TOWARD AN AUTOMATIC WAY TO IDENTIFY RED BLOTCH INFECTED VINES FROM HYPERSPECTRAL IMAGES ACQUIRED IN THE FIELD

Authors: Eve LAROCHE-PINEL[1,2], Benjamin CORALES[1,2], Erica SAWYER[1,2,3], Khushwinder SINGH[1,2], Kaylah VASQUEZ[1,2], Monica COOPER[4], Marc FUCHS[5], Luca BRILLANTE[1,2*]

[1] *Department of Viticulture & Enology, California State University Fresno, Fresno, CA, USA*
[2] *Viticulture and Enology Research Center, California State University Fresno, Fresno, CA, USA*
[3] *Department of Mathematics, California State University Fresno, Fresno, CA, USA*
[4] *University of California, Agriculture & Natural Resources, Napa, CA, USA*
[5] *Plant Pathology and Plant-Microbe Biology, Cornell University, Geneva, NY, USA*

*Corresponding author: luca.brillante@csufresno.edu

Abstract

Context and purpose of the study – Vineyards are affected by different virus diseases, which can lower yield and affect the quality of grapes. Grapevine red blotch disease is one of them, and no curative solution exists. Once infected, a vine must be removed and replaced with a virus-free vine (aka roguing). Screening vineyards to look for symptoms can be time-consuming and needs well-trained experts. To improve this process, we conducted an experiment identifying infected vines using a hyperspectral camera in the field.

Material and methods – We monitored one vineyard in Rutherford, California, at the symptomatic stage in September and October 2020 and in August (pre-symptomatic stage), September and October 2021. More than 700 vines were sampled and analyzed through Polymerase Chain Reaction (PCR). We imaged the same vine canopies using a Senop HSC hyperspectral camera mounted on a tripod and captured 230 bands from the visible (510 nm) to the near-infrared (900 nm). We segmented leaves from the background through a U-Net neural network model and extracted the canopy signal. We tested different machine learning algorithms, Random Forest (RF), Partial Least Square (PLS), Support Vector Machine (SVM), and their multiple-model ensembles, to predict the PCR results (Infected vs. Non-infected). We evaluated and interpreted each model using mean accuracies, confusion matrices, and feature importance computation. We also computed a spectral binning and used recursive feature elimination (RFE) selection.

Results – The stacking ensemble of PLS and SVM models had the highest overall (cross-validated) accuracy of 69.5% for the entire dataset, 61% for the pre-symptomatic, and 74.5% for the symptomatic dataset. In this dataset, the model correctly classified non-infected vines with 83% accuracy and infected vines with 65% accuracy. Absolute values of PLS coefficients were the most important for reflectance at wavelengths between 550-600 nm and 750-800 nm. Concerning the permutation importance of the SVM model, the greatest values were obtained for reflectance around 600 nm, 710 nm, and 830 nm. These wavelengths are related to pigments known to be affected by red blotch. Using the RFE on the binning dataset, the overall accuracy reached 73.3% using 23 bands for the entire dataset and 76% using 30 bands for the symptomatic dataset. This study proves that hyperspectral imaging can help reduce the spread of red blotch by identifying vines that may be infected and could be rogued or molecularly analyzed if higher certainty is desired.

Keywords: Disease detection, Grapevine Red Blotch virus, hyperspectral imaging, machine learning, imaging spectroscopy

1. Introduction

Grapevine red blotch virus (GRBV) has been found to impede the accumulation of sugars and phenolic compounds, according to Ricketts et al. (2017) and Martinez-Lüscher et al. (2019). In the absence of any preventive measures, this disease can lead to substantial economic losses of up to $68,548/ha (Ricketts et al., 2017) over a vineyard's

25-year lifespan. Although removing diseased vines is effective, it is time-consuming. While molecular diagnostic techniques are the preferred method for detecting viral infections in vines, their high cost makes it impractical to test all vines. Consequently, a more automated approach to rapidly detect and diagnose viral diseases would be beneficial to vineyard managers.

The detection of plant diseases that cause changes in the optical signatures of plants due to their effects on biophysical and biochemical plant properties can be facilitated by remote sensing, as demonstrated by Croft and Chen (2018) and Knipling (1970). Hyperspectral imaging spectrometry has proven to be a powerful remote sensing tool, leading to a surge in research activity in this area (Terentev et al., 2022).

The objective of this research was to evaluate the potential of a hyperspectral camera that acquires images in the visible (VIS) and Near-Infrared (NIR) domains for detecting grapevine infections. The study involved the capture of canopy images of healthy and infected grapevines at various symptomatic stages (both pre- and post-symptom manifestation), with infection status determined by molecular analysis. Different machine-learning algorithms were employed to predict the grapevine infection status based on the hyperspectral images.

2. Material and methods

Plants measurements – The study took place in a vineyard in Napa Valley, California. In September and October 2020, as well as in August, September, and October 2021, four leaves were gathered from designated plants (671 samples in total). To evaluate viral infection, PCR-based analysis was performed on the petioles of the four collected leaves per plant. Different datasets were used in this study to consider the variability of the symptoms across the season. A first dataset with all the samples was considered, a second dataset with only the early season sample (from August 2021) was tested, and a third one with only the late season symptomatic dataset was taken into account.

Images acquisition and pre-processing - To conduct this study, a Senop HSC-2 hyperspectral camera was fixed to a tripod. The camera acquires images in 230 bands ranging from 510 to 900nm. The images were captured between 10 am and 4 pm on the same day as the sampling. The canopy signal was isolated from the background (soil, sky, and shadows) using a U-net model. The second stage of the processing involved diminishing the spatial dimensions of the images from 1048 to 256 pixels by implementing a bicubic interpolation over a neighborhood of 4x4 pixels. Finally, the reflectance values were averaged by images, and a Savitzky-Golay filter was employed to eliminate noise and smooth the spectra (Savitzky & Golay, 1964).

Classification - To perform binary classification (infected vs. non-infected), three distinct machine learning models, along with their various combinations, were evaluated. The first model employed was a Random Forest algorithm (RF), which is a frequently used classification model based on decision trees. The second model was a Partial Least Square Discriminant Analysis (PLS-DA), which is a commonly used model for spectral data analysis. The final model implemented in this study was a Support Vector Machine (SVM) classifier, which is particularly suitable for binary classification tasks. Additionally, model ensembles were evaluated through simple averaging, and model stacking was performed by utilizing the PLS-DA scores as predictors for the other models.

Spectral binning was employed to decrease the number of bands by averaging the reflectance values over every 5 bands. The binned spectra were subjected to Recursive Feature Elimination (RFE) for feature selection. RFE involves iteratively eliminating less important features from a dataset to identify the most significant features.

3. Results and discussion

3.1. *The best result using raw data are obtained with the PLS-DA and SVM ensemble model*

Table 1 provides a comprehensive summary of the overall accuracy achieved using different datasets with various machine learning models, including model ensembles and stacking. The optimal model to apply in this scenario appears to be the SVM, which yielded an accuracy of 69.2% for the complete dataset, 55.3% for the early season, and 74.5% for the late season. However, a higher accuracy rate was obtained by implementing model ensembles that averaged PLS-DA and SVM predictions. For the whole dataset, the accuracy increased by 0.3 points to 69.5%, and for the early season, the accuracy improved by 5.7 points to 61.0%. The accuracy of the late season dataset remained consistent. Conversely, utilizing the PLS-DA scores as predictors for the other models led to a decrease in accuracy. This can be explained by the fact that PLS and SVM feature importance showed different and complementary results, the PLS coefficients were found to be particularly significant for reflectance in the 550-600 nm and 750-800 nm wavelength ranges. In terms of the SVM model's permutation importance, the highest values were observed for reflectance at approximately 600 nm, 710 nm, and 830 nm.

Confusion matrices were computed to explore the classification results for each class obtained from the PLS-DA and SVM model ensemble (Table 2). The accuracy of the non-infected vine images is 80% for the entire dataset, and 83% for the late season dataset, indicating good predictive performance. However, for the early season asymptomatic images, the accuracy of the non-infected class drops to 63%, with a false positive rate of 37%. The accuracy of the infected vine images is 56% for the entire dataset, decreasing to 53% for the early season dataset and increasing to 65% for the late season dataset. Symptoms at an early stage seem to be more difficult to identify by the model but this dataset has also a smaller number of samples (76) which can also impact the training of the algorithm and its accuracy.

3.2. *A higher accuracy is obtained by reducing the number of bands and applying RFE*

Using the RFE on the binning dataset and using the SVM model, the overall accuracy reached 73.3% using 23 bands for the entire dataset, 59.8% using 19 bands for the early season dataset and 76% using 30 bands for the symptomatic dataset. Early season symptoms seem to still be difficult to identify but using the late season dataset, the accuracy exceeds 75% overall using 30 bands which are primarily located in the green (524nm to 578 nm), red (600nm to 715nm) and NIR (740nm to 774nm) domains.

4. Conclusions

Among the tested approaches, SVM and SVM_PLS-DA ensemble models seem to be the best to predict red-blotch infection using hyperspectral images of vines canopy. Using a RFE algorithm, the overall accuracy reaches 76% for the late season dataset using 30 bands located in green, red and NIR domains which are known to be reactive to virus infection.

5. Acknowledgments

Funding support for this project originated from the California Departmnent of Food and Agriculture, Specialty Crop Block Grant Program CDFA-SCBGP #19-0001-039-SF, and from the California State University-Agriculture Research Institute System Grant #21- 01-123. Authors thank the National Grape Research Alliance for early support of the project, Bronco Wine Company and St. Supery Estate Vineyards and Winery for access to vineyards and co-operation, and Takuy Frnzian, Khushwinder Singh, Guadalupe Partida, and Selena Vengco for help with sample collection and image acquisition.

6. Literature cited

CROFT, H., CHEN, J., 2018. Leaf Pigment Content. In Comprehensive Remote Sensing, 1-9: 117–142. Elsevier, doi:10.1016/B978-0-12-409548-9.10547-0.

KNIPLING, E. B., 1970. Physical and physiological basis for the reflectance of visible and near-infrared radiation from vegetation. Remote Sensing of Environment, 1(3):155–159, doi:10.1016/S0034-4257(70)80021-9.

MARTÍNEZ-LÜSCHER, J., PLANK, C. M., BRILLANTE, L., COOPER, M. L., SMITH, R. J., AL RWAHNIH, M., YU, R., OBERHOLSTER, A., GIRARDELLO, R., KURTURAL, S. K., 2019. Grapevine red blotch virus may reduce carbon translocation leading to impaired grape berry ripening. Journal of Agricultural and Food Chemistry 2019 67 (9), 2437-2448, doi: 10.1021/acs.jafc.8b05555

RICKETTS, K. D., GÓMEZ, M. I., FUCHS, M. F., MARTINSON, T. E., SMITH,R. J., COOPER, M. L., MOYER, M. M., WISE, A., 2017. Mitigating the economic impact of grapevine red blotch: Optimizing disease management strategies in U.S. Vineyards. American Journal of Enology and Viticulture, 68:127–135, , doi: 10.5344/ajev.2016.16009.

SAVITZKY, ABRAHAM., GOLAY, M. J. E., 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry, 36(8), 1627–1639. doi: 10.1021/ac60214a047

TERENTEV, A., DOLZHENKO, V., FEDOTOV, A., EREMENKO, D., 2022. Current State of Hyperspectral Remote Sensing for Early Plant Disease Detection: A Review. In Sensors (Vol. 22, Issue 3). MDPI. doi: 10.3390/s22030757

Table 1: Accuracies (in %) of the different machine learning models and their ensemble and stack for each dataset. The best results are highlighted in bold.

	Model			Ensemble		Stack	
	RF	*PLS-DA*	*SVM*	*PLS-DA_SVM*	*PLS-DA_SVM_RF*	*PLS-DA_RF*	*PLS-DA_SVM*
All data	61.2	67.5	69.2	**69.5**	69.0	56.5	53.1
Early season	49.9	48.2	55.3	**61.0**	58.4	53.7	51.6
Late season	62.3	72.2	**74.5**	**74.5**	74.4	58.4	53.4

Table 2: Confusion matrices obtained using PLS-DA_SVM ensemble model (results are in %)

		All data		***Early season***		***Late season***	
Observed	**Non-Infected**	80	20	63	37	83	17
	Infected	44	56	47	53	35	65
		Non-Infected	**Infected**	**Non-Infected**	**Infected**	**Non-Infected**	**Infected**
		Predicted					