Italian Journal of Agrometeorology
Rivista Italiana di Agrometeorologia

Firenze University Press
www.fupress.com/ijam

OPEN ACCESS

Citation: I. Volpi, D. Guidotti, M. Mammini, S. Marchi (2021) Predicting symptoms of downy mildew, powdery mildew, and gray mold diseases of grapevine through machine learning. *Italian Journal of Agrometeorology* (2): 57-69. doi: 10.36253/ijam-1131

Received: November 10, 2020

Accepted: August 14, 2021

Published: December 27, 2021

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Competing Interests: The Author(s) declare(s) no conflict of interest.

Predicting symptoms of downy mildew, powdery mildew, and gray mold diseases of grapevine through machine learning

Iride Volpi, Diego Guidotti, Michele Mammini, Susanna Marchi

AEDIT s.r.l., Pontedera, Pisa, Italy
E-mail: volpi@aedit.it; guidotti@aedit.it; mammini@aedit.it; marchi@aedit.it

Abstract. Downy mildew, powdery mildew, and gray mold are major diseases of grapevine with a strong negative impact on fruit yield and fruit quality. These diseases are controlled by the application of chemicals, which may cause undesirable effects on the environment and on human health. Thus, monitoring and forecasting crop disease is essential to support integrated pest management (IPM) measures. In this study, two tree-based machine learning (ML) algorithms, random forest and C5.0, were compared to test their capability to predict the appearance of symptoms of grapevine diseases, considering meteorological conditions, spatial indices, the number of crop protection treatments and the frequency of monitoring days in which symptoms were recorded in the previous year. Data collected in Tuscany region (Italy), on the presence of symptoms on grapevine, from 2006 to 2017 were divided with an 80/20 proportion in training and test set, data collected in 2018 and 2019 were tested as independent years for downy mildew and powdery mildew. The frequency of symptoms in the previous year and the cumulative precipitation from April to seven days before the monitoring day were the most important variables among those considered in the analysis for predicting the occurrence of disease symptoms. The best performance in predicting the presence of symptoms of the three diseases was obtained with the algorithm C5.0 by applying (i) a technique to deal with imbalanced dataset (i.e., symptoms were detected in the minority of observations) and (ii) an optimized cut-off for predictions. The balanced accuracy achieved in the test set was 0.8 for downy mildew, 0.7 for powdery mildew and 0.9 for gray mold. The application of the models for downy mildew and powdery mildew in the two independent years (2018 and 2019) achieved a lower balanced accuracy, around 0.7 for both the diseases. Machine learning models were able to select the best predictors and to unravel the complex relationships among geographic indices, bioclimatic indices, protection treatments and the frequency of symptoms in the previous year.

Keywords: agrometeorology, ERA5, grape, IPM models, monitoring networks.

1. INTRODUCTION

Downy mildew, powdery mildew, and gray mold are major diseases of grapevine (*Vitis vinifera* L.), affecting leaves and fruits and causing yield loss and quality decrease of must and wine. Downy mildew is caused by

the Oomycete *Plasmopara viticola* (Berk. & Curt.) Berl. & de Toni, with sexual spores determining primary infections and asexual spores causing secondary infections (Gessler et al., 2011). This pathogen infects leaves, shoots, and bunches, damaging up to 75% of the crop in one season when no treatments are applied (Buonassisi et al., 2017), thus leading to great economic losses. Powdery mildew is caused by *Erisyphe necato*r Schwein., a polycyclic disease with two distinct phases: primary infections are caused by sexual spores (ascospores) and secondary infections are determined by asexual spores (conidia) (Gadoury and Pearson, 1988), on all green tissues of grapevines, mainly leaves and berries (Gadoury et al., 2001; Caffi et al., 2011). *Botrytis cinerea* Pers. is the causal agent of gray mold and in grapevine infects all green tissues, particularly ripening berries, with different infection pathways for conidia (inflorescences, young clusters and ripening berries) and mycelium (berry-to-berry) (Elmer et al., 2007).

Because these pathogens may cause severe symptoms on grapevines at the beginning of infection, control strategies have focused on early treatments, even in integrated pest management (IPM), as prevention to stop the pathogen outbreak before its establishment. Applying fungicide treatments during the growing season remains the most common practice to control these diseases, from early spring onward, with differences between years due to weather conditions and to the geographic location of the vineyard (Chen et al., 2020; Lu et al., 2020; Molitor et al., 2016). However, concerns about the negative impact of chemicals on environmental and human health have resulted in restrictions to regulate fungicide use, such as the EU directives (i.e., Directive 1107/2009/EU) (Valdés-Gómez et al., 2017). European Commission currently enforces national action plans for pesticide reduction, encouraging the use of monitoring networks (Directive 128/2009/EC), forecasting models, and dissemination tools to share this information among growers and technicians (Pertot et al., 2017). Therefore, a reliable monitoring and forecasting system is essential for deriving prediction indices in support of sustainable protection measures (e.g., Marchi et al., 2016).

To this aim, various weather-driven models, either mechanistic (Rossi et al., 2008; Caffi et al., 2011; Legler et al., 2011; Gonzales et al., 2015) or empirical (Orlandini et al., 1993; Rodríguez-Rajo et al., 2010; Hill et al., 2019), have been developed for predicting grapevine diseases and assisting farmers in decision-making for crop protection. Decision support systems (DSSs), based on predictive models that use weather data and infection information, may provide this service to farmers (Rossi et al., 2014; Pertot et al., 2017). In particular, DSSs may help determine the time window for fungicide application to optimize their effects and to reduce the number of interventions during the growing season. Nevertheless, currently available models are mainly focused on predicting the risk of the outbreak, rather than the pressure of the disease. This approach may cause unnecessary fungicide applications and the use of untargeted chemical compounds, in turn contravening control regulations based on the maximum number of treatments allowed for each season (mandatory in IPM).

Since these three diseases are strongly influenced by seasonal weather conditions, albeit with different pathways among vectors, varying annually and driven by composite interactions between the disease agent and the host plant (growth stage and grapevine cultivar), models that provide information on infection risk need to combine numerous weather variables, crop parameters, and disease traits. Increasing computing power is providing the means to capture and process abundant data, and to reveal associations among variables that describe the weather-pathogen-host interactions. In particular, machine learning (ML) techniques allow considering a large number of variables, integrating diverse data sources in close real time, in order to assess the interactions among disease agent, host plant, and climatic variability, before visible symptoms are present, with the aim of ensuring effective and sustainable fungicide management (Lee et al., 2019; Sperschneider, 2019).

The potential of statistical models and ML algorithms to predict the occurrence of grapevine diseases has been rarely assessed (Chen et al., 2020). Here, we investigated the ability of ML algorithms to clarify the occurrence of symptoms of these three important diseases of grapevine based on prevailing weather conditions both within and between locations and years, generating temporal- and spatial-explicit projections of the infections. These models were implemented using as inputs: bioclimatic and geographic indices, the frequency of monitored symptoms in the previous year, and the number of crop protection treatments during the growing season. The aim of the study was to calculate the overall probability of symptom appearance at field scale, using the ML approach and the area-wide IPM monitoring network of Regione Toscana, providing farmers with a tool able to address timely and accurate grapevine disease forecasting.

2. MATERIALS AND METHODS

2.1 Monitoring grapevine diseases

Data on disease symptoms were obtained from Agroambiente.info (http://www.agroambiente.info/), the

agricultural and environmental portal of Phytosanitary Service of Regione Toscana (Italy). Agroambiente.info stores data deriving from an area-wide IPM monitoring network, which covers most of the wine production area of Tuscany. Sampling is carried out weekly by trained field technicians, from the leaf development stage (mid-end of April) to harvest (mid-end of September), in a variable number of vineyards through years (112-179). In each vineyard, date and presence of symptoms of downy mildew, powdery mildew and gray mold are recorded inspecting leaves and/or bunches. In addition, the date of treatments is reported, as well as the active substance (maximum two active substances for each treatment), among those allowed by "Integrated Production Regulation" of Regione Toscana. A simplified index of disease severity for each disease is documented during the monitoring activity, though it was not used for the ML exercise. Considering that different cultivars may show variable susceptibilities to the three diseases, the monitoring network focuses only on the cv Sangiovese, which is the most widespread and important for Tuscany denomination of controlled origin red wines. A numeric identification code ("farm ID") is assigned to each of the selected vineyards.

In this study, we considered data from 2006 to 2019, excluding 2011 since no data was available for that year in the regional database. In each dataset of the three diseases, the observations were classified as "inf" or "no", according to the presence or absence of disease symptoms, respectively. Observations were classified as "inf" when symptoms were present on leaves and/or on bunches.

2.2 Variables associated with grapevine diseases

Variables were calculated for each vineyard and each disease to be used as features for the ML models. The set of variables included: bioclimatic indices, geographical indices, indices indicating the number of phytosanitary treatments applied, an index referring to the frequency of the presence of infection in the previous year, and the day-of-year (doy) (Tab. 1).

The package 'raster' (Hijmans, 2019a) of the R environment (R Core Team, 2020), was used to extract for each vineyard from the raster files of the Tuscany region: (i) the Euclidean distance from the sea (dis_sea) in m and (ii) the elevation above sea level (m), obtained from the Digital Elevation Model (dem).

Meteorological data were downloaded from the open access ERA5-Land dataset, the latest generation of ECMWF atmospheric reanalysis, which provides hourly data from 1981 to 2-3 months before present in a fixed grid and with a native resolution of 9 km (Copernicus Climate Change Service, 2017).

ERA5-Land dataset was selected over others (e.g., ERA5, ERA-Interim) because of its higher spatial resolution and its improved correlation with in situ measurements, especially concerning the water cycle (Muñoz-Sabater et al., 2021). Using reanalysis meteorological data, as ERA5-Land dataset, for modelling has the main advantages of providing data with a better temporal and spatial coverage with respect to the data collected with real weather stations that do not have a uniform spatial and temporal coverage and may be subjected to breaks (Padulano et a., 2021). Indeed, concerning the density of the weather monitoring network of Tuscany Region, the distance from each vineyard to its nearest station ranged between 120 m to 27000 m with an average value of 6640 m.

Meteorological data from the ERA5-Land dataset used to calculate daily maximum, minimum and average air temperature (°C) and daily precipitation (mm) for the period from 2006 to 2019 were: "2-m temperature", defined as the hourly temperature of air at 2 m above the ground, sea or inland waters, and "total precipitation", defined as accumulated liquid and frozen water, including rain and snow, that falls to the Earth's surface. The distance between each ERA5 grid-box and each georeferenced monitoring site was calculated through the R package 'geosphere' (Hijmans, 2019 b), with the aim of associating each sampling site with an ERA5 grid-box.

Bioclimatic indices were calculated starting from daily data on air temperature and precipitation, considering three different periods: (i) from November to January for the indices describing the weather conditions during overwintering (average of minimum, maximum and mean daily temperature), (ii) from November to March for monthly mean air temperature and cumulative precipitation, (iii) from April to October (monitoring period) for the bioclimatic indices describing the weather conditions in the interval from 14 to 7 days before the monitoring day or during the 7 days before the monitoring day. We considered these two time steps to identify the environmental conditions of the period during which the pathogen penetration into the host tissues was most probable (avg_14_7, avg_max_14_7, avg_min_14_7, cum_rain_14_7) (Chen et al., 2020; Carisse et al., 2009; Barka et al., 2002).

The phytosanitary treatments were included in the ML models as counts of the applications carried out in each vineyard from the beginning of the vegetative season, considering three periods: (i) cumulative number of treatments carried out until 14 days before the monitor-

Table 1. Set of variables associated with the three grapevine diseases.

Indices	Period	Description	Unit	Disease
w_mean_avg	November - January	Average of mean daily temperatures	°C	downy mildew, powdery mildew, gray mold
w_min_avg	November - January	Average of minimum daily temperatures	°C	downy mildew, powdery mildew, gray mold
w_rain_cum	November - January	Cumulative precipitation	mm	downy mildew, powdery mildew, gray mold
tavg_11	November	Average of mean daily temperatures	°C	downy mildew, powdery mildew, gray mold
tavg_12	December	Average of mean daily temperatures	°C	downy mildew, powdery mildew, gray mold
tavg_1	January	Average of mean daily temperatures	°C	downy mildew, powdery mildew, gray mold
tavg_2	February	Average of mean daily temperatures	°C	downy mildew, powdery mildew, gray mold
tavg_3	March	Average of mean daily temperatures	°C	downy mildew, powdery mildew, gray mold
psum_11	November	Cumulative precipitation	mm	downy mildew, powdery mildew, gray mold
psum_12	December	Cumulative precipitation	mm	downy mildew, powdery mildew, gray mold
psum_1	January	Cumulative precipitation	mm	downy mildew, powdery mildew, gray mold
psum_2	February	Cumulative precipitation	mm	downy mildew, powdery mildew, gray mold
psum_3	March	Cumulative precipitation	mm	downy mildew, powdery mildew, gray mold
avg_14_7	April - October	Average of mean daily temperatures from 14 days to 7 days before the monitoring day	°C	downy mildew, powdery mildew, gray mold
avg_max_14_7	April - October	Average of max daily temperatures from 14 days to 7 days before the monitoring day	°C	downy mildew, powdery mildew, gray mold
avg_min_14_7	April - October	Average of min daily temperatures from 14 days to 7 days before the monitoring day	°C	downy mildew, powdery mildew, gray mold
cum_ rain_14_7	April - October	Cumulative precipitation from 14 days to 7 days before the monitoring day	mm	downy mildew, powdery mildew, gray mold
cum_rain_7	April - October	Cumulative precipitation from April to 7 days before the monitoring day	mm	downy mildew, powdery mildew, gray mold
gdd_apr_7	April - October	Cumulative degree day (mean air temperature) from April to 7 days before the monitoring day, with a lower threshold of 10 °C	°C	downy mildew, gray mold
gdd_jan_7	January - October	Cumulative degree day (mean air temperature) from January to 7 days before the monitoring day, with a lower threshold of 10 °C	°C	downy mildew, gray mold
gdd_7	April - October	Cumulative degree day from April to 7 days before the monitoring day, with a lower threshold of 6 °C and an upper threshold of 30.5 °C [1, 2]	°C	powdery mildew
dem100	n.a.	Elevation a.s.l.	m	downy mildew, powdery mildew, gray mold
dis_sea	n.a.	Euclidean distance from sea	m	downy mildew, powdery mildew, gray mold
count_tr_0_7	n.a.	Number of treatments in the 7 days before the monitoring day	n°	downy mildew, powdery mildew, gray mold
count_tr_7_14	n.a.	Number of treatments from 14 days to 7 days before the monitoring day	n°	downy mildew, powdery mildew, gray mold
count_tr_14	n.a.	Cumulative number of treatments 14 days before the monitoring day	n°	downy mildew, powdery mildew, gray mold
perc_inf	n.a.	Percentage of the observation in which was reported the presence of symptoms in the previous year	%	downy mildew, powdery mildew, gray mold
doy	n.a.	Day of the year	n.a.	downy mildew, powdery mildew, gray mold

[1] Allen (1976).
[2] Carisse et al. (2009).

ing day (count_tr_14), (ii) number of treatments carried out from 14 to 7 days before the monitoring day (count_ tr_7_14), and (iii) number of treatments carried out in the 7 days before the monitoring day (count_tr_0_7).

In addition, the models included as a variable the frequency of monitoring days in which symptoms were recorded in the previous year, to consider the potential presence of the pathogens overwintering in the vine-

yard. The latter variable was calculated as the percentage of observations in which the presence of symptoms was observed in each year and in each vineyard, and it was assigned to the following monitoring year (perc_inf).

2.3 Data analysis

The three datasets on the symptoms observed of downy mildew, powdery mildew and gray mold covered a period from 2006 to 2019, excluding 2011 since no data were available for that year. The datasets had a different number of observations: 18857 for downy mildew, 14848 for powdery mildew, and 4960 for gray mold.

The dataset of each disease was partitioned with the aim of training and testing the ML models. The two datasets, downy mildew and powdery mildew, were divided in one training set and two test sets. In particular, data collected in the period from 2006 to 2017 were partitioned with an 80/20 proportion in "training" and "test 1", respectively. The partition was carried out using the R package 'healthcareai' (Thatcher et al., 2020), considering the group "farm ID x year", which allowed ensuring that observations from each vineyard in each year were not contained in both training set and test set. A further test ("test 2") included data collected in 2018 and 2019 to evaluate the performance of the model on two independent years. Since less data were available in comparison with the other two diseases, the dataset on gray mold infection (2006-2019) was partitioned only in training set and test set with an 80/20 proportion in "training" and "test 20%", considering the group "farm ID x year".

The class "inf" was present in a different percentage of the total observations for the three diseases: 37% in training set, 35% in "test 1", and 58% in "test 2" for downy mildew; (ii) 16% in training set, 15% in "test 1", and 28% in "test 2" for powdery mildew; (iii) 10% in training set and 8% in test set for gray mold.

Spearman's correlation among the variables associated with each disease was calculated with the R package 'Hmisc' (Harrell, 2019), to remove redundant features, highlighting variables that were highly correlated. Thus, in the case that the Spearman's correlation coefficient between two variables was higher than 0.9 (absolute value), we selected the one with the highest importance, using a filter approach based on the Receiver Operating Characteristic (ROC) curve analysis, a plot of true positive rate (TPR) versus false positive rate (FPR) at various threshold settings.

Machine learning models selected for comparison were: (i) Random forest (RF), based on several decision trees, which operates as an ensemble to produce an output with low bias and lower variance than each single tree; and (ii) C5.0 based on single binary decision tree or a collection of rules with a boosted procedure. Both algorithms were tree-based models, being able to handle complex non-linear relationships and outperforming other ML algorithms in earth science and ecology applications (Thessen, 2016).

The train function of the R package 'caret' (Kuhn, 2020) was used to train and tune the two models, RF and C5.0, by means of the ROC metric, using a 10-fold cross-validation clustered by the grouping factor "farm ID x year". Thus, while running the train function of 'caret', through the 10-fold cross-validation the training set is partitioned in 10 equal size subsamples of which 9 subsamples are used to train the model and a single subsample is retained as the validation data for testing the performance of the model with the aim of tuning the model parameters.

The models were evaluated using a confusion matrix among observed and predicted classes (Tab. 2) and a set of performance metrics on the training set and test set (Tab. 3).

The best performing algorithm (evaluated on training set and test set), was further optimized: (i) applying subsampling techniques for class imbalance during the training with the R package 'caret', and (ii) selecting the cut-off, to be applied on the probability outputs of the models for classification, which optimized the informedness (*Specificity*+*Sensitivity*-1) of the trained model, using the R package 'MLeval' (John, 2020).

For the best performing algorithm, the importance of variables in the modelling mechanism was extracted using the function varImp() of the R package "caret".

3. RESULTS

The correlation analysis allowed removing highly correlated variables associated with each disease. Variables removed were: avg_14_7, avg_max_14_7, gdd_apr_7, gdd_jan_7, tavg_12, w_min_avg for downy mildew; avg_14_7, avg_max_14_7, doy, w_min_avg for powdery mildew and avg_14_7, avg_max_14_7, doy, gdd_apr_7, tavg_12, w_mean_avg, w_min_avg, w_rain_cum for gray mold.

The results of the cross-validation on the training set highlighted ROC-AUC values higher than 0.8 for both RF and C5.0 (Tab. 4).

AUC-PR was higher than 0.6 for downy mildew and gray mold, while it was around 0.6 for powdery mildew. The sensitivity was around 0.7 for downy mildew, while it was around 0.4 for powdery mildew and gray mold.

Table 2. Confusion matrix based on the number of observed and predicted classes. For each disease the class "inf" indicated the presence of symptoms on leaves or on grapes.

Predicted symptoms	Observed symptoms	
	inf	no
inf	True Positive (TP)	False Positive (FP)
no	False Negative (FN)	True Negative (TN)

Table 3. List of metrics used to evaluate the performance of the classification algorithms.

Evaluation metric	Data	Description
AUC-ROC	Training	Area under the ROC (Receiver operating characteristic) curve
AUC-PR	Training	Area under the Precision-Recall curve
Sensitivity = Recall	Training/Test	$\frac{TP}{TP+FN}$
Specificity	Training/Test	$\frac{TN}{TN+FP}$
Positive Predictive Value (PPV) = Precision	Test	$\frac{TP}{TP+FP}$
Negative Predictive Value (NPV)	Test	$\frac{TN}{TN+FN}$
F1	Test	$\frac{2 \times Precision \times Recall}{Precision + Recall}$
Accuracy	Test	$\frac{TN+TP}{TN+TP+FN+FP}$
Balanced accuracy	Test	$\frac{Sensitivity + Specificity}{2}$

Table 4. Performance of the two machine learning algorithms (RF and C5.0), tuned through the 10-fold cross-validation, on the training sets of the three diseases.

	RF			C5.0		
	Downy mildew	Powdery mildew	Gray mold	Downy mildew	Powdery mildew	Gray mold
AUC-ROC	0.85	0.87	0.91	0.84	0.87	0.90
AUC-PR	0.77	0.60	0.69	0.75	0.59	0.66
Sensitivity	0.66	0.43	0.45	0.65	0.50	0.48
Specificity	0.84	0.96	0.99	0.85	0.94	0.98

Table 5. Results of RF and C5.0 in predicting the presence of symptoms of the three diseases on the test sets "test 1".

	RF			C5.0		
	Downy mildew	Powdery mildew	Gray mold	Downy mildew	Powdery mildew	Gray mold
TP	701	129	44	705	158	52
FP	256	75	9	245	92	9
TN	1573	2024	967	1584	2007	967
FN	285	243	44	281	214	36
Sensitivity	0.71	0.35	0.50	0.71	0.42	0.59
Specificity	0.86	0.94	0.99	0.87	0.96	0.99
PPV	0.73	0.63	0.83	0.74	0.63	0.85
NPV	0.85	0.89	0.96	0.85	0.90	0.96
F1	0.72	0.45	0.62	0.73	0.51	0.70
Accuracy	0.81	0.87	0.95	0.81	0.88	0.96
Balanced accuracy	0.78	0.66	0.74	0.79	0.69	0.79

The specificity was around 0.8 for downy mildew, while it was higher than 0.9 for the other two diseases. The two algorithms performed similarly on the training set, with slightly better results for RF.

For all the three diseases, the algorithm C5.0 performed better than RF on the test set ("test 1"), reporting in particular a higher sensitivity and a higher balanced accuracy (Tab. 5).

The C5.0 algorithm predicted the presence of symptoms of downy mildew with a balanced accuracy of 78%. The overall predicted "inf" were correct for 71% of the cases, whereas the percentage of correctly predicted "no" on the total prediction of no symptoms was 87%. The percentage of cases in which "inf" was correctly identified was 74%, while for "no" it was 85%. The presence of symptoms of powdery mildew was predicted by C5.0 with a balanced accuracy of 69%. The overall predicted "inf" were correct for 42% of the cases, whereas the percentage of correctly predicted "no" on the total prediction of no symptoms was 96%. The percentage of cases in which "inf" was correctly identified was 63%, while for "no" it was 90%. The presence of symptoms of gray mold was predicted by C5.0 with a balanced accuracy of 79%. The overall predicted "inf" were correct for 59% of the cases, whereas the percentage of correctly predicted "no" on the total prediction of no symptoms was 99%. The percentage of cases in which "inf" was correctly identified was 85%, while for "no" it was 96%.

The subsampling technique "down" was selected as the best according to the performance on the test set of the three diseases (Tab. S1). Applying the subsampling technique to the algorithm C5.0 increased the percentage of cases in which "inf" was correctly identified and the balanced accuracy for all the three diseases (Tab. 6). Moreover, the informedness of the C5.0 algorithm with down-sampling was the highest when applying a cut-off equal to: (i) 0.46 for the prediction of downy mildew and

Table 6. Results of C5.0 on the test sets "test 1" applying: (i) down-sampling during training (C5.0 'down'), and (ii) the optimized cut-off for classification on the probability outputs of C5.0 'down'.

	C5.0 'down'			C5.0 'down' & cut-off opt.		
	Downy mildew	Powdery mildew	Gray mold	Downy mildew	Powdery mildew	Gray mold
TP	758	236	77	791	254	84
FP	378	252	64	446	329	137
TN	1451	1847	912	1383	1770	839
FN	228	136	11	195	118	4
Sensitivity	0.77	0.63	0.87	0.80	0.68	0.95
Specificity	0.79	0.88	0.93	0.76	0.84	0.86
PPV	0.67	0.48	0.55	0.64	0.46	0.38
NPV	0.86	0.93	0.99	0.88	0.86	0.99
F1	0.71	0.55	0.67	0.71	0.56	0.54
Accuracy	0.78	0.84	0.93	0.77	0.69	0.87
Balanced accuracy	0.78	0.76	0.90	0.78	0.70	0.91

Table 7. Results of the application of the optimized cut-off for classification on the probability outputs of C5.0 'down' on both the overall "test 2" and on the two years considered separately (2018 and 2019).

	C5.0 'down' & cut-off opt.					
	Downy mildew			Powdery mildew		
	test 2	2018	2019	test 2	2018	2019
TP	1662	1126	536	304	168	136
FP	385	47	338	201	89	112
TN	1503	855	648	1409	824	585
FN	964	721	243	313	221	92
Sensitivity	0.63	0.61	0.69	0.49	0.43	0.60
Specificity	0.8	0.95	0.66	0.87	0.9	0.84
PPV	0.81	0.96	0.61	0.6	0.65	0.55
NPV	0.61	0.54	0.73	0.82	0.79	0.86
F1	0.71	0.75	0.65	0.54	0.52	0.57
Accuracy	0.7	0.72	0.67	0.77	0.76	0.78
Balanced accuracy	0.71	0.78	0.67	0.68	0.67	0.72

powdery mildew symptoms and (ii) 0.36 for the prediction of gray mold. The application of the optimized cut-off on the results of the C5.0 algorithm with down-sampling improved the sensitivity, being 0.80 for downy mildew, 0.68 for powdery mildew and 0.95 for gray mold.

Results on "test 2", highlighted a prediction accuracy around 0.7 for both the symptoms of downy mildew and powdery mildew, in both 2018 and 2019 (Tab. 7).

Table 8. Importance of variables in the modelling process of the algorithm C5.0 'down' for the three diseases. The importance is calculated as the percentage of splits associated with each predictor (metric = 'splits').

C5.0 'down'					
Downy mildew		Powdery mildew		Gray mold	
perc_inf	12.3	perc_inf	7.3	perc_inf	11.2
doy	11.1	dis_sea	7.2	cum_rain_7	7.9
cum_rain_7	7.6	cum_rain_7	6.6	gdd_jan_7	7.8
dem100	6.2	gdd_7	6.5	cum_rain_14_7	7.3
psum_3	5.9	dem100	6.5	psum_12	7.3
psum_2	5.9	count_tr_14	5.7	avg_min_14_7	6.6
dis_sea	5.0	avg_min_14_7	5.7	psum_2	6.6
psum_12	4.3	psum_2	5.4	tavg_11	6.2
count_tr_14	4.3	psum_12	5	dis_sea	5
tavg_2	3.9	psum_3	4.5	psum_3	4.6
psum_1	3.8	tavg_11	4.3	count_tr_0_7	4.4
w_rain_cum	3.8	tavg_3	4.1	psum_11	4.3
psum_11	3.7	psum_1	4	dem100	4.2
tavg_1	3.6	psum_11	3.9	psum_1	3.5
tavg_11	3.3	count_tr_0_7	3.8	tavg_1	3
count_tr_0_7	3.0	w_mean_avg	3.8	count_tr_7_14	3
avg_min_14_7	2.9	w_rain_cum	3.8	tavg_3	2.9
cum_rain_14_7	2.7	cum_rain_14_7	3.2	tavg_2	2.3
tavg_3	2.4	tavg_1	3.2	count_tr_14	2.3
count_tr_7_14	2.4	tavg_2	3.1		
w_mean_avg	1.9	count_tr_7_14	2.2		

The importance of the variables in the modelling process for the three diseases is reported in Tab. 8.

Around 50% of the splits were associated with the first six most important variables for downy mildew, namely: the percentage of observations of the year before in which symptoms were present, day of year, the cumulative precipitation from April to 7 days before the monitoring day, the elevation, and the precipitation of March and February.

For powdery mildew, the first eight most important variables covered around 50% of the splits, being: the percentage of the observation of the previous year in which symptoms appeared, the distance from sea, the cumulative precipitation from April until 7 days before the monitoring day, cumulative degree day (gdd) from April until 7 days before the monitoring day, the elevation, the count of the treatments carried out until 14 days before the monitoring day, the average minimum temperature between 14 and 7 days before the monitoring day, the precipitation of February.

For gray mold, the first six most important variables were associated to around 50% of the splits: the

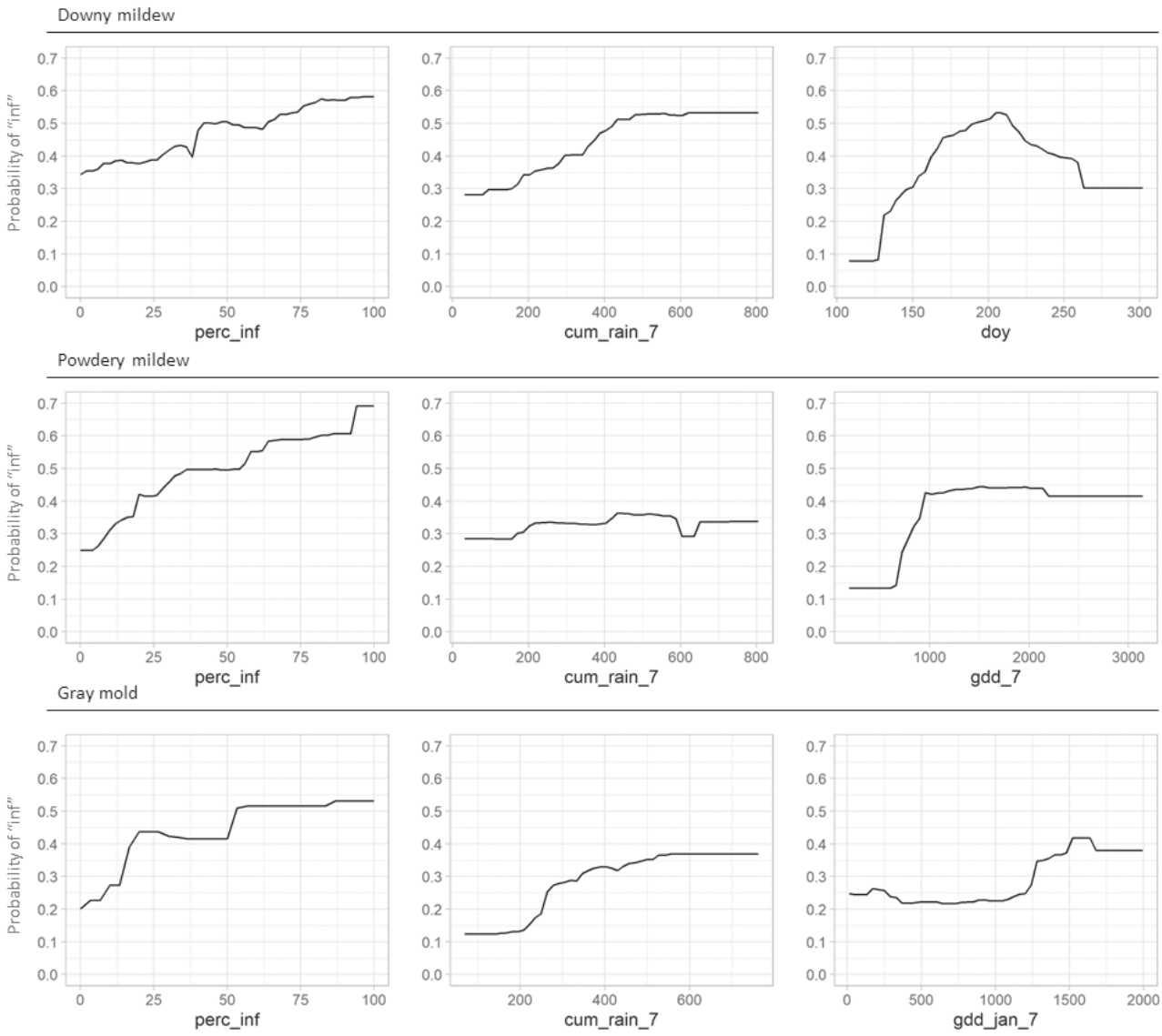

Fig. 1. Partial dependence plots for the marginal effect on the prediction of the class "inf" of the variables: (i) perc_inf, cum_rain_7 and doy for downy mildew; (ii) perc_inf, cum_rain_7 and gdd_7 for powdery mildew and (iii) perc_inf, cum_rain_7 and gdd_jan_7 for gray mold.

percentage of the observation of the previous year in which symptoms appeared, the cumulative precipitation from April until 7 days before the monitoring day and between 14 and 7 days before the monitoring day, the precipitation of December, the average minimum temperature between 14 and 7 days before the monitoring day.

Among the three models, common variables on the top of the list were: perc_inf and cum_rain_7. Similar variables, such as doy and gdd, were ranked within the top variables for all the three models. Partial dependence plots (pdp) (Fig. 1) represent the marginal effect of the latter variables on the probability of predicting the presence of symptoms for the three diseases. In particular, with increasing values of perc_inf and cum_rain_7, the probability of predicting "inf" increased for powdery mildew, until about 500 mm for cum_rain_7. Concerning downy mildew, with increasing values of doy, the probability of the class "inf" increased, until about doy 200, while decreasing after this threshold. For powdery mildew, the probability of the class "inf" markedly increased with ggd_7, until 1000, while the probability of the class "inf" for gray mold increased with ggd_jan_7, between about 1200 and 1600.

4. DISCUSSION

Results of the application of ML algorithms, trained on historical data, for the prediction of the appearance of symptoms of downy mildew, powdery mildew, and gray mold in grapevine, demonstrated a better performance of the algorithm C5.0 in comparison with the RF, in the test set ("test 1"), for all the three diseases. Similar results were found by Volpi et al. (2020), who applied ML algorithms for predicting the probability of infestation by *Bactrocera oleae* on olive trees, founding that C5.0 had a higher ROC compared to k-nearest neighbors (k-NN), Classification and Regression Trees (CART), Random Forest (RF) and Neural Network (NN).

The three datasets on grapevine diseases were unbalanced, since the observations in which the symptoms of diseases were recorded were a minority of the total observations, in particular for powdery mildew and gray mold (<20%). Class imbalance problems may lead to partial behaviour of the classifier towards the majority class and sampling methods are most commonly applied to balance the class distribution of the training data (Kaur et al., 2019). Moreover, the output of C5.0 classification for each observation is a probability between 0 and 1 of being classified as "inf" or "no" and the standard cut-off applied for classification is 0.5, which is not the most appropriate for imbalanced datasets (Zou et al., 2016). Therefore, the application of both the down-sampling technique and a cut-off for classification, optimized to improve the informedness of the model, increased the sensitivity of the model, thus increasing the amount of true positives and decreasing the amount of false negatives, which has a high cost for the prediction of plant diseases.

The final models achieved a good performance in predicting the presence of symptoms of the three diseases on "test 1", with a balanced accuracy of 0.8 for downy mildew, 0.7 for powdery mildew and 0.9 for gray mold, highlighting a lower occurrence of wrong classifications for the gray mold model.

The application of the models for downy mildew and powdery mildew on the two independent years (2018 and 2019) achieved a lower balanced accuracy than on "test 1", being, however, around 0.7 for the two diseases. This slightly lower performance of the ML model on unseen data may be due to the known bias-variance tradeoff of ML models, being complex models more subjected to high variance (Abu-Mostafa et al., 2012).

Differently to other ML techniques (i.e., Bayesian network) in which the causal relationships among the variables are linked to previous knowledge (Lu et al., 2020), the effect of the variables on the prediction in tree-based ML models is entirely data-driven. However, it is possible to interpret the C5.0 model by exploring the importance of variables in the modelling mechanism and the effect of variables on the prediction.

Indeed, it was possible to highlight, for all the diseases, a higher frequency in the top-ranking positions, in terms of importance, of indices related to precipitation rather than to air temperature. In particular, the cumulative precipitation from the beginning of April to 7 days before the day of observation was among the most important variables for the three diseases.

For downy mildew and gray mold, the probability of infection increased with increasing values of cumulative precipitation (approximately until 500 mm); while, for powdery mildew, the relationship was less clear. In particular, downy mildew is typically diffused in viticultural areas characterized by temperate climate and frequent precipitation during spring and summer (Lafon and Clerjeau, 1988), and precipitation was reported as a key driver for both primary and secondary infections (Rossi et al., 2008). Climate conditions at the end of spring, particularly precipitation, were found to be decisive for the development of downy mildew symptoms (Chen et al., 2020). Precipitation events have a positive effect in spreading the infection of powdery mildew (dispersing cleistothecia and releasing ascospores) and, though free water is detrimental to conidial germination, in rainy seasons the environmental conditions become favourable for the infection due to mild temperatures, limited direct sunlight, and high humidity (Gadoury et al., 2012). Furthermore, more severe gray mold epidemics were reported under wet growing seasons, since the wetness duration is a key factor for both the development of early season and late season infections (Ciliberti et al., 2015 a, b).

The frequency of symptoms observed in the previous year (the year before the one considered for ML application) was the most important variable in the modelling mechanism for the appearance of symptoms of the three diseases. In particular, the risk of symptom development in the current year increased with the occurrence of severe infection in the previous year. Severe infections may be a source of overwintering pathogens, potentially leading to new infections under optimal environmental conditions. Indeed, downy mildew is able to overwinter mainly on infected shoots, while powdery mildew in grapevine buds, and gray mold in grapevine debris (Pertot et al., 2017; Jaspers et al., 2013; Rügner et al., 2002).

Variables describing the progress of the season, such as doy for downy mildew and cumulative degree days for powdery mildew and gray mold, were among the most

important variables for predicting the development of disease symptoms. In particular, the probability of infection increased for downy mildew from April to about mid-July and then decreased. Previous studies reported that the progress of disease relates to the phenological development of grapevine (Molitor et al., 2016; Carmichael et al., 2018; Bove et al., 2020). In addition, ggd_7 was a key variable for predicting the occurrence of powdery mildew symptoms; Carisse et al. (2009) used this variable to predict the proportion of seasonal airborne inoculum.

However, even if gdd or doy were among the most important variable in the modelling mechanism, the use of a multivariate approach through ML algorithms with respect to a univariate cumulative GDD index is recognized to be more suited to model non-linear patterns and variable interactions often characterizing real-world ecological patterns (Yo et al., 2017).

The effect of the number of chemical treatments was more important for powdery mildew than for the other diseases, since only for powdery mildew an index indicating the frequency of treatments was among the top variables. Further studies are needed to evaluate new approaches to include the effect of treatments in modelling predictions, considering the type of chemical and the mechanism of action.

Results from this work highlighted that a ML algorithm trained on historical data, may be efficiently used to predict the appearance of symptoms of downy mildew, powdery mildew, and gray mold in grapevine, providing an innovative control tool, even in association with traditional models. The simplicity of the approach requires, however, the availability of symptom records, which is the monitoring of disease occurrence. Massive datasets of disease symptoms or pest attacks may allow not only regional-level analyses, like in the present study, but also the recognition of specific and localized risk factors, which take into account additional variables, conferring susceptibility or resistance to a given disease or pest. The ML algorithms can be implemented with additional weather data that are used in other models for disease prediction (Rossi et al., 2008; Chen et al., 2020). Yet, climatic inputs can be further enriched to forecast the occurrence of downy mildew, powdery mildew, and gray mold under different climate scenarios and assess the future trajectories of these diseases.

The integration of ML models in decision support systems also represents a practical application to plan the reduction of fungicide treatments. In particular, the use of ML is a promising approach to implement early warning systems, identifying periods when climatic conditions are favourable to promote disease development and alerting on symptoms that are associated with high risk of infection (Caffi et al., 2010; Pellegrini et al., 2010). As future activity, the integration of mechanistic and ML models (e.g., in Bayesian networks) could be tested to evaluate the effect of including previous knowledge in the modelling mechanism.

5 CONCLUSION

The application of ML algorithms, trained on historical data, was proved useful for the prediction of the appearance of symptoms of downy mildew, powdery mildew, and gray mold in grapevine. The grape disease monitoring network enabled the observation of a wide range of symptoms. This, in combination with ERA5-Land dataset allowed the development of early detection algorithms to support the implementation of IPM in viticulture. Compared to ground weather stations, ERA5 data had the advantage of providing information for locations that are not covered by traditional agrometeorological networks. Nevertheless, for becoming fully operative, this approach needs an efficient monitoring system at the landscape scale and intensive field surveys at the local scale.

ACKNOWLEDGEMENTS

We thank the Phytosanitary Service of Regione Toscana.

REFERENCES

Abu-Mostafa Y. S., Magdon-Ismail M., Lin H.T., 2012. Learning From Data. AMLBook.

Allen C., 1976. A modified sine wave method for calculating degree days. Environmental Entomology, 5: 388–396.

Barka E. A., Gognies S., Nowak J., Audran J. C., Belarbi A. 2002. Inhibitory effect of endophyte bacteria on *Botrytis cinerea* and its influence to promote the grapevine growth. Biological Control, 24: 135-142.

Bove F., Savary S., Willocquet L., Rossi V., 2020. Designing a modelling structure for the grapevine downy mildew pathosystem. European Journal of Plant Pathology, 158: 599-614.

Buonassisi D., Colombo M., Migliaro D., Dolzani C., Peressotti E., Mizzotti C., Velasco R., Masiero S., Perazzolli M., Vezzulli S., 2017. Breeding for grapevine downy mildew resistance: a review of "omics" approaches. Euphytica, 213:1–21.

Caffi T., Rossi V., Bugiani R., 2010. Evaluation of a warning system for controlling primary infections of grapevine downy mildew. Plant Disease, 94: 709-716.

Caffi T., Rossi, V., Legler S. E., Bugiani R., 2011. A mechanistic model simulating ascosporic infections by *Erysiphe necator*, the powdery mildew fungus of grapevine. Plant Pathology, 60: 522-531.

Carmichael P. C., Siyoum N., Jongman M., Korsten L., 2018. Prevalence of *Botrytis cinerea* at different phenological stages of table grapes grown in the northern region of South Africa. Scientia Horticulturae, 239: 57-63.

Carisse O., Bacon R., Lefebvre A., Lessard K., 2009. A degree-day model to initiate fungicide spray programs for management of grape powdery mildew [Erysiphe necator]. Canadian Journal of Plant Pathology, 31: 186–194.

Chen M., Brun F., Raynal M., Makowski D. 2020. Forecasting severe grape downy mildew attacks using machine learning. Plos One, 15(3), e0230254

Ciliberti N., Fermaud M., Languasco L., Rossi V., 2015 (a). Influence of fungal strain, temperature, and wetness duration on infection of grapevine inflorescences and young berry clusters by Botrytis cinerea. Phytopathology, 105: 325–333.

Ciliberti N., Fermaud M., Roudet J., Rossi V., 2015 (b). Environmental conditions affect Botrytis cinerea infection of mature grape berries more than the strain or transposon genotype. Phytopathology, 105: 1090–1096.

Copernicus Climate Change Service (C3S) (2017): ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store (CDS). https://cds.climate.copernicus.eu/cdsapp#!/home.

Elmer P.A.G., Michailides T.J., 2007. Epidemiology of Botrytis cinerea in orchard and vine crops. In: Elad Y, Williamson B, Tudzynski P, Delen N, editors. Botrytis: Biology, Pathology and Control. Springer Netherlands, pp. 243–272.

Gadoury D.M. and Pearson R. C., 1988. Initiation, development, dispersal, and survival of cleistothecia of Uncinula necator in New York vineyards. Phytopathology, 78:1413-1421.

Gadoury D.M., Seem R.C., Pearson R.C., Wilcox W.F., 2001. Effects of powdery mildew on vine growth, yield, and quality of Concord grapes. Plant Disease, 85:137–140.

Gadoury D.M., Cadle-Davidson L., Wilcox W.F., Dry I.B., Seem R.C., Milgroom M.G., 2012. Grapevine powdery mildew (*Erysiphe necator*): A fascinating system for the study of the biology, ecology and epidemiology of an obligate biotroph. Molecular Plant Pathology, 13: 1–16.

Gessler C., Pertot I., Perazzolli M., 2011. Plasmopara viticola: a review of knowledge on downy mildew of grapevine and effective disease management. Phytopathologia Mediterranea, 50:3–44.

González-Domínguez E. Caffi T., Ciliberti N. Rossi V. 2015. A mechanistic model of Botrytis cinerea on grapevines that includes weather, vine growth stage, and the main infection pathways. PLoS One, 10(10), e0140444.

Harrell F.E. Jr, 2019. Hmisc: Harrell Miscellaneous. R package version 4.2-0. https://CRAN.R-project.org/package=Hmisc.

Hijmans R.J., 2019 a. raster: Geographic Data Analysis and Modeling. R package version 2.8-19. https://CRAN.R-project.org/package=raster.

Hijmans R.J., 2019 b. geosphere: Spherical Trigonometry. R package version 1.5-10. https://CRAN.R-project.org/package=geosphere.

Hill G. N., Beresford R. M., Evans K. J., 2019. Automated analysis of aggregated datasets to identify climatic predictors of botrytis bunch rot in wine grapes. Phytopathology, 109: 84-95.

Jaspers M. V., Seyb A.M., Trought M.C.T., Balasubramaniam R., 2013. Overwintering grapevine debris as an important source of Botrytis cinerea inoculum. Plant Pathology, 62: 130–138.

John C.R., 2020. MLeval: Machine Learning Model Evaluation. R package version 0.3. https://cran.r-project.org/package=MLeval.

Kaur H., Pannu H.S., Malhi A.K., 2019. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM Computing Surveys, 52. Article 79.

Kuhn M., 2020. caret: Classification and Regression Training. R package version 6.0-85. https://CRAN.R-project.org/package=caret.

Lafon R., Clerjeau M., 1988. Downy mildew. In: Pearson, R.C., Goheen, A.C. (Eds.), Compendium of Grape Diseases. APS Press, St. Paul, Minnesota, USA, pp. 11–13.

Lee, D.S., Bae, Y.S., Byun, B.K., Lee, S., Park, J.K., Park, Y.S., 2019. Occurrence prediction of the citrus flatid planthopper (Metcalfa pruinosa (Say, 1830)) in South Korea using a random forest model. Forests, 10, 583.

Legler S.E., Caffi T., Rossi V., 2011. A non linear model for temperature-dependent development of Erysiphe necator chasmothecia on grapevine leaves. Plant Pathology, 61: 96-105.

Lu W., Newlands N. K., Carisse O., Atkinson D. E., Cannon A. J. 2020. Disease Risk Forecasting with

Bayesian Learning Networks: Application to Grape Powdery Mildew (Erysiphe necator) in Vineyards. Agronomy, 10(5), 622.

Marchi S., Guidotti D., Ricciolini M., Petacchi R., 2016. Towards understanding temporal and spatial dynamics of *Bactrocera oleae* (Rossi) infestations using decade-long agrometeorological time series. International Journal of Biometeorology, 60: 1681-1694.

Molitor D., Baus O., Hoffmann L., Beyer M., 2016. Meteorological conditions determine the thermal-temporal position of the annual Botrytis bunch rot epidemic on Vitis vinifera L. cv. Riesling grapes. Oeno One, 50: 231–244.

Muñoz-Sabater J., Dutra E., Agustí-Panareda A., Albergel C., Arduini G., Balsamo G., Boussetta S., Choulga M., Harrigan S., Hersbach H., Martens B., Miralles D., Piles M., Rodríguez-Fernández N., Zsoter E., Buontempo C., Thépaut J.N., 2021. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. Earth System Science Data Discussions [preprint], in review.

Orlandini S., Gozzini B., Rosa M., Egger E., Storchi P., Maracchi G., Miglietta F., 1993. PLASMO: a simulation model for control of Plasmopara viticola on grapevine 1. EPPO Bulletin, 23: 619-626.

Padulano R., Rianna G., Santini M., 2021. Datasets and approaches for the estimation of rainfall erosivity over Italy: A comprehensive comparison study and a new method. Journal of Hydrology: Regional Studies, 34: 100788.

Pellegrini A., Prodorutti D., Frizzi A., Gessler C., Pertot I., 2010. Development and evaluation of a warning model for the optimal use of copper in organic viticulture. Journal of Plant Pathology, 43-55.

Pertot I., Caffi T., Rossi V., Mugnai L., Hoffmann C., Grando M. S., Gary C., Lafond D., Duso C., Thiery D., Mazzoni V., Anfora G., 2017. A critical review of plant protection tools for reducing pesticide use on grapevine and new perspectives for the implementation of IPM in viticulture. Crop Protection, 97: 70-84.

R Core Team, 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rodríguez-Rajo F. J., Jato V., Fernández-González M., Aira M. J. 2010. The use of aerobiological methods for forecasting Botrytis spore concentrations in a vineyard. Grana, 49: 56-65.

Rossi V., Caffi T., Giosuè S., Bugiani R., 2008. A mechanistic model simulating primary infections of downy mildew in grapevine. Ecological Modelling, 212: 480–491.

Rossi V., Salinari F., Poni S., Caffi T., Bettati T. 2014. Addressing the implementation problem in agricultural decision support systems: the example of vite.net®. Computers and Electronics in Agriculture, 100: 88-99.

Rügner A., Rumbolz J., Huber B., Bleyer G., Gisi U., Kassemeyer H.H., Guggenheim R., 2002. Formation of overwintering structures of Uncinula necator and colonization of grapevine under field conditions. Plant Pathology, 51: 322–330.

Sperschneider J., 2019. Machine learning in plant–pathogen interactions: empowering biological predictions from field scale to genome scale. New Phytologist, 228: 35–41.

Thessen, A.E., 2016. Adoption of machine learning techniques in ecology and earth science. One Ecosystem 1, 1–38.

Valdés-Gómez H., Araya-Alman M., Pañitrur-De la Fuente C., Verdugo-Vásquez N., Lolas M., Acevedo-Opazo C., Gray C., Calonnec A., 2017. Evaluation of a decision support strategy for the control of powdery mildew, Erysiphe necator (Schw.) Burr., in grapevine in the central region of Chile. Pest Management Science, 73: 1813-1821.

Volpi I., Guidotti D., Mammini M., Petacchi R., Marchi S., 2020. Managing complex datasets to predict *Bactrocera oleae* infestation at the regional scale. Computers and Electronics in Agriculture, 179: 105867.

Yo M.A.R., Illig M.A.C.R., 2017. Statistically reinforced machine learning for nonlinear patterns and variable interactions. Ecosphere, 8: 1-16.

Zou Q., Xie S., Lin Z., Wu M., Ju Y., 2016. Finding the Best Classification Threshold in Imbalanced Classification. Big Data Research, 5: 2–8.

Table S1. Results of the application on the dataset "test 1" of the algorithm C5.0 trained with different subsampling techniques: down-sampling (down), up-sampling (up), Synthetic Minority Over-sampling Technique (SMOTE) and Random Over-Sampling Examples (ROSE).

	Downy mildew				Powdery mildew				Gray mold			
	down	up	SMOTE	ROSE	down	up	SMOTE	ROSE	down	up	SMOTE	ROSE
TP	758	657	748	706	236	155	170	256	77	47	63	55
FP	378	259	385	523	252	95	114	334	64	12	56	55
TN	1451	1570	1444	1306	1847	2004	1985	1765	912	964	920	921
FN	228	329	238	280	136	217	202	116	11	41	25	33
Sensitivity	0.77	0.66	0.76	0.72	0.63	0.42	0.46	0.69	0.87	0.53	0.72	0.62
Specificity	0.79	0.86	0.79	0.71	0.88	0.95	0.95	0.84	0.93	0.99	0.94	0.94
PPV	0.67	0.72	0.66	0.57	0.48	0.62	0.59	0.43	0.55	0.80	0.53	0.50
NPV	0.86	0.83	0.86	0.82	0.93	0.90	0.91	0.94	0.99	0.96	0.97	0.96
F1	0.71	0.69	0.71	0.64	0.55	0.50	0.52	0.53	0.67	0.64	0.61	0.56
Accuracy	0.78	0.79	0.78	0.71	0.84	0.87	0.87	0.82	0.93	0.95	0.92	0.92
Balanced accuracy	0.78	0.76	0.77	0.71	0.76	0.69	0.70	0.76	0.90	0.76	0.83	0.78

Link to a GitHub repository containing the R script for ML models and a subset of data as example: https://github.com/aeditsrl/grapevine_ML