of those differences. The signed-ranks are then used to calculate a test statistic, which is compared to a critical value from a table of values or a p-value is calculated. If the test statistic is greater than the critical value, or the p-value is less than the significance level, then the null hypothesis is rejected, indicating that the two samples are significantly different. In this hypothesis test, the null hypothesis H_0 is that the median difference between the two populations is zero while one- or two-sided alternate hypotheses may be formed. The Wilcoxon signed-rank test is employed herein to ascertain whether there are statistical differences between the proposed methodologies, and more specifically: (i) if DCAE outperforms DAE; (ii) if there is a difference between using the encoded or the best decoded space as the input to the machine learning models; and (iii) if using simultaneously multiple pre-treatments is better than the best single pre-treatment. In all tests a confidence interval of $\alpha = 0.05$ is used.

3. Results

3.1. Dataset

In total, the dataset is comprised of 233 individual measurements, each corresponding to a single berry and consisting of: (i) in situ spectra extracted from the VNIR hyperspectral cubes, (ii) in situ spectra collected using the handheld VNIR–SWIR spectrometer, (iii) the sugar content in °Brix, (iv) the grape variety, and (v) the date corresponding to the measurement.

The major statistical moments of the °Brix content across all varieties are provided in Table 3. As these results indicate, there is sufficient variability in terms of the maturity of the grapes, capturing the most critical phases to select the optimal harvest time which depends on the variety. For red wine varieties the optimal values are usually between 23–26°Brix, whereas white varieties are slightly lower at about 21–24°Brix.

The spectral signals extracted from the hyperspectral camera and the reflectance spectra recorded by the point spectrometer are presented in Fig. 7. It should be noted that although overlaid, they depict different physical quantities with the hyperspectral camera presenting the raw recorded spectrum in digital number (DN), while the point spectrometer providing the reference reflectance data. It is possible to discern that the white grape varieties (namely, Chardonnay, Malagouzia and Sauvignon-Blanc) have similar spectra while the major absorbance peaks at about 680 and 770 nm for are identifiable in both spectral sources. On the other hand, for Syrah (the red grape variety) there are notable differences compared to the white grape varieties in terms of the shape of the spectral signatures, and there is only a slight absorbance peak at approximately 770 nm. In all cases, the largest differences in the shapes is to be found in the near infrared range and particular at above 800 nm; the spectrometer has high reflectance values compared to the raw measurement from the hyperspectral camera. This is attributed to the fact that the hyperspectral camera records using the sun as an illumination source and its blackbody radiation curve has lower intensity in these range, whereas the reflectance of the spectrometer is a calibrated value which takes into account the white reference measurement.

Table 4
Mean similarity between $\mathbf{X}'$ and $\mathbf{X}_{sp}$ for each autoencoder strategy considered.

Method	Pre-treatment	Mean similarity		
		ρ	ρ_c	SAM
DAE	SNV	0.8288	0.6575	0.8517
	SG1+SNV	0.8185	0.6364	0.8744
	SG2+SNV	0.7374	0.4745	1.0636
	Multi	0.7948	0.5892	0.9349
DCAE	SNV	0.8328	0.6648	0.8427
	SG1+SNV	0.8007	0.5918	0.9249
	SG2+SNV	0.7264	0.4512	1.0947
	Multi	0.7937	0.5854	0.9338

3.2. Autoencoder

The results of the different autoencoder architectures considered were evaluated using the similarity metrics between the transformed spectra and the reference reflectance data, and are provided in Table 4. As indicated, according to all the metrics, the highest similarity may be found when using either the dense autoencoders (i.e., DAE) or the convolutional autoencoders (i.e., DCAE) but only employing a single pre-treatment and more precisely the SNV transform. The worst pre-treatment appears to be SG2+SNV which yields the lowest mean similarity in both scenarios, something that ostensibly also affects the multi-channel scenario whose similarity values are potentially skewed due to the inclusion of also this pre-treatment method. This indicates that the process of de-noising is more involved when the spectral derivatives are concerned, and particularly so when the second derivative is used; this effect is less pronounced in the SG1+SNV scenario.

Fig. 8 presents an example of the input and output of the DCAE using multiple input spectral pre-treatments for a spectral signature extracted from a single pixel of a hyperspectral cube depicting the Chardonnay variety. The input to the DCAE is comprised of three channels corresponding to three spectra pre-treatments applied in the recorded uncalibrated (show in subfigures a, d, and g). Its output, as well as the reference spectra with which it is compared, is presented in subfigures b, e, and h; the difference between the two is showcased in the third column and specifically in subfigures c, f, and i. It is noteworthy that the autoencoder has identified the spectral information present in the ranges above 800 nm where the input data exhibit less information than the reference spectra.

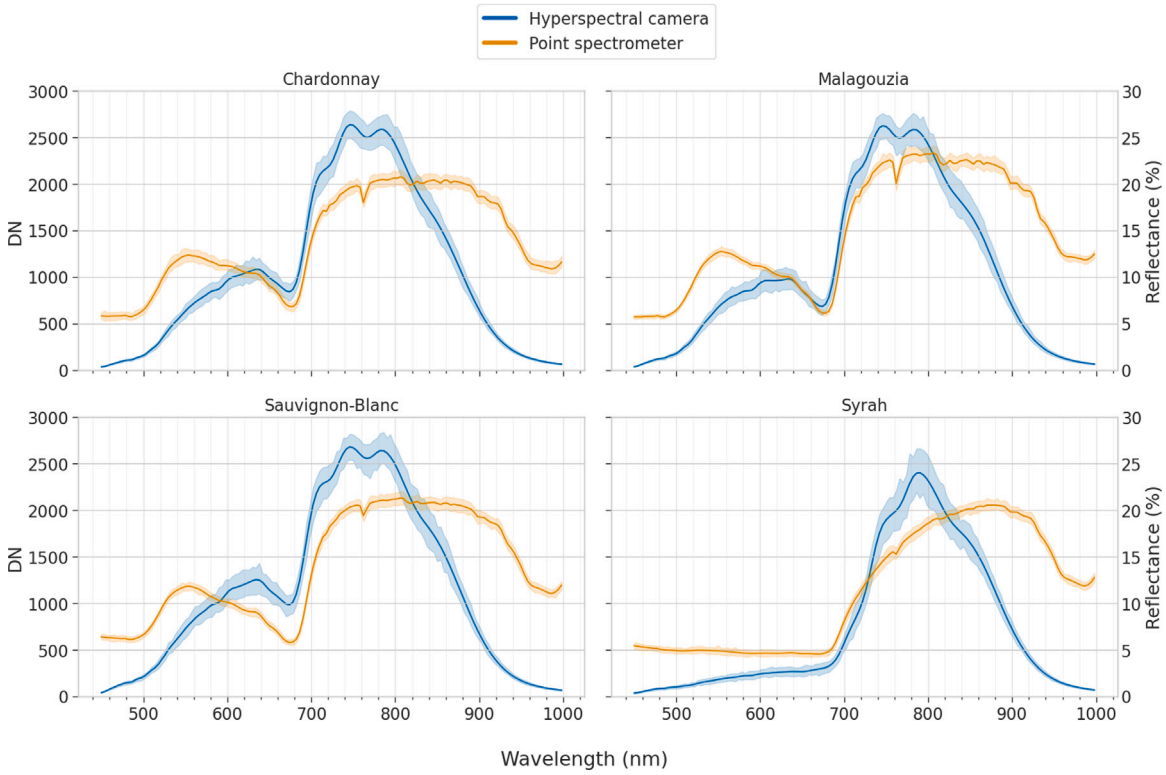


Fig. 7. Mean spectra with the 95% confidence interval per each of the four grape varieties showcasing the raw uncalibrated data from the hyperspectral camera (source, *y*-axis labels on the left) and the target reflectance spectra from the spectrometer (reference, *y*-axis labels on the right).

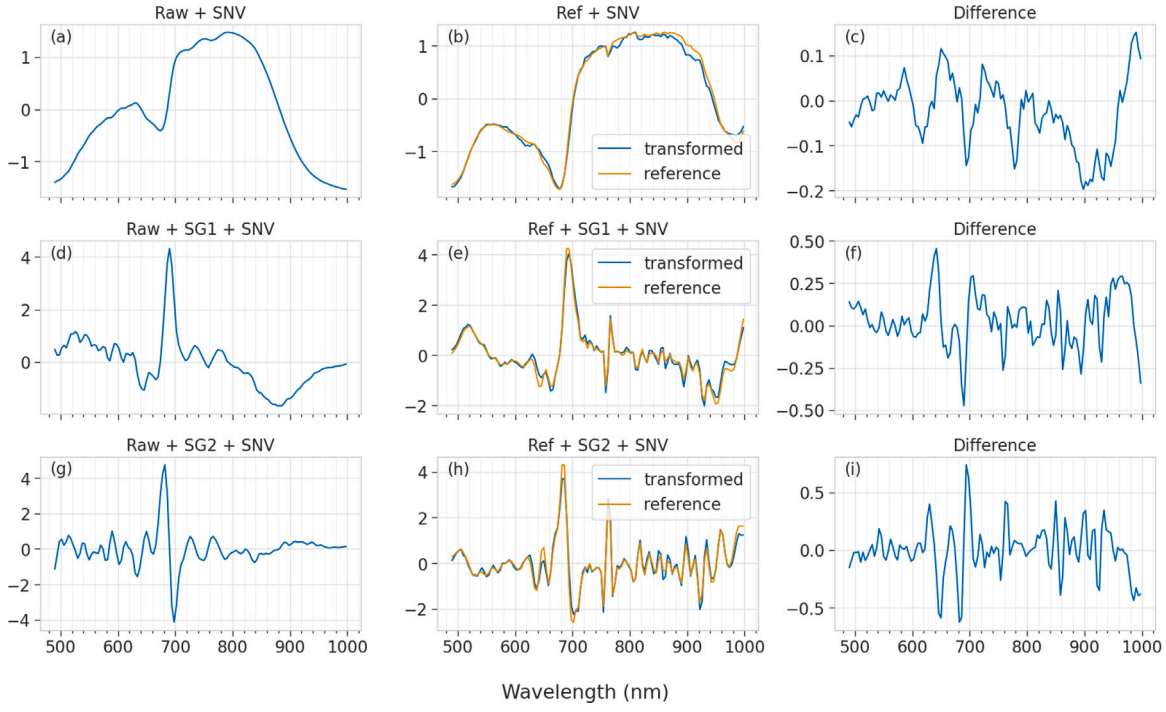


Fig. 8. Example of the denoising procedure for the DCAE approach using three channels corresponding to spectral pre-treatments (namely, SNV, SG1 + SNV, and SG2 + SNV) for a single measurement. The first column (i.e., subfigures a, d, and g) depicts the uncalibrated signals recorded from the hyperspectral camera which is the input to the DCAE, the second column (i.e., subfigures b, e and h) shows the output of DCAE which is the transformed signal and the reference (calibrated) spectrum, while the third column (i.e., subfigures c, f and i) illustrates the difference between them.

Table 5
Best sugar content (°Brix) prediction results in the independent test set across the four grape varieties, autoencoder methodology, and input space used; the best results per each variety are denoted with bold.

Autoencoder	Space	Channels	Pre-treatment	Model	Accuracy		
					RMSE	R^2	RPIQ
Chardonnay							
DAE	Decoded	single	SG2+SNV	Cubist	1.72	0.72	2.50
	Encoded	single	SG2+SNV	Random Forest	**1.66**	**0.74**	**2.60**
DCAE	Decoded	single	SG1+SNV	Random Forest	1.84	0.68	2.34
	Encoded	single	SG1+SNV	Cubist	1.79	0.69	2.40
Malagouzia							
DAE	Decoded	multi	SG2+SNV	XGBoost	**1.72**	**0.62**	**1.98**
	Encoded	single	SNV	Random Forest	1.79	0.59	1.90
DCAE	Decoded	single	SG2+SNV	XGBoost	1.86	0.55	1.82
	Encoded	single	SNV	XGBoost	1.78	0.59	1.91
Sauvignon-Blanc							
DAE	Decoded	single	SNV	Random Forest	2.53	0.68	1.71
	Encoded	single	SNV	Random Forest	2.73	0.63	1.58
DCAE	Decoded	multi	SG1+SNV	XGBoost	**2.29**	**0.74**	**1.89**
	Encoded	single	SNV	Random Forest	2.43	0.70	1.78
Syrah							
DAE	Decoded	single	SG1+SNV	Random Forest	1.92	0.58	2.81
	Encoded	single	SG1+SNV	XGBoost	1.70	0.67	3.18
DCAE	Decoded	single	SNV	Support Vector Regression	**1.66**	**0.69**	**3.25**
	Encoded	single	SNV	Random Forest	1.92	0.58	2.81

3.3. Modeling

The results in the independent test set are provided in Table 5; due to the large number of experiments conducted, the table presents the best results per each variety, autoencoder methodology, and the input space (i.e., the encoded latent space or the decoded spectra) and pre-processing method employed. Moreover, it is noted that the best results in each variety are denoted with boldface. The rest of the results are summarized in the form of boxplots in Fig. 9 which enable the comparison of the various methodologies presented herein. It is worth remarking that there is considerable variance in the results, indicating that careful tuning and exploration of the various pre-treatments and machine learning algorithms ought to take place.

To ascertain if there is a statistical difference between the two autoencoder methodologies, namely DAE or DCAE, the Wilcoxon signed-rank test is employed using the RMSE as the measurement values. The null hypothesis H_0 is that the median of their differences is zero, while the one-sided alternative hypothesis H_1 is that the DCAE outperforms DAE. The test statistic yields a value of $W = 2642$ which is equivalent to a p-value of 0.012; when compared to the confidence interval of $\alpha = 0.05$, it suggests that the H_0 is rejected while H_1 holds, thus DCAE statistically outperforms DAE.

Another interesting comparison pertains to the input provided to the machine learning algorithms, and the investigation whether the encoded (latent) space is more suitable for predictions as opposed to the denoised transformed spectra. To this end, the null hypothesis H_0 of the Wilcoxon signed-rank test is that the median of their RMSE differences is zero, with the two-sided alternative hypothesis H_2 being that there is a statistically significant difference between them. The result of the test is $W = 368$ corresponding to a p-value of $0.661 > \alpha$, indicating that the null hypothesis hold, i.e., there is no statistically significant difference between the usage of the encoded or the decoded spaces. Thus, as also evidenced by Fig. 9, both approaches ought to be tested to identify the optimal model.

As a final comparison, it was investigated whether the best single input channel outperforms the combined usage of multiple input channels. In this case, the null hypothesis H_0 states that the two approaches are equivalent, with the one-sided alternative hypothesis H_1 being that the best single input channel is better. The test statistic is equal to $W = 2890$, with the equivalent p-value being $0.0003 < \alpha$, hence signifying that the multi-input channels yield statistically worse results.

4. Discussion

While the application of hyperspectral imaging in the laboratory has successfully demonstrated its capacity to robustly estimate various maturity indicators in fruits (including grapes), there are limited studies in real-life field conditions. This is mainly due to two factors: (i) many commercial hyperspectral imaging devices employ push-broom (or line-scanning) technologies thus require either a linear translation unit or very stable (e.g., lack of wind) conditions for robust measurements, and (ii) the field conditions demonstrate significant variability due to varying illumination conditions both within a single scene (due to e.g., shadows) and among different scenes (e.g., partial cloud presence obstructing momentarily the sun), In this paper, we address the first limitation by using a snapshot hyperspectral camera, and the second limitation by proposing the usage of the autoencoder framework to transfer raw recorded spectra to calibrated reflectance values. Another major advantage of this technique, is its capacity to overcome the inherent difficulty of continuously calibrating the camera using white reflectance panels in the field, which is still present even when custom set-ups are employed which utilize artificial light sources as these may fluctuate (e.g., depending on the ambient temperature) or the sensor's response may change as it overheats from continuous usage.

To this end, firstly, various autoencoder architectures are examined, namely the deep autoencoder framework which employs fully connected layers and the deep convolutional autoencoder framework utilizing the convolutional operation. Furthermore, the combined usage of multiple spectral input sources (as formed through common pre-processing techniques) is investigated. As a first step, the paper examines the similarity between the output of the autoencoders' which are the transformed spectra $\mathbf{X}'$ and the reference reflectance measurements $\mathbf{X}_{sp}$ (Table 4). The closest similarity is observed when using the SNV or the SG1+SNV pre-treatments, i.e., only a simple scatter correction technique or the calculation of the first derivative. The use of the second derivative, denoted as SG2+SNV, yields the worst similarity, which may be attributed to the fact they are calculated by taking the derivative of the first derivative. They are thus *ipso facto* in general more sensitive to noise in the signal and may potentially amplify the noise and cause distortions in the signal, such as amplifying high-frequency noise or smoothing out sharp peaks in the signal. Specifically this impedes the process of transforming the raw measurements by the hyperspectral camera $\mathbf{X}_{hs}$ into the more noisy second-derivative spectra of $\mathbf{X}_{sp}$, particularly in the infrared where the signal-to-noise ratio is

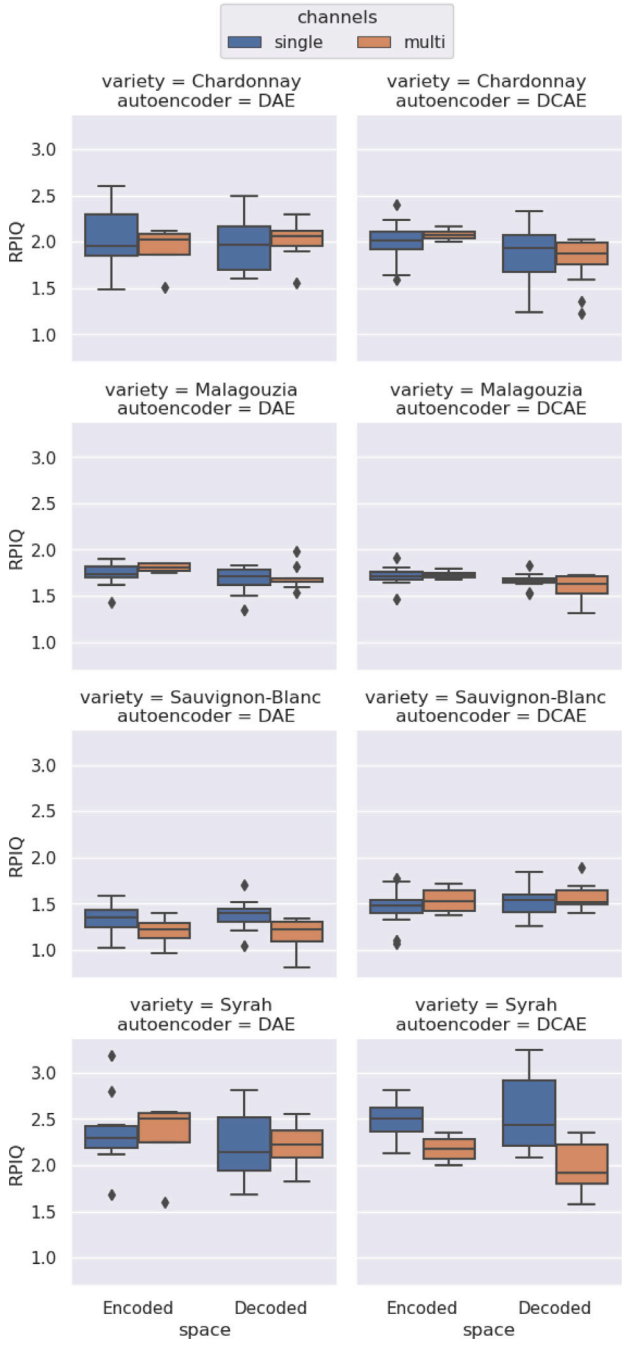


Fig. 9. Prediction results in the independent test set summarized across the four grape varieties, grouped per each autoencoder methodology, number of input channels and input space used; higher values are better.

higher. At the same time, it is noteworthy that there is no significant difference between DAE and DCAE despite the latter employing about 10 times more trainable parameters.

Following this de-noising framework, the second step involves the application of machine learning models to predict the sugar content. The input to the model may be either the compressed latent space from the autoencoder, or the denoised spectra produced by the autoencoder (i.e., the decoder's output $\mathbf{X}'$). Interestingly, despite the fact the similarity metrics employed indicated that there is not a significant difference between DAE and DCAE, nevertheless the performance of the machine learning models indicates otherwise. More specifically, although DAE yields the best models for Chardonnay and Malagouzia whilst DCAE is the best method for Sauvignon-Blanc and Syrah, the Wilcoxon signed-rank test signified that the median RMSE of DCAE is statistically lower than the one of DAE. This may signify that DCAE better captures the information which is considered as spectrally important for the models to more robustly determine the °Brix content, and is thus more favorable for the herein presented application. When the use of the encoded space is compared to the decoded spectra, no statistically significant differences were the observed. Despite this, it bears mentioning that both approaches ought to be tested, considering that as shown in Fig. 9 in some cases the encoded space fares better (e.g., Chardonnay with DCAE and Syrah with DAE) while in other cases the decoded space yields enhanced prediction results (e.g., Sauvignon-Blanc with DCAE). In the third statistical comparison that took place, it was demonstrated that the use of multiple input channels is statistically worse than the best single input channel. This is probably due to the inclusion of the second derivative, which as evidenced in Table 4, is more difficult to predict. The effect of incorporating the second derivative in the multi-input channels appears to deteriorate the performance of the machine learning because the autoencoder attempts to encode the noisy part of the spectrum and the transformed spectra have a lower resemblance to the reference values.

All in all, the best models exhibit excellent performance (RPIQ > 2) in the Chardonnay and Syrah varieties and provide good estimations (RPIQ ≈ 2) for the Malagouzia and Sauvignon-Blanc varieties. The RMSE values of the best models range between 1.66 °Brix and 2.29 °Brix which although they are larger than the resolution of the refractometer that provided the ground truth measurements (at 0.2 °Brix), they nevertheless provide a fair estimate for the sugar content of the grapes in terms of their maturity stage. Thus, the effect of the accuracy of estimation of the ground truth measurements in the developed models is considered negligible. Compared to other studies which predict the sugar content via point spectroscopy (like e.g., in Kalopesa et al., 2023) the results are understandably slightly worse, but are still comparable. In the same varieties, Kalopesa et al. (2023) report a mean RPIQ of 3.70, whereas here the mean RPIQ is 2.43. Rodríguez-Pulido et al. (2022) report an RMSE of 1.40 °Brix in the examined varieties (Syrah and Tempranillo) when transferring their laboratory model in the field, which is comparable to the RMSE of 1.66 °Brix for Syrah reported herein. Their application however uses a push-broom portable hyperspectral camera and requires continuous white reference measurements close to the target which is less than ideal for robotic platforms. This testifies to the robustness of the herein introduced methodology.

With respect to the limitations of the present study, it should be noted that whereas the proposed framework successfully manages to map the uncalibrated signal from the hyperspectral camera to the calibrated signature of the handheld spectrometer, nevertheless this transformation is device specific and hence the training of the autoencoder has to re-take place if a different hyperspectral camera is used, even if it uses the exact same spectral range and bands. Moreover, although the data collected in the present study was without artificial light sources, nevertheless the integration time was picked in the field during each measurement, and hence a more automated and structured practice is needed. Closely related to this, is the inclusion of more outlying values (e.g., very dimly or very brightly illuminated targets) to the examine the robustness of the proposed methodology. The effect of the ambient temperature and of the prolonged in-situ operation of the hyperspectral camera which causes heating on the spectral acquisitions should also be studied in the future. Undoubtedly, the current study focused only on four varieties; considering the plethora of grape varieties and growing conditions present in viticulture, further work is necessary on more diverse datasets to validate the generalization ability of the proposed framework. Lastly, it should be noted that to incorporate this method on a moving robotic platform (compared to the stationary approach employed herein) more analysis is required,

although due to the sensor's low integration time (ca. 20 ms) the effect of the movement on the hyperspectral cube is expected to be limited.

In the future, automatic segmentation techniques (like e.g. in Lu et al., 2022) for real-time detection of the grape bunch could be deployed in combination with our proposed methodology to enable in situ detection of grape maturity via robotic platforms. An optimization procedure to automatically select the most optimal integration time from a given scene could also be studied, to further automate the procedure. Finally, other maturity indicators that are essential to gauge the grapes' ripeness like pH and total acidity, ought to be included in future studies, in addition to °Brix (Chávez-Segura and Vejarano, 2022).

5. Conclusions

In this paper a novel autoencoder-based framework was presented which enables hyperspectral imaging to become more operational in the field, by collecting raw signatures and transforming them into reference reflectance data accounting for the different illumination conditions and shadowing effects. The proposed technique assists in overcoming the aforementioned difficulties and enables the mounting of hyperspectral cameras on robotic platforms for continuous monitoring, without the need to use artificial light sources or continuous white reference measurements (i.e., device calibration). In terms of the most suitable autoencoder architecture, it was demonstrated that both dense and convolutional autoencoders can sufficiently transform the raw spectra to standardized values. In the particular case of the estimation of sugar content in vinified grapes examined in this paper, the convolutional autoencoders result in higher overall accuracy while the use of either the encoded (latent) space or the decoded transformed spectra lead to similar performances. When the use of multiple spectra pre-treatments was investigated, it was shown that the simple SNV pre-treatment performs better in terms of reconstruction error and prediction performance when compared to the use of spectral derivatives, which also led to lower accuracy when all pre-treatments were combined simultaneously and served as input to the autoencoder.

CRediT authorship contribution statement

Nikolaos L. Tsakiridis: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing, Visualization. **Nikiforos Samarinas:** Methodology, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Stylianos Kokkas:** Data curation, Visualization. **Eleni Kalopesa:** Data curation, Validation, Writing – review & editing. **Nikolaos V. Tziolas:** Conceptualization, Data curation, Writing – review & editing. **George C. Zalidis:** Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Nikolaos L. Tsakiridis reports financial support was provided by European Commission.

References

Awad, M., Khanna, R., 2015. Support vector regression. In: Efficient Learning Machines. Apress, pp. 67–80. http://dx.doi.org/10.1007/978-1-4302-5990-9_4.

Baiano, A., Terracone, C., Peri, G., Romaniello, R., 2012. Application of hyperspectral imaging for prediction of physico-chemical and sensory characteristics of table grapes. Comput. Electron. Agric. 87, 142–151. http://dx.doi.org/10.1016/j.compag.2012.06.002.

Bank, D., Koenigstein, N., Giryes, R., 2020. Autoencoders. arXiv e-prints, arXiv:2003.05991. http://dx.doi.org/10.48550/arXiv.2003.05991.

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.-M., McBratney, A., 2010. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. TrAC Trends Anal. Chem. 29 (9), 1073–1081. http://dx.doi.org/10.1016/j.trac.2010.05.006.

Benelli, A., Cevoli, C., Ragni, L., Fabbri, A., 2021. In-field and non-destructive monitoring of grapes maturity by hyperspectral imaging. Biosyst. Eng. 207, 59–67. http://dx.doi.org/10.1016/j.biosystemseng.2021.04.006.

Berman, D., Buczak, A., Chavis, J., Corbett, C., 2019. A survey of deep learning methods for cyber security. Information 10 (4), 122. http://dx.doi.org/10.3390/info10040122.

Bisson, L.F., Waterhouse, A.L., Ebeler, S.E., Walker, M.A., Lapsley, J.T., 2002. The present and future of the international wine industry. Nature 418 (6898), 696–699. http://dx.doi.org/10.1038/nature01018.

Breiman, L., 2001. Mach. Learn. 45 (1), 5–32. http://dx.doi.org/10.1023/a:1010933404324.

Chávez-Segura, G., Vejarano, R., 2022. White grape quality monitoring via hyperspectral imaging: from the vineyard to the winery. In: White Wine Technology. Elsevier, pp. 17–27. http://dx.doi.org/10.1016/b978-0-12-823497-6.00003-x.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, ACM, New York, NY, USA, pp. 785–794. http://dx.doi.org/10.1145/2939672.2939785, URL: http://doi.acm.org/10.1145/2939672.2939785.

Chen, S., Zhang, F., Ning, J., Liu, X., Zhang, Z., Yang, S., 2015. Predicting the anthocyanin content of wine grapes by NIR hyperspectral imaging. Food Chem. 172, 788–793. http://dx.doi.org/10.1016/j.foodchem.2014.09.119.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297. http://dx.doi.org/10.1007/bf00994018.

Eurostat, 2022. Crop Production in EU Standard Humidity [APRO_CPSH1]. Eurostat, URL: https://ec.europa.eu/eurostat/databrowser/view/APRO_CPSH1_custom_4136337/default/table?lang=en. (Accessed 12 December 2022).

FAO, 2021. World Food and Agriculture – Statistical Yearbook 2021. FAO, Rome, Italy, http://dx.doi.org/10.4060/cb4477en.

Ferrag, M.A., Maglaras, L., Moschoyiannis, S., Janicke, H., 2020. Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. J. Inf. Secur. Appl. 50, 102419. http://dx.doi.org/10.1016/j.jisa.2019.102419.

Ferrer-Gallego, R., Rodríguez-Pulido, F.J., Toci, A.T., García-Estevez, I., 2022. Phenolic composition, quality and authenticity of grapes and wines by vibrational spectroscopy. Food Rev. Int. 38 (5), 884–912. http://dx.doi.org/10.1080/87559129.2020.1752231.

Fountas, S., Mylonas, N., Malounas, I., Rodias, E., Santos, C.H., Pekkeriet, E., 2020. Agricultural robotics for field operations. Sensors 20 (9), 2672. http://dx.doi.org/10.3390/s20092672.

Gabrielli, M., Lançon-Verdier, V., Picouet, P., Maury, C., 2021. Hyperspectral imaging to characterize table grapes. Chemosensors 9 (4), 71. http://dx.doi.org/10.3390/chemosensors9040071.

Global Industry Analysts, I., 2022. Wine: Global Strategic Business Report. Technical Report, Research And Markets, URL: https://researchandmarkets.com/reports/338680/.

Gomes, V., Reis, M.S., Rovira-Más, F., Mendes-Ferreira, A., Melo-Pinto, P., 2021. Prediction of sugar content in port wine vintage grapes using machine learning and hyperspectral imaging. Processes 9 (7), 1241. http://dx.doi.org/10.3390/pr9071241.

Gutiérrez, S., Fernández-Novales, J., Diago, M.P., Tardaguila, J., 2018. On-the-go hyperspectral imaging under field conditions and machine learning for the classification of grapevine varieties. Front. Plant Sci. 9, http://dx.doi.org/10.3389/fpls.2018.01102.

Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313 (5786), 504–507. http://dx.doi.org/10.1126/science.1127647.

Hollander, M., Wolfe, D.A., Chicken, E., 2015. Nonparametric Statistical Methods. Wiley, Hoboken, NJ, USA, http://dx.doi.org/10.1002/9781119196037.

International Organisation of Vine and Wine, 2017. Distribution of the world's grapevine varieties. URL: https://www.oiv.int/public/medias/5888/en-distribution-of-the-worlds-grapevine-varieties.pdf. (Accessed 12 December 2022).

International Organisation of Vine and Wine, 2022a. State of the world vine and wine sector 2021. URL: https://www.oiv.int/sites/default/files/documents/ENG_State%20of%20the%20world%20vine%20and%20wine%20sector%20April%202022%20v6.pdf. (Accessed 12 December 2022).

International Organisation of Vine and Wine, 2022b. World wine production outlook — OIV first estimates. URL: https://www.oiv.int/sites/default/files/documents/EN_OIV_2022_World_Wine_Production_Outlook_1.pdf. (Accessed 12 December 2022).

Kalopesa, E., Karyotis, K., Tziolas, N., Tsakiridis, N., Samarinas, N., Zalidis, G., 2023. Estimation of sugar content in wine grapes via in situ VNIR–SWIR point spectroscopy using explainable artificial intelligence techniques. Sensors 23 (3), 1065. http://dx.doi.org/10.3390/s23031065.

Kruse, F., Lefkoff, A., Boardman, J., Heidebrecht, K., Shapiro, A., Barloon, P., Goetz, A., 1993. The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data. Remote Sens. Environ. 44 (2–3), 145–163. http://dx.doi.org/10.1016/0034-4257(93)90013-n.

Kuhn, M., Johnson, K., 2013. Regression trees and rule-based models. In: Applied Predictive Modeling. Springer, New York, pp. 173–220. http://dx.doi.org/10.1007/978-1-4614-6849-3_8.

Kussul, N., Lavreniuk, M., Skakun, S., Shelestov, A., 2017. Deep learning classification of land cover and crop types using remote sensing data. IEEE Geosci. Remote Sens. Lett. 14 (5), 778–782. http://dx.doi.org/10.1109/lgrs.2017.2681128.

Kvalseth, T.O., 1985. Cautionary note about R^2. Amer. Statist. 39 (4), 279. http://dx.doi.org/10.2307/2683704.

Lin, L.I.-K., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45 (1), 255. http://dx.doi.org/10.2307/2532051.

Liu, S., Fan, S., Lin, L., Huang, W., 2022. An improved method for predicting soluble solids content in apples by heterogeneous transfer learning and near-infrared spectroscopy. Comput. Electron. Agric. 203, 107455. http://dx.doi.org/10.1016/j.compag.2022.107455.

Liu, T., Li, Z., Yu, C., Qin, Y., 2017. NIRS feature extraction based on deep auto-encoder neural network. Infrared Phys. Technol. 87, 124–128. http://dx.doi.org/10.1016/j.infrared.2017.07.015.

Lorenz, D.H., Eichhorn, K.W., Bleiholder, H., Klose, R., Meier, U., Weber, E., 1995. Growth Stages of the Grapevine: Phenological growth stages of the grapevine (Vitis vinifera L. ssp. vinifera)—Codes and descriptions according to the extended BBCH scale. Aust. J. Grape Wine Res. 1 (2), 100–103. http://dx.doi.org/10.1111/j.1755-0238.1995.tb00085.x.

Lu, S., Liu, X., He, Z., Zhang, X., Liu, W., Karkee, M., 2022. Swin-transformer-YOLOv5 for real-time wine grape bunch detection. Remote Sens. 14 (22), 5853. http://dx.doi.org/10.3390/rs14225853.

Lv, S., Zhao, S., Li, D., Pang, B., Lian, X., Liu, Y., 2023. Spatial–spectral joint hyperspectral anomaly detection based on a two-branch 3D convolutional autoencoder and spatial filtering. Remote Sens. 15 (10), 2542. http://dx.doi.org/10.3390/rs15102542.

Mantripragada, K., Dao, P.D., He, Y., Qureshi, F.Z., 2022. The effects of spectral dimensionality reduction on hyperspectral pixel classification: A case study. In: da Silva, C.R. (Ed.), PLoS One 17 (7), e0269174. http://dx.doi.org/10.1371/journal.pone.0269174.

Masci, J., Meier, U., Cireşan, D., Schmidhuber, J., 2011. Stacked convolutional auto-encoders for hierarchical feature extraction. In: Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pp. 52–59. http://dx.doi.org/10.1007/978-3-642-21735-7_7.

McGovern, P.E., 2013. Ancient Wine: The Search for the Origins of Viniculture. Princeton University Press, pp. 1–365.

Mishra, P., Lohumi, S., Khan, H.A., Nordon, A., 2020. Close-range hyperspectral imaging of whole plants for digital phenotyping: Recent applications and illumination correction approaches. Comput. Electron. Agric. 178, 105780. http://dx.doi.org/10.1016/j.compag.2020.105780.

Nalepa, J., Myller, M., Imai, Y., Honda, K.-I., Takeda, T., Antoniak, M., 2020. Unsupervised segmentation of hyperspectral images using 3-D convolutional autoencoders. IEEE Geosci. Remote Sens. Lett. 17 (11), 1948–1952. http://dx.doi.org/10.1109/lgrs.2019.2960945.

Niimi, J., Boss, P.K., Jeffery, D., Bastian, S.E.P., 2017. Linking sensory properties and chemical composition of vitis vinifera cv. Cabernet sauvignon grape berries to wine. Am. J. Enol. Viticult. 68 (3), 357–368. http://dx.doi.org/10.5344/ajev.2017.16115.

Nissen, J.N., Johansen, J., Allesøe, R.L., Sønderby, C.K., Armenteros, J.J.A., Grønbech, C.H., Jensen, L.J., Nielsen, H.B., Petersen, T.N., Winther, O., Rasmussen, S., 2021. Improved metagenome binning and assembly using deep variational autoencoders. Nature Biotechnol. 39 (5), 555–560. http://dx.doi.org/10.1038/s41587-020-00777-4.

Palsson, B., Ulfarsson, M.O., Sveinsson, J.R., 2021. Convolutional autoencoder for spectral–spatial hyperspectral unmixing. IEEE Trans. Geosci. Remote Sens. 59 (1), 535–549. http://dx.doi.org/10.1109/tgrs.2020.2992743.

Polder, G., Blok, P.M., de Villiers, H.A.C., van der Wolf, J.M., Kamp, J., 2019. Potato virus Y detection in seed potatoes using deep learning on hyperspectral images. Front. Plant Sci. 10, http://dx.doi.org/10.3389/fpls.2019.00209.

Power, A., Truong, V.K., Chapman, J., Cozzolino, D., 2019. From the laboratory to the vineyard—evolution of the measurement of grape composition using NIR spectroscopy towards high-throughput analysis. High-Throughput 8 (4), 21. http://dx.doi.org/10.3390/ht8040021.

Prasanna, V., Prabha, T.N., Tharanathan, R.N., 2007. Fruit ripening phenomena–an overview. Crit. Rev. Food Sci. Nutr. 47 (1), 1–19. http://dx.doi.org/10.1080/10408390600976841.

Reddi, S.J., Kale, S., Kumar, S., 2019. On the convergence of adam and beyond. CoRR, arXiv:1904.09237.

Rodríguez-Pulido, F.J., Mora-Garrido, A.B., González-Miret, M.L., Heredia, F.J., 2022. Research progress in imaging technology for assessing quality in wine grapes and seeds. Foods 11 (3), 254. http://dx.doi.org/10.3390/foods11030254.

Shoeibi, A., Ghassemi, N., Alizadehsani, R., Rouhani, M., Hosseini-Nejad, H., Khosravi, A., Panahiazar, M., Nahavandi, S., 2021. A comprehensive comparison of handcrafted features and convolutional autoencoders for epileptic seizures detection in EEG signals. Expert Syst. Appl. 163, 113788. http://dx.doi.org/10.1016/j.eswa.2020.113788.

Signoroni, A., Savardi, M., Baronio, A., Benini, S., 2019. Deep learning meets hyperspectral image analysis: A multidisciplinary review. J. Imag. 5 (5), 52. http://dx.doi.org/10.3390/jimaging5050052.

Silva, R., Gomes, V., Mendes-Faia, A., Melo-Pinto, P., 2018. Using support vector regression and hyperspectral imaging for the prediction of oenological parameters on different vintages and varieties of wine grape berries. Remote Sens. 10 (2), 312. http://dx.doi.org/10.3390/rs10020312.

Silva, R., Melo-Pinto, P., 2021. A review of different dimensionality reduction methods for the prediction of sugar content from hyperspectral images of wine grape berries. Appl. Soft Comput. 113, 107889. http://dx.doi.org/10.1016/j.asoc.2021.107889.

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2014. Striving for simplicity: The all convolutional net. arXiv. http://dx.doi.org/10.48550/ARXIV.1412.6806. URL: https://arxiv.org/abs/1412.6806.

Su, Y., Li, J., Plaza, A., Marinoni, A., Gamba, P., Chakravortty, S., 2019. DAEN: Deep autoencoder networks for hyperspectral unmixing. IEEE Trans. Geosci. Remote Sens. 57 (7), 4309–4321. http://dx.doi.org/10.1109/tgrs.2018.2890633.

Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.-A., 2008. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning. ICML '08, Association for Computing Machinery, New York, NY, USA, pp. 1096–1103. http://dx.doi.org/10.1145/1390156.1390294.

Wang, X., Wang, Z., Zhang, Y., Jiang, X., Cai, Z., 2021. Latent representation learning based autoencoder for unsupervised feature selection in hyperspectral imagery. Multimedia Tools Appl. 81 (9), 12061–12075. http://dx.doi.org/10.1007/s11042-020-10474-8.

Wendel, A., Underwood, J., 2017. Illumination compensation in ground based hyperspectral imaging. ISPRS J. Photogram. Remote Sens. 129, 162–178. http://dx.doi.org/10.1016/j.isprsjprs.2017.04.010.

Willett, J.B., Singer, J.D., 1988. Another cautionary note about R^2: Its use in weighted least-squares regression analysis. Amer. Statist. 42 (3), 236. http://dx.doi.org/10.2307/2685031.

Xu, M., Sun, J., Cheng, J., Yao, K., Wu, X., Zhou, X., 2022. Non-destructive prediction of total soluble solids and titratable acidity in Kyoho grape using hyperspectral imaging and deep learning algorithm. Int. J. Food Sci. Technol. 58 (1), 9–21. http://dx.doi.org/10.1111/ijfs.16173.

Zhang, L., Cheng, B., 2019. A stacked autoencoders-based adaptive subspace model for hyperspectral anomaly detection. Infrar. Phys. Technol. 96, 52–60. http://dx.doi.org/10.1016/j.infrared.2018.11.015.

Zhang, L., Jin, J., Wang, L., Rehman, T.U., Gee, M.T., 2022. Elimination of leaf angle impacts on plant reflectance spectra using fusion of hyperspectral images and 3D point clouds. Sensors 23 (1), 44. http://dx.doi.org/10.3390/s23010044.

Zhang, C., Wu, W., Zhou, L., Cheng, H., Ye, X., He, Y., 2020. Developing deep learning based regression approaches for determination of chemical compositions in dry black goji berries (Lycium ruthenicum Murr.) using near-infrared hyperspectral imaging. Food Chem. 319, 126536. http://dx.doi.org/10.1016/j.foodchem.2020.126536.

Zhang, X., Yang, J., Lin, T., Ying, Y., 2021. Food and agro-product quality evaluation based on spectroscopy and deep learning: A review. Trends Food Sci. Technol 112, 431–441. http://dx.doi.org/10.1016/j.tifs.2021.04.008.