



Crops and Soils Research Paper

Cite this article: De Nart D et al (2024). Vine variety identification through leaf image classification: a large-scale study on the robustness of five deep learning models. The Journal of Agricultural Science 162, 19–32. https://doi.org/10.1017/S0021859624000145

Received: 23 November 2022
Revised: 15 December 2023
Accepted: 29 January 2024
First published online: 12 February 2024

Keywords: computer vision; cultivar recognition; leaves; viticulture

Corresponding author: R. Carraro; Email: roberto.carraro@crea.gov.it

Vine variety identification through leaf image classification: a large-scale study on the robustness of five deep learning models

D. De Nart1, M. Gardiman2, V. Alba2, L. Tarricone2, P. Storchi2, S. Roccotelli2, M. Ammoniaci2, V. Tosi2, R. Perria2 and R. Carraro2

1Council for Agricultural Research and Economics, Research Centre for Engineering and Agro-Food Processing, via Giacomo Venezian, 26, Milano, Italy and 2Council for Agricultural Research and Economics, Research Centre for Viticulture and Enology, via XXVIII Aprile, 26, Conegliano, Italy

Abstract

Varietal identification plays a pivotal role in viticulture for several purposes. Nowadays, such identification is accomplished using ampelography and molecular markers, techniques requiring specific expertise and equipment. Deep learning, on the other hand, appears to be a viable and cost-effective alternative, as several recent studies claim that computer vision models can identify different vine varieties with high accuracy. Such works, however, limit their scope to a handful of selected varieties and do not provide accurate figures for external data validation. In the current study, five well-known computer vision models were applied to leaf images to verify whether the results presented in the literature can be replicated over a larger data set consisting of 27 varieties with 26 382 images. It was built over 2 years of dedicated field sampling at three geographically distinct sites, and a validation data set was collected from the Internet. Cross-validation results on the purpose-built data set confirm literature results. However, the same models, when validated against the independent data set, appear unable to generalize over the training data and retain the performances measured during cross validation. These results indicate that further enhancement have been done in filling such a gap and developing a more reliable model to discriminate among grape varieties, underlining that, to achieve this purpose, the image resolution appears to be a crucial factor in the development of such models.

Introduction

Grapevine is one of the most important fruit species and is cultivated in more than 90 countries (FAOSTAT, 2019). World under vines is estimated at 7.3 million ha (OIV, 2021), and grape production in 2019 was estimated to be ~77 million tonnes (FAOSTAT, 2019). Grapes are used to produce wine, related fermented and distillate products, dried fruit (raisins), juice and fresh fruit (table grapes). Nonetheless, the major grape destination is winemaking (Terral et al., 2010).

The genus Vitis L. encompasses 60–80 species, 20–25 of which originated from North America, about 60 from Asia and only Vitis vinifera L. from Europe (Galet, 1988; This et al., 2006; Terral et al., 2010; Keller, 2020; WFO, 2022). This latter, evolved from a dioecious wild form, V. vinifera subsp. sylvestris (Gmelin) Hegi (Garcia and Revilla, 2013) has been domesticated between 6000 and 4000 years BC (Zohary, 1995; Arroyo-García et al., 2006; Pagnoux et al., 2021) and is of greater economic importance. The other species have been used only for breeding activities to obtain mainly rootstocks and fungus-resistant hybrids (This et al., 2004).

The number of cultivated grapevine varieties is estimated to be ~6000 (Lacombe, 2012) and this number is being increasing due to breeding activities. The most cultivated varieties worldwide are about 400 (Galet, 2000), and a great number of vine genetic resources are mainly maintained inside germplasm repositories as a source of genetic variability. Although there are ~25 000 prime names registered in the Vitis International Variety Catalogue (Maul and Töpfer, 2015) of which 13 500 are referred to V. vinifera L., there are many synonyms, homonyms and incorrect or unknown denominations in grapevine biodiversity.

Thus, the characterization and identification of varieties are of great importance, not only for taxonomic purposes, but also for rational management and use of collections, breeding tasks, compliance with national and international guidelines and obtaining plant breeder rights within the UPOV (International Union for the Protection of New Varieties of Plants) system (UPOV, 1991). In this scope, many methods have been proposed until now, but the most effective are based on morphological traits of vines (ampelography, phyllometry) and genetic analyses. Besides them, other approaches have been suggested in the last few decades,

© The Author(s), 2024. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.



such as chemotaxonomic (metabolomic profile of grape aromas) and phenol characteristics (Roggero *et al.*, 1988; Mattivi *et al.*, 1990; Preiner *et al.*, 2017).

Among genetic analyses, microsatellite markers are the most effective and widely used for grapevine variety identification purposes, due to their great number, high polymorphism and codominant expression (Thomas *et al.*, 1994; Lin and Walker, 1998; Boursiquot and This, 1999; Sefc *et al.*, 2001; This *et al.*, 2004; Sefc *et al.*, 2009). Despite this technique is lab dependent, the results can be acquired quickly, and the analysis could be performed on vegetative or woody organ samples (Migliaro *et al.*, 2012). Anyway, it is not generally effective in detecting intravarietal variability, i.e. clones or biotypes, even though in recent times some microsatellite markers, specific for somatic mutants of the berry colour, have been discovered (Migliaro *et al.*, 2017).

Ampelography, the description and classification of grapevines (This *et al.*, 2006), was the first method used earlier. Even though it has been used profitably for many years, it is not suitable for describing juvenile forms (due to their different traits compared to adult plants), and difficulties may occur in varieties with a very similar phenotype. The expression of some characteristics could also be affected by the environment and thus compromise the analyses. Aiming to reduce the subjectivity of the observations and avoid inaccurate comparisons, the descriptors have been standardized at the international level (OIV, 1983, 2009). Despite that, it remains a technique that requires good training and a lot of experience, and therefore is restricted to a small number of ampelographers.

Several methods have been developed to overcome the subjectivity of morphological observations, by using biometric measures. These approaches typically consider leaves because of their particular and distinctive patterns, which are commonly referred to in taxonomical classifications. Characteristic traits such as shape, colour and distances among landmark points can be determined with a broad set of techniques. In viticulture, the first studies based on biometric measures of leaves began during the 19th century and later improved during the last century (Goethe, 1887; Ravaz, 1902; Rodrigues, 1952; Galet, 1976; Chitwood, 2021) taking into consideration particular distances, angles and the ratio of a fixed set of leaf landmarks. Compared to ampelography, a biometric method avoids subjectivity of the observations, and is adequate for computational input (Costacurta *et al.*, 1992; Alessandri *et al.*, 1996; Soldavini *et al.*, 2007; Zhang *et al.*, 2010; Bodor *et al.*, 2012) and statistical analyses can be carried out. Different biometric approaches to data analysis have been proposed such as multivariate statistical, morphometric or neural network analyses, but in general performed on a small number of varieties and seldom with uncertainty results, especially when the number of varieties is high (Boursiquot *et al.*, 1987; Costacurta *et al.*, 1998, 2003; Mancuso, 2001; Mancuso *et al.*, 2002; Bodor *et al.*, 2017; Klein *et al.*, 2017; Pereira *et al.*, 2017; Kupe *et al.*, 2021). However, the above grapevine identification techniques are time-consuming and require adequate equipment and specialists with a lot of expertise.

To overcome these difficulties artificial intelligence methods, based on convolutional neural networks (CNNs), have been developed. Such models learn in a supervised way a set of filters that allows them to extract from the input image a set of relevant features for the purpose of image classification. CNNs are the *de facto* standard for image classification tasks in artificial intelligence, several deep learning frameworks, like Keras and Torch, allow us to implement them with moderate coding effort, and

such models have proven their validity in a broad range of usage scenarios including human, animal and plant classifications (Seng *et al.*, 2018; Chai *et al.*, 2021). CNNs have recently been proposed for grapevine identification in studies on leaf or bunch samples of different varieties (Pereira *et al.*, 2019; Škrabánek *et al.*, 2020; Liu *et al.*, 2021; Nasiri *et al.*, 2021; Yang and Xu, 2021; Koklu *et al.*, 2022) with promising results. These studies, however, present some limitations in the form of oversimplified data sets, or lack of external validation to assess the robustness of the trained models. Nasiri *et al.* (2021), for instance, consider only six varieties, while Liu *et al.* (2021) do not validate their model against external data. Moreover, because of the high accuracy values scored by their considered models, these authors limit the scope of their work to now-classical convolutional models like VGG-16 (Simonyan and Zisserman, 2015) or GoogLeNet (Szegedy *et al.*, 2015) without investigating different solutions or more recent iterations of these models.

The current study attempts to address the above-mentioned research limitations, i.e. little variability examined and absence of external validation, by applying a set of well-known modern computer vision models to a large field image data set to demonstrate their applicability to a production scenario, as well as to an external, public domain data set to evaluate their ability to generalize across different sampling procedures and techniques.

Materials and methods

Leaf data set construction

The leaf images of 27 grapevine varieties of true identity have been taken in diverse vineyards located in three different environments (northern, central and southern Italy), in the summers of 2020 and 2021. The resulting data set consists of 26 382 images, with an unbalanced number of samples for each variety (Fig. 1). All images were taken during the period from the flowering to the maturity of the berries, both in field, directly on the canopy and in lab on detached leaves, using a solid background white colour.

Only the upper side of one whole leaf was captured in each photo and leaves with some evident deformations, i.e. disease symptoms or any other growth malformation, were not considered. The sample included only adult leaves, growing in the middle portion of the shoots, during the berry set and the veraison period, the most effective for the leaf characterization according to the OIV methodology (OIV, 2009). Thus, too young or too old leaves, attached on the upper or lower shoot side respectively, were excluded from the trial.

The resolution of the images ranged from a minimum of 1920 × 1280 pixels (mobile phone) to a maximum of 5184 × 3456 pixels (camera), thus including in the training set a number of different sensors to allow the model to abstract over the apparatus used to produce the images.

External testing data set

To measure the model's performance on radically different data from the samples generated by our data collection procedure, the Grapevine Leaves data set by Vlah (2021) was considered. This data set is hosted on the Kaggle platform and made available to the public under CC BY-NC-SA 4.0 license; it consists in over 1000 images of grapevine adult leaves, with a resolution of 1536 × 2048 pixels, taken in a German vineyard, located in Geisenheim,

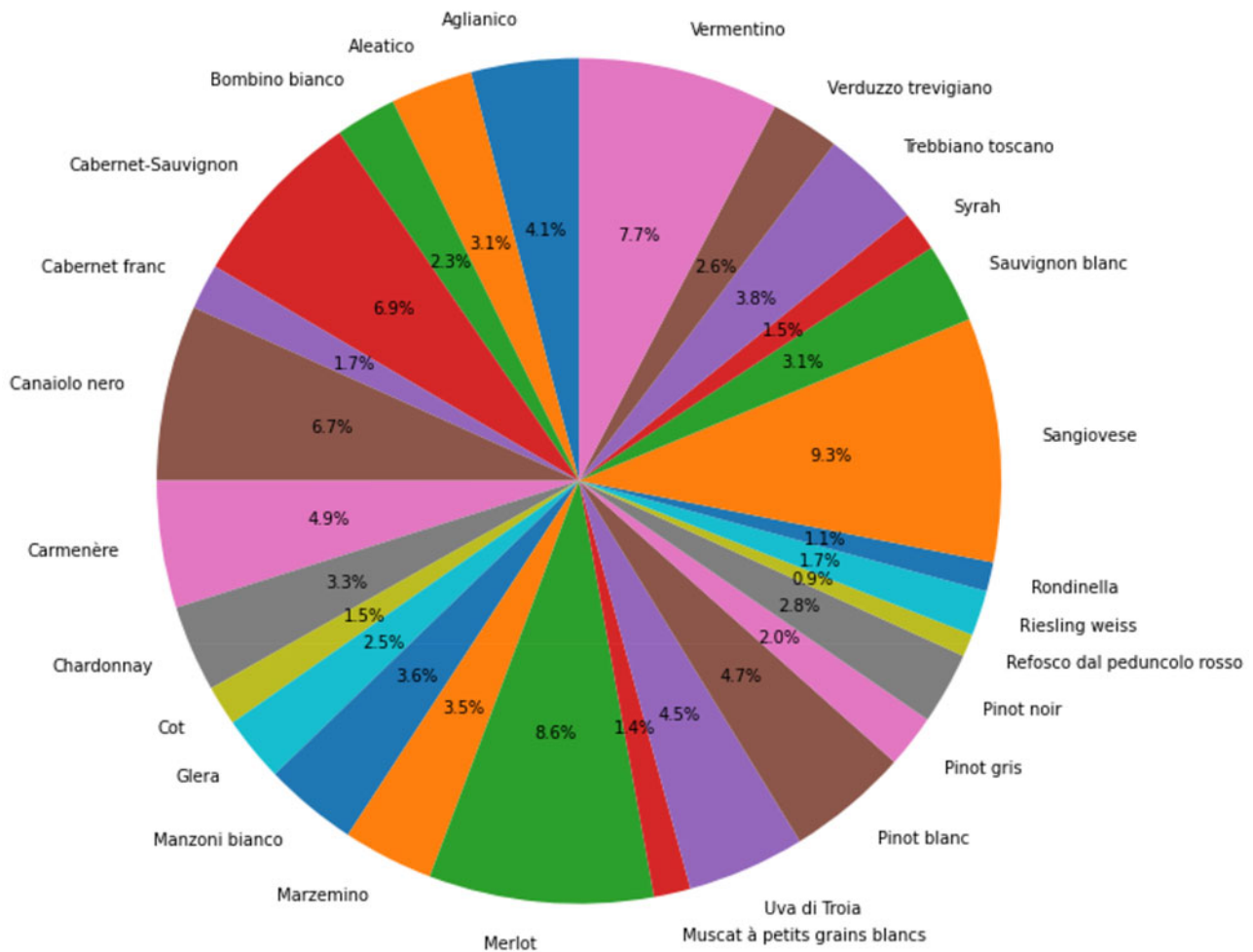


Figure 1. Class cardinality proportions within the data set.

in August 2020 using an Apple iPhone 7 camera. This data set shares 8 of the 11 grapevine varieties with our own: Cabernet franc, Cabernet-Sauvignon, Chardonnay, Merlot, Pinot noir, Riesling weiss, Sauvignon blanc and Syrah. This data set was acquired in different years and different geographical areas, under diverse environmental conditions that may influence the expression of leaf ampelographic characteristics (OIV, 2009; Chitwood *et al.*, 2016); moreover, Vlah's leaf image collection was made by a different camera, taking the images only in field. Therefore, these elements provide sufficient data diversity with our training data that can be considered a useful external data sample, suitable to test model's robustness.

Image classification models and training

Five well-established CNN models were considered for the experiment: ResNet 50 (He *et al.*, 2016), MobileNet V2 (Sandler *et al.*, 2018), Inception Net V3 (Szegedy *et al.*, 2016), Inception ResNet V2 (Szegedy *et al.*, 2017) and EfficientNet (Tan and Le, 2019) in its variants B0, B3 and B5. Although none of these models is to be considered state-of-the-art in the field of image classification (Wortsman *et al.*, 2022), they all provide reasonably good overall performance and they have widely available implementations in a number of deep learning packages such as TensorFlow, Keras and

PyTorch. Each of the models studied is very complex, with millions of parameters to be learned during training. For example, the smallest model considered here, MobileNet V2, has ~3.4 million trainable parameters. Hence it is of vital importance to optimize the training procedure to achieve good results in acceptable times. A stratified cross-validation procedure (Zeng and Martinez, 2000) was used to randomly partition the original data set into ten equal-sized subsets, called folds, each one respecting the original data set's class proportions: i.e. the most represented class remained the most represented class across all the ten partitions, and the other classes were represented proportionally. Once folds were computed, an iterative process began and at each iteration the partitions, built in the previous step, were grouped into three larger partitions: the training set, validation set and test set, which are referred to as splits. The training set was made of eight folds of the data, and the other two splits of a single fold each. Each one of the splits, being either a fold or the union of eight folds, respected the class proportions of the original data set. Of the three splits, the training set was used to feed the model during training, the validation set was used to check model progress during training and finally the test set was used to perform model evaluation. At each iteration, a new model instance was trained over the training set and evaluated on the test set, producing a set of class predictions for each image in

the latter split. The procedure was repeated many times until each fold was used once as a test set, which means that every image in the data set received a class prediction by a model that was not fed with it during its training. Such predictions were used to evaluate the model performance metrics over the whole data set.

The stochastic gradient descent (Bottou, 2010) training procedure was used for all models, with triangular learning rate scheduling (Smith, 2017). This is an iterative procedure also and it requires the training data to be processed multiple times, each one called an epoch. Due to its iterative nature, the training procedure could, theoretically, go on forever and it is up to the data scientist to stop it when an appropriate fit is achieved. Since it is impossible to know *a priori* the optimal number of training epochs, they were determined empirically by introducing the validation set. The training procedure was stopped when the performance measured on the validation set achieved a maximum and no further progress could be observed. Such a maximum point can be considered as the best fit, since it reasonably provides a sweet spot between underfitting and overfitting. Since the validation set was used to tune the number of training epochs, its data were, as a matter of fact, embedded in the trained model, even though it was not actually processed at training time, hence the need for a third split to perform an unbiased evaluation.

All models were trained for a maximum of 20 epochs with online data augmentation, which means generating multiple versions of the same image as it is passed to the model during the training. This solution, with respect to a pre-computed set of perturbed images, highlighted two advantages: it is more memory efficient (fewer images to be loaded in the GPU memory) and introduces a higher degree of randomization over different epochs, allowing the model to achieve a higher tolerance towards sub-optimal images. The training images were augmented using random rotations, vertical flips, horizontal flips and brightness adjustments to increase variability in the training data, with replication padding to avoid disrupting the original pixel colour distributions. The various transformations were applied stochastically in cascade, meaning that a wing image could be, for instance, both flipped and rotated, to maximize the randomness of transformations and, hopefully, the model robustness against noisy data.

Analysis

Three well-known classification metrics (Powers, 2011) were used to assess the model performance:

- (1) Accuracy, which in binary classification is defined as the number of true positives over the total number of considered predictions and can therefore be extended to the multiclass scenario by defining it as the fraction of correctly classified samples:

$$Accuracy = \frac{|correct\ predictions|}{|predictions|} \quad (1)$$

It is used to evaluate the overall model performance regardless of how errors are distributed among different classes and which type they belong to. Since neural networks evaluate a probability for each class, it is a frequent practice, in a multiclass setting like ours, to consider the scores produced by the model as a ranking. In this case, the prediction is considered correct if the ground truth class is among the first n classes of the ranking, that is, classes with the highest probability scores. When used in this fashion, Accuracy is commonly referred to as Accuracy@ n

where n is the number of labels to be considered part of the prediction; in this work Accuracy@3 and Accuracy@5 were used, as it is a common practice in image classification benchmarks (Krizhevsky *et al.*, 2012).

- (2) Precision, also known as positive-predictive value, is the fraction of positive values that are true positives. It is used to evaluate the model performance with respect to a given class. It represents a measure of how good the model is at avoiding false positives.
- (3) Recall, also known as specificity, is the fraction of positive samples correctly identified by the system. It represents a measure of how good the model is at avoiding false negatives, and, like Precision, it is used to evaluate the class-wise performance.

Precision and Recall, being complementary to each other, are frequently accompanied by their harmonic mean called *F1* score, which can be evaluated as follows:

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (2)$$

where *TP* is the number of true positives, and *FP* and *FN*, respectively, are the number of false positives and false negatives.

F1 score ranges from 0 to 1 and a higher score indicates better overall prediction quality; moreover, being a harmonic mean, it is a lower value than the algebraic mean and it dramatically drops when one of the two values gets close to zero. Therefore, to have a high *F1* value both Precision and Recall must be close to 1; in other words, having one of the two scores close to 1 is not enough to achieve a high or even average value if the other metric indicates a poor performance.

In addition, confusion matrices were used to visualize the overall classification quality for each model. A confusion matrix is a square matrix where each row contains the instances of a given class, i.e. the variety samples, and each column reports the variety name predicted by the model. Correct classifications appear on the diagonal of such a matrix. Misclassifications (situations in which the model does not match the correct class) are, instead, scattered in the remainder of the matrix (Figs 2 and 3).

Results

Cross-validation results

The cross validation was performed as described in Materials and methods section for all models on the complete set of in-house collected data. Accuracy results for all the considered models are shown in Table 1. Values were calculated for each test split considered in the cross-validation procedure and then averaged. The Inception ResNet and Inception Net architectures appeared to perform better than the others, with the EfficientNet models placing between ResNet and MobileNet.

Due to our experimental design, our cross-validation procedure was built on pre-computed stratified partitions of the data set. Hence, each image was present in the test partition of the data exactly once, making a union of the test split predictions and evaluating global metrics over the ten replicas of all models considered. By performing this aggregation some distributional information was lost. However, as shown in Table 1, the Accuracy standard deviation among different folds was <0.01 for MobileNet and Inception Net, while slightly above 0.01 for

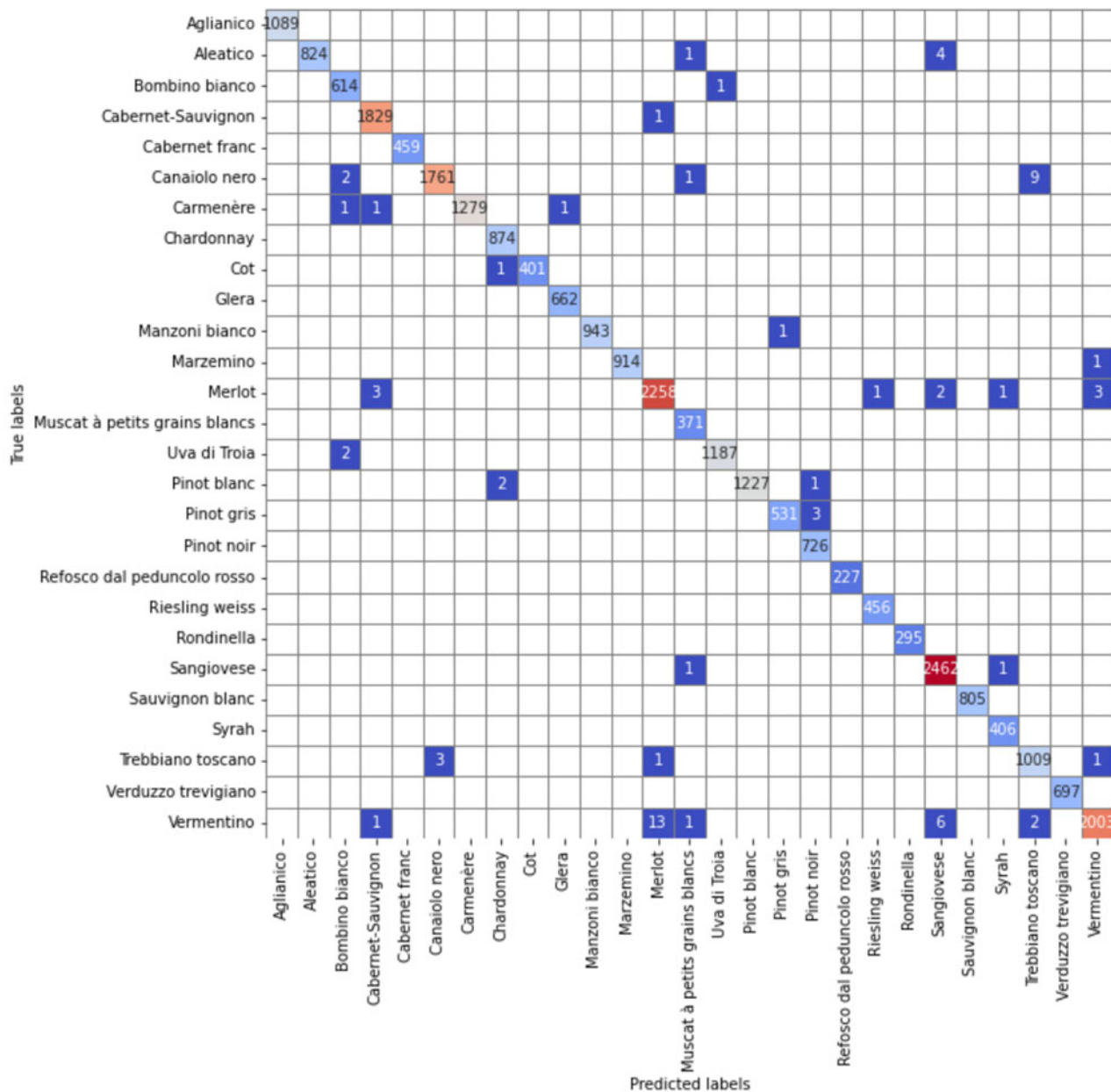


Figure 2. Inception Net V3 confusion matrix on the cross-validated data.

Inception ResNet and the EfficientNet models, implying that all models, except for ResNet 50, achieved homogeneous performance over different folds. By merging the fold evaluation results, a confusion matrix was computed, and class-specific metrics evaluated, namely Precision, Recall, and *F*-score, for each model on relevant-sized samples.

For all models, confusion matrices are very diagonal, with no more than 9% of the samples residing outside the diagonal. This is particularly evident for the best performing models, like Inception Net, whose confusion matrix is shown in Fig. 2. The errors made by such a model are very few and episodic, with the notable exception of recurrent misclassifications between the Merlot and Vermentino varieties, although in very low numbers, more precisely 16 samples, of which 13 were Vermentino samples labelled as Merlot and three vice versa.

When considering models with lower accuracy, classification errors become less episodic, and some error patterns emerge. For instance, in the EfficientNet B5 confusion matrix shown in Fig. 3, the confusion between Vermentino and Merlot is more evident, but also other confusion clusters emerge, such as Canaiolo nero-Trebbiano toscano, Trebbiano toscano-Merlot, Trebbiano toscano-Vermentino and the one among the three Pinot cultivars. It can be easily observed how most errors occur over instances of the Canaiolo nero, Merlot, Sangiovese, Trebbiano toscano and Vermentino varieties. However, these classes, except for Trebbiano toscano, have a high cardinality, in fact they are, along with Cabernet-Sauvignon, the largest classes in the data. Cabernet-Sauvignon, despite being a numerous class, shows consistently good classification accuracy across all models, implying that its distinctive leaf features are easily learned by all models.

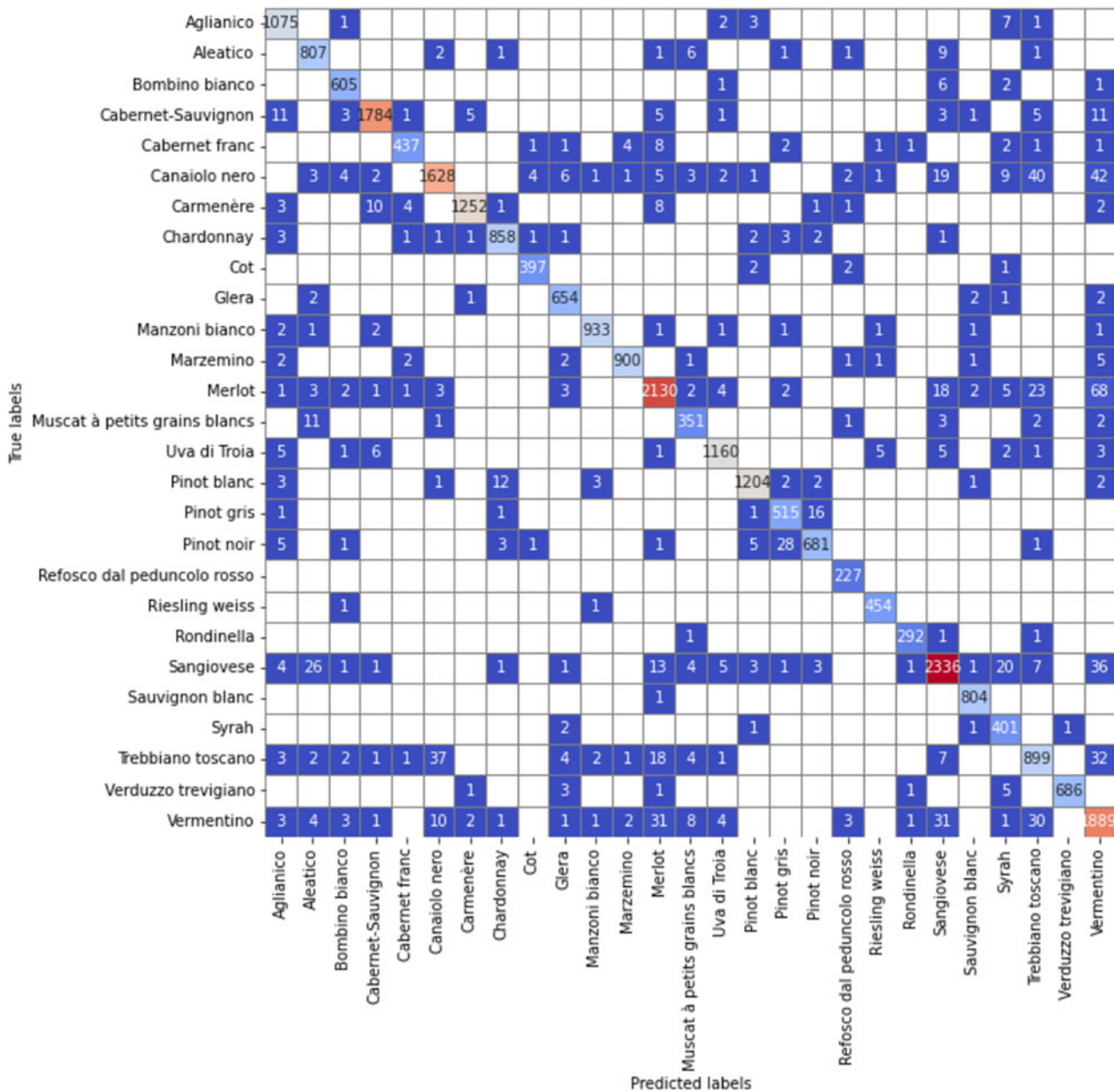


Figure 3. EfficientNet B5 confusion matrix on the cross-validated data.

Table 1. Mean and standard deviation of the accuracy, the number of true positives over the total number of considered predictions, for the cross validation

Model	Average Accuracy	Standard deviation of Accuracy
Inception Net V3	0.997	0.002
Inception ResNet V2	0.993	0.011
MobileNet V2	0.976	0.004
EfficientNet B5	0.961	0.012
EfficientNet B3	0.948	0.010
EfficientNet B0	0.924	0.016
ResNet 50	0.913	0.222

To better illustrate error distribution, for each model, Precision and Recall were considered for each class, and plot their values as shown in Fig. 4 where the measured values for the models ResNet 50, EfficientNet B5 and Inception Net V3 are displayed. Classes in the top right corner of each chart are predicted with little or no errors, while moving to the left side Precision decreases, and moving to the bottom Recall decreases. It can be easily noticed how ResNet, the worst scoring model, has no classes in the chart’s upper left corner, but rather a distribution that forms a sort of circle around said point, implying that there are classes with remarkably high Precision score and classes with very high Recall, but not both. On the other hand, EfficientNet B5 achieves overall better scores, resulting in several classes converging towards the upper right corner of the chart with few problematic ones, especially Trebbiano toscano, remaining quite far from it. Finally,

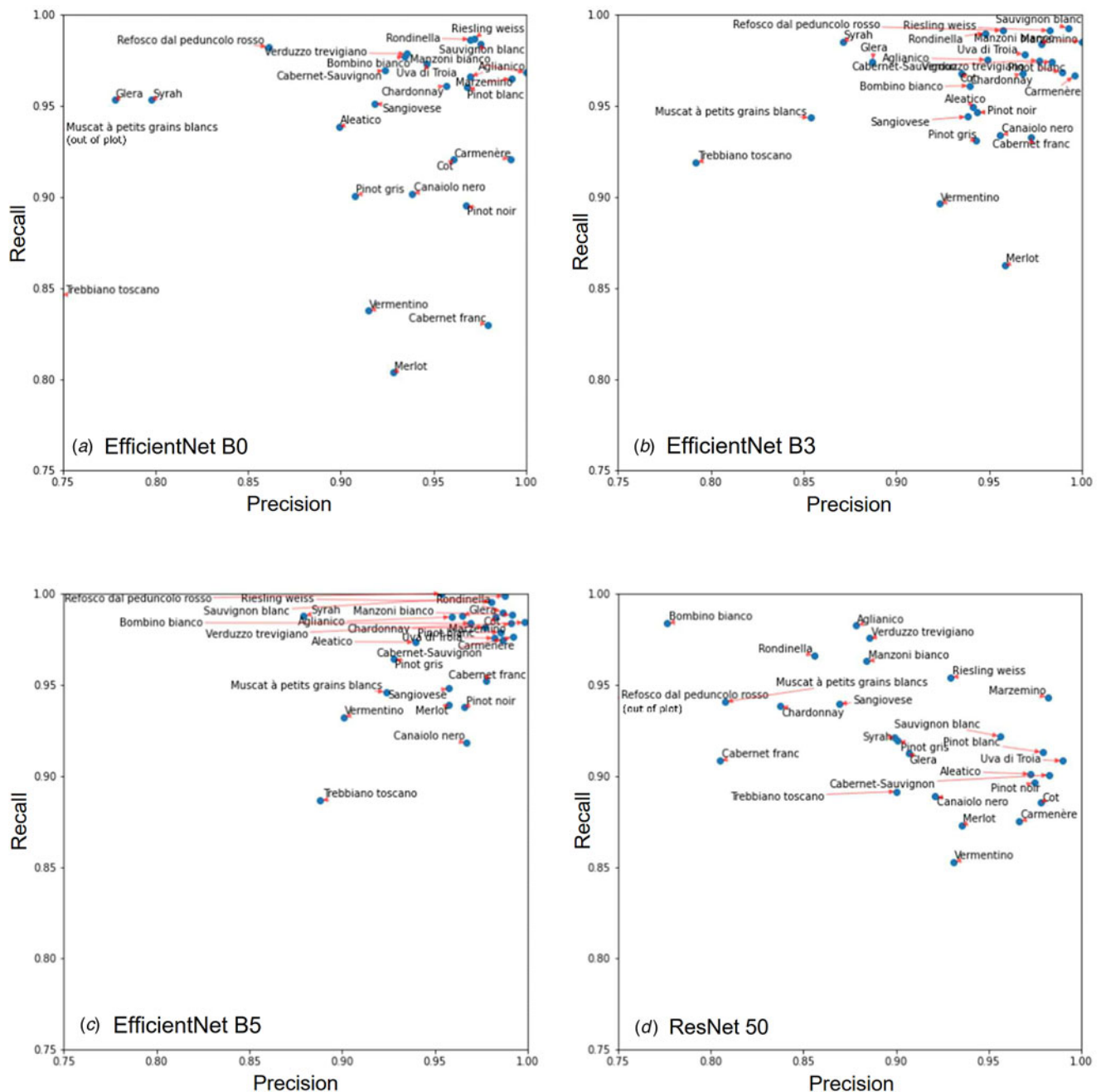


Figure 4. Positioning of classes with respect to measured Precision and Recall.

Inception Net V3, the best scoring model, has all classes firmly placed in the upper right corner of the Precision–Recall chart.

To further illustrate differences in class performances, the *F1* score was considered, i.e. the harmonic mean of Precision and Recall. All evaluated *F1* scores are presented in Table 2, which provides further evidence of how errors are not evenly distributed among the considered classes. Furthermore, it explains as such differences appear to be systematic, as varieties like Cabernet-Sauvignon, Marzemino and Uva di Troia achieve a *F1* score greater than 0.94 for all models, while varieties like Glera, Tebbiano toscano, Vermentino and Merlot appear to be consistently more difficult to correctly classify. These results underline that all models fitted the training data reasonably well and some varieties are consistently harder data to classify than others, as

some cultivars are notoriously very similar to each other and thus not easy to identify by ampelographers.

Convolutional features' analysis

To gather further insights on the training process outcome, it is possible to map how varieties are distributed in the learned feature space. Leveraging the models' layered architecture it is possible to ignore, for all models, the last layer, i.e. the one effectively implementing classification, and consider the features extracted by the convolutional layers as vectors representing the visual information in the image. Since these features typically come in thousands, to reduce the multidimensionality of data into a lower dimensional space a principal component analysis

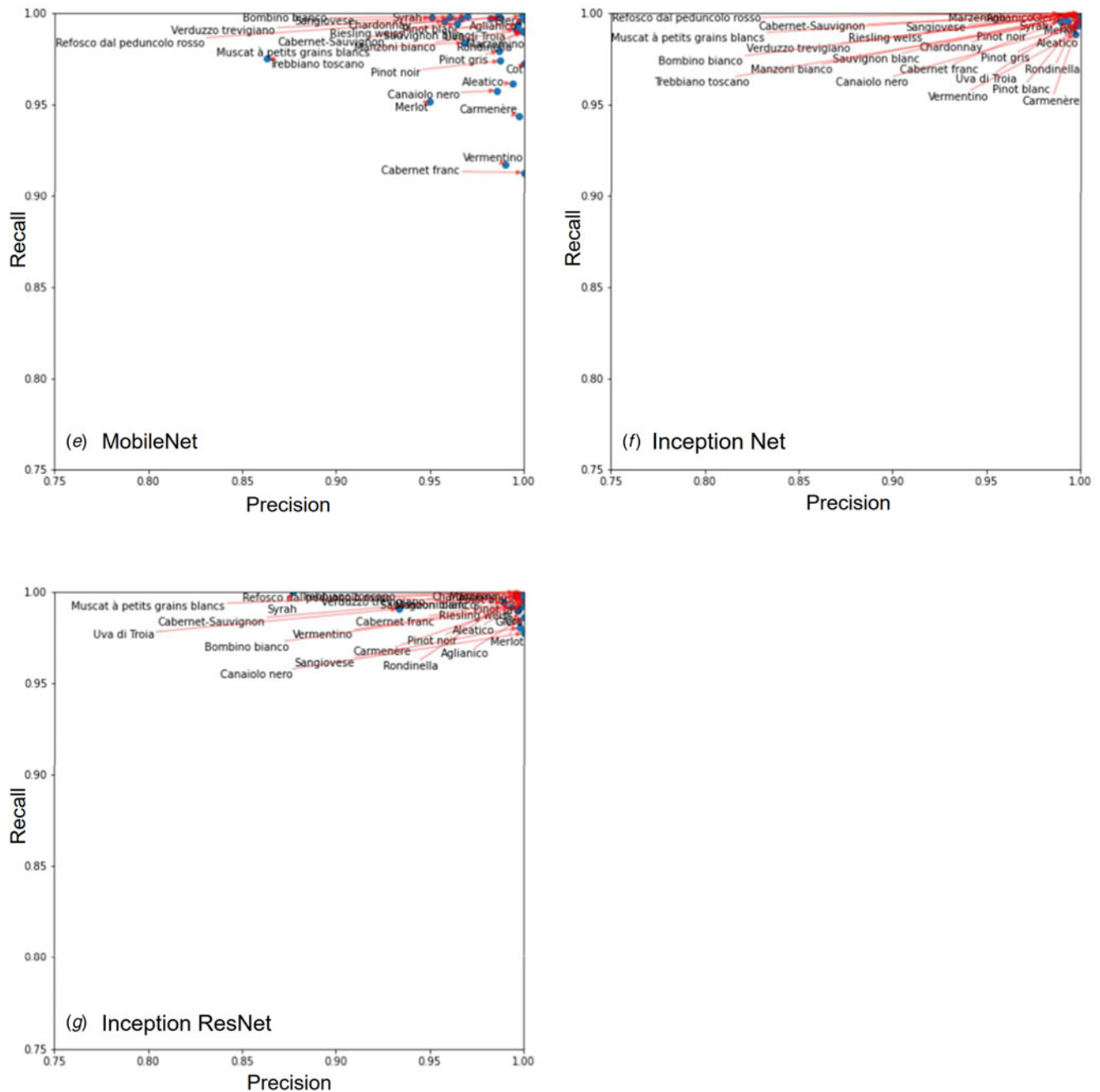


Figure 4. Continued.

has been performed aiming to find out patterns and relationships between the grapevine varieties more effectively, in an intelligible overview of our data. The feature space learned by the Inception Net V3 model projected into three dimensions is shown in Fig. 5. Ideally, visually similar varieties should occupy the same region of such a space or at least sit close to each other, and the more the point clouds representing two distinct varieties are distant, the easier it is for the model to discriminate between these two varieties. The varieties Uva di Troia, Cabernet-Sauvignon, and, to a lesser extent, Carmenère spread out across the three principal components implying that they span a wide variety of features and some of their individuals are starkly different from the rest of the data, hence very well recognizable. Other varieties, like

Pinot or Muscat, instead form a very dense point clouds occupying a considerably smaller share of space, implying that the provided data for these varieties is more self-consistent. It is also evident that there is significant overlap between several classes, and even though only a three-dimensional projection is visible of a much higher dimensional space, it is nevertheless a hint of the fact that the lines between certain varieties are blurry, and the model may confound them.

External data set validation

Finally, all considered models were trained on the full data set described in the ‘Cross-validation results’ section and tested

Table 2. Measure of the overall prediction quality (F1 scores) of considered varieties achieved by the models

Variety	EfficientNet B0	EfficientNet B3	EfficientNet B5	Inception ResNet V2	Inception Net V3	MobileNet V2	ResNet 50
Aglianico	0.97	0.96	0.97	0.99	1.00	1.00	0.93
Aleatico	0.92	0.95	0.96	1.00	1.00	0.98	0.94
Bombino bianco	0.96	0.95	0.98	1.00	1.00	0.98	0.87
Cabernet-Sauvignon	0.95	0.98	0.98	1.00	1.00	0.98	0.94
Cabernet franc	0.90	0.95	0.96	1.00	1.00	0.95	0.85
Canaiolo nero	0.92	0.94	0.94	0.99	1.00	0.97	0.90
Carmenère	0.95	0.98	0.98	1.00	1.00	0.97	0.92
Chardonnay	0.96	0.97	0.98	1.00	1.00	0.99	0.89
Cot	0.94	0.95	0.99	1.00	1.00	0.99	0.93
Glera	0.86	0.93	0.98	1.00	1.00	0.99	0.91
Manzoni bianco	0.98	0.99	0.99	1.00	1.00	0.99	0.92
Marzemino	0.98	0.98	0.99	1.00	1.00	0.99	0.96
Merlot	0.86	0.91	0.95	0.99	0.99	0.95	0.90
Muscat à petits grains blancs	0.80	0.90	0.93	1.00	0.99	0.98	0.87
Uva di Troia	0.96	0.97	0.98	0.96	1.00	0.99	0.95
Pinot blanc	0.96	0.98	0.98	1.00	1.00	1.00	0.94
Pinot gris	0.90	0.94	0.95	1.00	1.00	0.98	0.91
Pinot noir	0.93	0.94	0.95	1.00	1.00	0.98	0.93
Refosco dal peduncolo rosso	0.92	0.97	0.98	1.00	1.00	0.98	0.74
Riesling weiss	0.98	0.99	0.99	1.00	1.00	0.99	0.94
Rondinella	0.98	0.97	0.99	1.00	1.00	0.99	0.91
Sangiovese	0.93	0.94	0.95	0.99	1.00	0.98	0.90
Sauvignon blanc	0.98	0.99	0.99	1.00	1.00	1.00	0.94
Syrah	0.87	0.92	0.93	0.93	1.00	0.97	0.91
Trebbiano toscano	0.79	0.85	0.89	0.99	0.99	0.92	0.90
Verduzzo trevigiano	0.96	0.98	0.99	1.00	1.00	0.98	0.93
Vermentino	0.87	0.91	0.92	0.99	0.99	0.95	0.89

with Vlah's external data set. The Accuracy@1, Accuracy@3 and Accuracy@5 results are shown in Table 3. The performance difference between cross validation and this new evaluation is evident and it appears clear that no model can offer satisfactory performance over this data set when considering only the single most probable class. However, when considering the three most probable classes, the Accuracy score improves significantly (up to 0.75) and it is further improved by considering the five most probable classes (up to 0.83).

These numbers suggest that some models, EfficientNet B5 in particular, managed to learn robust features that allowed them to recognize characteristic traits and features that are invariant across different data sets. Other models, on the other hand, like EfficientNet B0, learned features that are way too specific with respect to the training data and do not allow them to generalize over differently sampled data.

Discussion

The main goal of this research is to create a tool that can identify a grapevine variety only with one or few leaf images acquired in

vineyard by the user, not requiring in this way specific expertise and equipment. This is a hard goal, due to the variability of the leaf morphological traits that is affected also by the cultivation environment. To reduce this variability, only adult leaves grown in the middle part of the shoot, during the period from the flowering to the berry maturity, were considered, which have very similar characteristics and many parameters that can be effective in discriminating the cultivars. Although, some varieties may have adult leaves that could differ in shape, as illustrated in Fig. 6, the nets were fed with all the acquired leaves.

In the current study, cross-validation results confirm the findings of many authors (Pereira *et al.*, 2019; Škrabánek *et al.*, 2020; Liu *et al.*, 2021; Nasiri *et al.*, 2021; Yang and Xu 2021), suggesting that well-established computer vision models can fit extremely well a data set of grape leaves. Images not previously used by the model during the training phase but acquired through the same process that generated the training set can be classified with high accuracy. This is especially interesting for cultivars that have a very similar morphological aspect and are difficult to distinguish also for ampelographers. That is, for example, the group of Pinot cultivars. Pinot blanc and Pinot gris, studied in

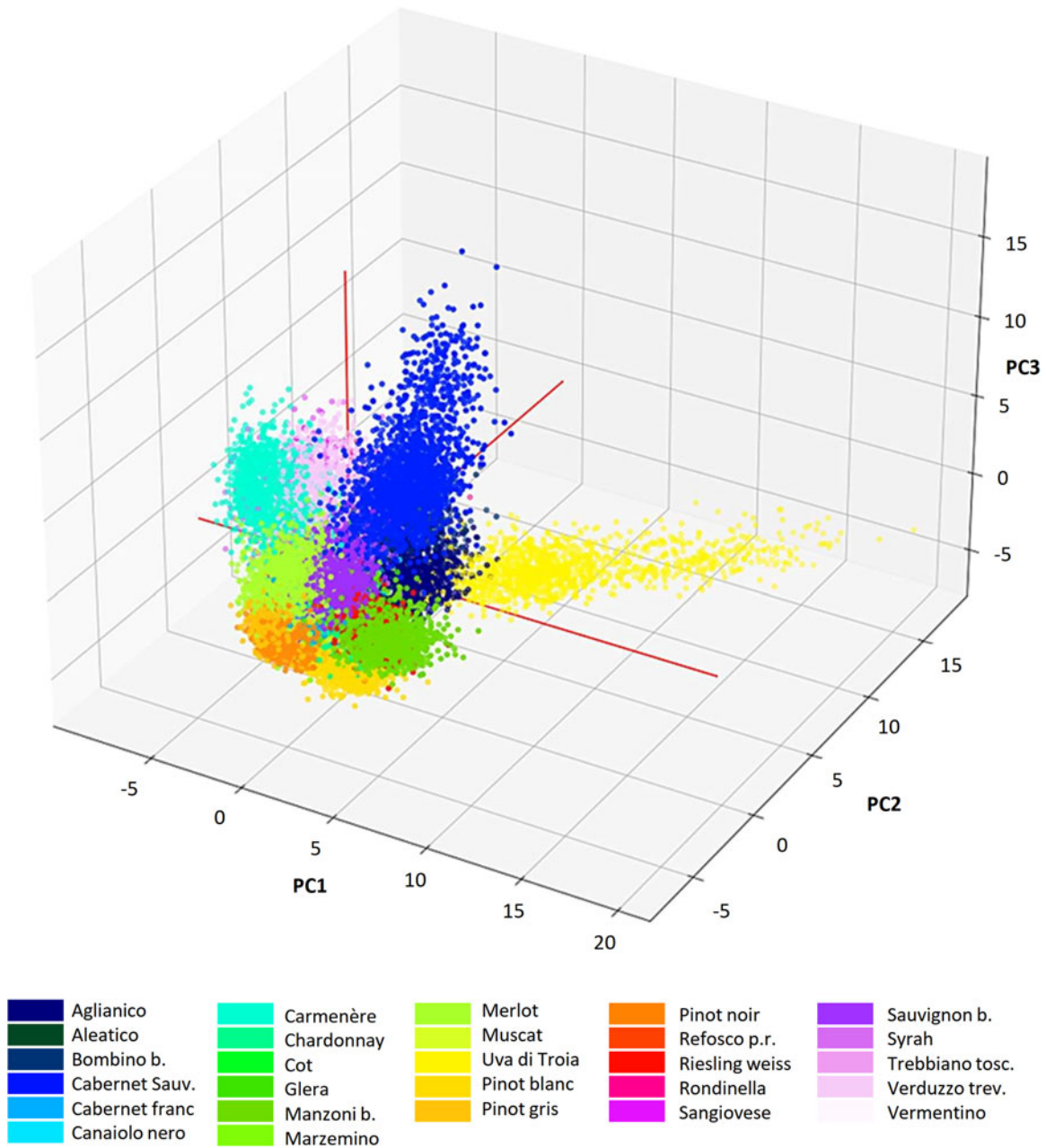


Figure 5. Vector space learned by the Inception Net V3 model.

Table 3. Model Accuracy scores including the external Kaggle data set

Model	Input size	Trainable parameters	Accuracy@1	Accuracy@3	Accuracy@5
MobileNet V2	224 × 224	3.4M	0.29	0.49	0.61
EfficientNet B0	224 × 224	11M	0.19	0.34	0.40
ResNet 50	224 × 224	23M	0.28	0.47	0.55
Inception Net V3	299 × 299	24M	0.32	0.53	0.63
Inception ResNet V2	299 × 299	56M	0.28	0.46	0.56
EfficientNet B3	300 × 300	12M	0.32	0.53	0.68
EfficientNet B5	456 × 456	30M	0.35	0.75	0.83

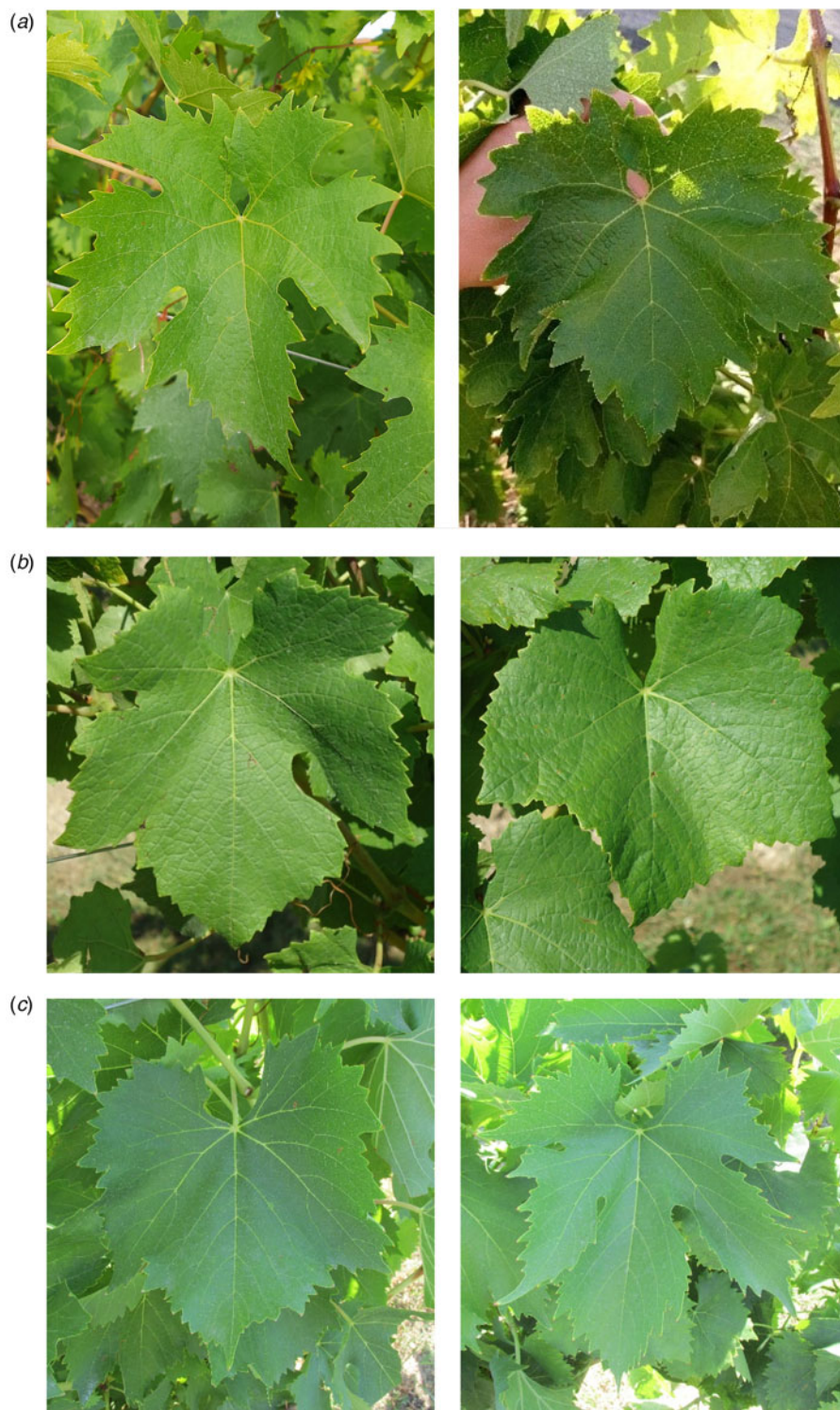


Figure 6. Two samples of Trebbiano toscano (a), Pinot noir (b), and Sangiovese with remarkably different visual features.

this topic, are cultivars generated by a bud mutation of Pinot noir, and maintained by means of vegetative propagation (Vezzulli *et al.*, 2012; Pelsy *et al.*, 2015). The main different characteristic is the colour of the berries, instead the leaves are highly similar. Nevertheless, the Inception Net V3 model is able to distinguish between the three varieties and only three misclassifications occurred (Fig. 2). Instead, with the worse classifying net (EfficientNet B5) more than 50 misclassifications have been done (Fig. 3). These observations are however too episodic to

draw strong hypothesis on why one architecture outperforms another. On the other hand, given a collection of vine leaf images, what is clear is how many of the considered models can achieve over 0.9 Accuracy in a cross-validation scenario. These results are possible because cross validation guarantees us that training, validation and test data are truly homogeneous, as the data generated by the experimental sampling procedure were split into non-overlapping sets to perform evaluations. In fact, when splitting data with cross validation, leaf images in the test portion

of the data are not only taken with the same devices used to acquire training data, but also are taken under the same light conditions, over the same time span and, more importantly, they come from the same vineyards, which implies that test leaves have undergone the same environmental conditions that at least a fraction of their training counterparts had. However, due to the overwhelming range of environmental conditions a plant can be exposed to, achieving such a high level of homogeneity between training data and unknown data is improbable, and therefore it should be considered as a bias. Our experiments on an external data set, generated by a different process applied at a different time in a different geographical zone, show a drastic decrease in the model performance, implying that the considered models are not able to generalize over training data enough to replicate cross-validation performances because of such a bias.

The evaluation presented in Table 3 suggests that the number of trainable parameters is not directly proportional to the model's robustness, implying that the performance decrease is not due to underfitting, but rather overfitting with the training data. Usually, in CNNs, a larger model size implies a higher degree of abstraction, i.e. the inference of higher level aggregated features, possibly image-wide, like, for instance, the overall leaf shape, or relational features, including the distance between certain shapes. This is because more trainable parameters generally imply more filters and more pooling layers which allow the model to process more the input image before feeding the feature vector to the final layer of classification. Our results apparently imply how this kind of abstraction can indeed provide us with better results in a cross-validation scenario, but it does not infer features which provide our model with robustness over environmental conditions.

On the other hand, models that consider larger input images appear to be more robust, which allow us to hypothesize that, from a machine learning perspective, small local visual features, such as leaf margin shapes and patterns, are more robust vine variety predictors than broad, high-level features like the overall shape or vein topology.

To overcome the criticality highlighted in the current research, future works will explore (a) new models with larger input layers, (b) how to overcome the classical leaf image classification approach presented in this paper by experimenting new training methods, such as Triplet Loss (Dong and Shen, 2018; Ge, 2018) which allow us to build different feature spaces and (c) hybrid model architectures that include, in addition, grapes and shoot images, and other information such as day of the year, geographical coordinates or weather variables to be used as predictors or to implement *a posteriori* heuristics. Moreover, considering the recent innovations and developments in autonomous robotic systems in viticulture (Moreno and Andújar, 2023; Rançon *et al.*, 2023) it is foreseeable that leaf images of a large number of varieties could be taken in a short time. This would make it possible to analyse a significant number of cultivated varieties and to improve and generalize the results of the proposed approach.

Conclusions

The results of the current analysis confirm the claim that these computer vision models can fit a large data set of grape leaves extremely well and are able to correctly classify cultivars when images are acquired under strictly controlled similar conditions. They also suggest that the performances of the same models worsen significantly when applied to an external data set gathered under different environmental conditions and using different

devices. Moreover, the results suggest that current image classification models do not cope well with the intrinsically high variability of environmental conditions that can be found in a field scenario, as even an expert curated data set with several thousand samples apparently does not guarantee a satisfactory model robustness for practical field usage. The conducted evaluation highlighted that model size, i.e. the number of trainable parameters, is not a proxy for model robustness, on the other hand input size appears a driving factor towards achieving a higher robustness, thus image resolution appears to be a crucial factor in developing new models for this task. These observations suggest how fine leaf features carry significant information with respect to the classification task which may get lost when images are downsampled to the most common input layer sizes like 224 square pixels.

A further criticality lies in the distribution of cultivars according to their visual features presented in the 'Discussion' section which suggests how models tend to learn a feature space in which some cultivars are highly adjacent if not overlapping. While this is cognitively sound as it reflects the visual similarity between these varieties, it also hinders model robustness, as the decision boundaries among them are prone to overfitting and may rely on spurious, non-relevant features. These two insights suggest how the considered classes, i.e. different cultivars, cannot be considered as equidistant or somehow even spaced in terms of visual similarity and how the distinction between said classes may lie in fine features which are easily lost with image pre-processing and downscaling.

Author contributions. D. D. N., P. S., L. T., M. G. and R. C. conceived and designed the study. M. A., V. A., V. T., S. R., M. G., R. C. and R. P. conducted data gathering. D. D. N. performed statistical analyses and prepared graphics. D. D. N., R. C. and M. G. wrote the article. R. C. supervised the work.

Funding statement. The authors gratefully acknowledge the financial support of the Italian Ministry of Agricultural, Food and Forestry Policies (MiPAAF) that funded the project AgriDigit, sub-project SUVISA (D. M. 36510, 20/12/2018).

Competing interests. None.

References

- Alessandri S, Vignozzi N and Vignini A (1996) Ampelocads (ampelographic computer-aided digitizing system): an integrated system to digitize, file and process biometrical data from *Vitis* spp. leaves. *American Journal of Enology and Viticulture* 47, 257–267.
- Arroyo-García R, Ruiz-García L, Bolling L, Ocete R, López MA, Arnold C, Ergul A, Söylemezoglu G, Uzun HI, Cabello F, Ibáñez J, Aradhya MK, Atanassov A, Atanassov I, Balint S, Cenis JL, Costantini L, Goris-Lavets S, Grando MS, Klein BY, McGovern PE, Merdinoglu D, Pejic I, Pelsy F, Primikiriou N, Risovannaya V, Roubelakis-Angelakis KA, Snoussi H, Sotiri P, Tamhankar S, This P, Troshin L, Malpica JM, Lefort F and Martinez-Zapater JM (2006) Multiple origins of cultivated grapevine (*Vitis vinifera* L. ssp. *sativa*) based on chloroplast DNA polymorphisms. *Molecular Ecology* 15, 3707–3714.
- Bodor P, Baranyai L, Bálo B, Tóth E, Strever A, Hunter J and Bisztray G (2012) GRA.LE.D. (GRApevine LEaf Digitalization) software for the detection and graphic reconstruction of ampelometrical differences between *Vitis* leaves. *South African Journal of Enology and Viticulture* 33, 1–6.
- Bodor P, Hajdu E, Baranyai L, Deák T, Bisztray GD and Bálo B (2017) Traditional and landmark-based geometric morphometric analysis of table grape clone candidates. *Mitteilungen Klosterneuburg* 67, 20–27.
- Bottou L (2010) Large-scale machine learning with stochastic gradient descent. In Lechervallier Y and Saporta G (eds), *COMPSTAT'2010*. Switzerland: Springer, pp. 177–186.

- Boursiquot J and This P** (1999) Essai de définition du cépage. *Progrès Agricole et Viticole* **116**, 359–361.
- Boursiquot J, Faber M, Blachier O and Truel P** (1987) Computerization and statistical analysis of ampelographic data. *Agronomie* **7**, 13–20.
- Chai J, Zeng H, Li A and Ngai EW** (2021) Deep learning in computer vision: a critical review of emerging techniques and application scenarios. *Machine Learning with Applications* **6**, 100134.
- Chitwood DH** (2021) The shapes of wine and table grape leaves: an ampelometric study inspired by the methods of Pierre Galet. *Plants, People, Planet* **3**, 155–170.
- Chitwood DH, Rundell SM, Li DY, Woodford QL, Yu TT, Lopez JR, Greenblatt D, Kang J and Londo JP** (2016) Climate and developmental plasticity: interannual variability in grapevine leaf morphology. *Plant Physiology* **170**, 1480–1491.
- Costacurta A, Calò A and Giust M** (1992) Analisi ampelografiche ed ampelometriche mediante sistemi di rilevatori computerizzati. In *Proceedings of the Congress on Grapevine Germplasm*, pp. 565–572. Alghero, IT.
- Costacurta A, Calò A, Carraro R, Giust M and Lorenzoni C** (1998) Varietal identification using procedures of stepwise discrimination. *Acta Horticulture* **528**, 51–58.
- Costacurta A, Crespan M, Milani N, Carraro R, Flamini R, Aggio L, Ajmone-Marsan P and Calò A** (2003) Morphological, aromatic and molecular characterization of Muscat vines and their phylogenetic relationships. *Rivista di Viticoltura ed Enologia* **2–3**, 13–28.
- Dong X and Shen J** (2018) Triplet loss in Siamese network for object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 459–474.
- FAOSTAT** (2019) *Crops and Livestock Products*. Rome: FAO. Available at <https://www.fao.org/faostat/en/#data/QCL> (accessed 08 November 2022).
- Galet P** (1976) *Précis d'ampélographie pratique*. Montpellier, FR: Déhan.
- Galet P** (1988) *Cépages et vignobles de France. Les vignes américaines*. Montpellier, FR: Déhan.
- Galet P** (2000) *Dictionnaire encyclopédique des cépages*. Paris, FR: Hachette.
- Garcia A and Revilla E** (2013) The current status of wild grapevine populations (*Vitis vinifera* ssp. *silvestris*) in the Mediterranean basin. In Sladonja B (ed.), *The Mediterranean Genetic Code – Grapevine and Olive*. IntechOpen, pp. 51–72. London, UK: IntechOpen Limited.
- Ge W** (2018) Deep metric learning with hierarchical triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 269–285. Berlin, DE: Springer Nature.
- Goethe H** (1887) *Handbuch der Ampelographie*. Berlin, DE: Paul Parey Verlag.
- He K, Zhang X, Ren S and Sun J** (2016) Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778. Piscataway, NJ, USA: Institute of Electrical and Electronics Engineers.
- Keller ME** (2020) *The Science of Grapevine*. Amsterdam, NL: Elsevier.
- Klein LL, Caito M, Chapnick C, Kitchen C, O'Hanlon R, Chitwood DH and Miller AJ** (2017) Digital morphometrics of two North American grapevines (Vitis: Vitaceae) quantifies leaf variation between species, within species, and among individuals. *Frontiers in Plant Science* **8**, 373.[Q4]
- Koklu M, Unlarsen M, Ozkan I, Aslan M and Sabanci K** (2022) A CNN–SVM study based on selected deep features for grapevine leaves classification. *Measurement* **12**, 110425.
- Krizhevsky A, Sutskever I and Hinton GH** (2012) ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* **25**, 1097–1105.
- Kupe M, Sayıncı B, Demir B, Ercisli S, Baron M and Sochor J** (2021) Morphological characteristics of grapevine cultivars and closed contour analysis with elliptic Fourier descriptors. *Plants* **10**, 1350.
- Lacombe T** (2012) Contribution à l'étude de l'histoire évolutive de la vigne cultivée (*Vitis vinifera* L.) par l'analyse de la diversité génétique neutre et de gènes d'intérêt (PhD thesis). Centre International d'Etudes Supérieures en Sciences Agronomiques, Montpellier, France.
- Lin H and Walker M** (1998) Identifying grape rootstocks with simple sequence repeat (SSR) DNA markers. *American Journal of Enology and Viticulture* **49**, 403–407.
- Liu Y, Su J, Shen L, Lu N, Fang Y, Liu F, Song Y and Su B** (2021) Development of a mobile application for identification of grapevine (*Vitis vinifera* L.) cultivars via deep learning. *International Journal of Agricultural and Biological Engineering* **14**, 172–179.
- Mancuso S** (2001) Clustering of grapevine (*Vitis vinifera* L.) genotypes with Kohonen neural networks. *Vitis* **40**, 59–63.
- Mancuso S, Boselli M and Masi E** (2002) Distinction of 'Sangiovese' clones and grapevine varieties using elliptic Fourier analysis (EFA), neural networks and fractal analysis. *Advances in Horticultural Science* **15**, 61–65.
- Mattivi F, Scienza A, Failla O, Villa P, Anzani R, Tedesco G, Gianazza E and Righetti P** (1990) *Vitis vinifera*: a chemotaxonomic approach. Anthocyanins in the skin. In *Proceedings of the 5th International Symposium on Grape Breeding*, pp. 119–133. St. Martin/Pfalz, Fribourg, DE: Vitis Journal.
- Maul E and Töpfer R** (2015) Vitis International Variety Catalogue (VIVC): a cultivar database referenced by genetic profiles and morphology. *BIO Web of Conferences* **5**, 01009.
- Migliaro D, Morreale G, Gardiman M, Landolfo S and Crespan M** (2012) Direct multiplex PCR for grapevine genotyping and varietal identification. *Plant Genetic Resources* **11**, 182–185.
- Migliaro D, Crespan M, Muñoz-Organero G, Velasco R, Moser C and Vezzulli S** (2017) Structural dynamics at the berry colour locus in *Vitis vinifera* L. somatic variants. *Acta Horticulturae* **1157**, 27–32.
- Moreno H and Andújar D** (2023) Proximal sensing for geometric characterization of vines: a review of the latest advances. *Computers and Electronics in Agriculture* **210**, 107901.
- Nasiri A, Taheri-Garavand A, Fanourakis D, Zhang Y and Nikoloudakis N** (2021) Automated grapevine cultivar identification via leaf imaging and deep convolutional neural networks: a proof-of-concept study employing primary Iranian varieties. *Plants* **10**, 1628.
- OIV** (1983) *The OIV Descriptor List for Grape Varieties and Vitis Species*. Paris, FR: International Organisation of vine and wine.
- OIV** (2009) *Second Edition of the Descriptor for Grape Varieties and Vitis Species*. Paris, FR: International Organisation of vine and wine.
- OIV** (2021) *State of the World Vitivinicultural Sector in 2020*. Paris, FR: International Organisation of vine and wine.
- Pagnoux C, Bouby L, Valamoti SM, Bonhomme V, Ivorra S, Gkatzogia E, Karathanou A, Kotsachristou D, Kroll H and Terral JF** (2021) Local domestication or diffusion? Insights into viticulture in Greece from Neolithic to Archaic times, using geometric morphometric analyses of archaeological grape seeds. *Journal of Archaeological Science* **125**, 105263.
- Pelsy F, Dumas V, Bévilacqua L, Hocquigny S and Merdinoglu D** (2015) Chromosome replacement and deletion lead to clonal polymorphism of berry color in grapevine. *PLoS Genetics* **11**, e1005081.
- Pereira CS, Morais R and Reis MJCS** (2017). Recent advances in image processing techniques for automated harvesting purposes: a review. *Intelligent Systems Conference (IntelliSys)*, London, UK, 566–575.
- Pereira CS, Morais R and Reis MJCS** (2019) Deep learning techniques for grape plant species identification in natural images. *Sensors* **19**, 4850.
- Powers DM** (2011) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technology* **2**, 37–63.
- Preiner D, Tomaz I, Markovic Z, Stupic D, Andabaka Z, Sikuten I, Cenbauer D, Maletic E and Kotic JK** (2017) Differences in chemical composition of 'Plavac Mali' grape berries. *Vitis* **56**, 95–102.
- Rançon F, Keresztes B, Deshayes A, Tardif M, Abdelghafour F, Fontaine G, Da Costa JP and Germain C** (2023) Designing a proximal sensing camera acquisition system for vineyard applications: results and feedback on 8 years of experiments. *Sensors* **23**, 847.
- Ravaz L** (1902) *Les vignes américaines: Porte-greffes et producteurs directs*. Montpellier and Paris, FR: Goulet.
- Rodrigues AE** (1952) Um metodo filométrico de caracterização. Fundamentos. Descrição – Técnica Operatória. Lisboa, PT: Serviço editorial da repartição de estudos, Informação e Propaganda, Direcção General dos Serviço Agrícolas, Ministério da Economia.
- Roggero JP, Larice JL, Rocheville-Divorne C, Archier P and Coen S** (1988) Composition anthocyanique des cépages. I. Essay de classification par analyse en composantes principales et par analyse factorielle discriminante. *Revue Française d'Oenologie* **112**, 41–48.
- Sandler M, Howard A, Zhu M, Zhmoginov A and Chen L** (2018) MobileNetV2: inverted residuals and linear bottlenecks. In *Proceedings of*

- the IEEE conference on computer vision and pattern recognition. Piscataway, NJ, USA: Institute of Electrical and Electronics Engineers.
- Sefc KM, Lefort F, Grando M, Scott K, Steinkellner H and Thomas M** (2001) Microsatellite markers for grapevine: a state of the art. In Roubelakis Angelakis K (ed). *Grapevine Molecular Physiology & Biotechnology*, pp. 433–463. Berlin, DE: Springer Science+Business Media.
- Sefc K, Pejić I, Maletić E, Thomas M and Lefort F** (2009) Microsatellite markers for grapevine: tools for cultivar identification; pedigree reconstruction. In: Roubelakis-Angelakis, K.A. (eds) *Grapevine Molecular Physiology & Biotechnology*, pp. 565–596. Dordrecht: Springer.
- Seng KP, Ang LM, Schmidtke LM and Rogiers SY** (2018) Computer vision and machine learning for viticulture technology. *IEEE Access* 6, 67494–67510.
- Simonyan K and Zisserman A** (2015) Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations, pp. 1–14. San Diego, USA, Computational and Biological Learning Society.
- Škrabánek P, Doležel P, Matoušek R and Junek P** (2020) RGB images driven recognition of grapevine varieties. In *International Workshop on Soft Computing Models in Industrial and Environmental Applications*, pp. 216–225. Berlin, DE: Springer Nature.
- Smith LN** (2017) Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pp. 464–472. Piscataway, NJ, USA: Institute of Electrical and Electronics Engineers.
- Soldavini C, Schneider A, Stefanini M, Dallserra M and Policarpo M** (2007) Superampelo un software per la descrizione ampelografica e ampelometrica della vite. *Italus Hortus* 14, 39–40.
- Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A** (2015) Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9. Piscataway, NJ, USA: Institute of Electrical and Electronics Engineers.
- Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z** (2016) Rethinking the Inception architecture. In *European Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826. Piscataway, NJ, USA: Institute of Electrical and Electronics Engineers.
- Szegedy C, Ioffe S, Vanhoucke V and Alemi AA** (2017) Inception-V4, Inception-ResNet and the impact of residual connections on learning. In *Thirty-first AAAI Conference on Artificial Intelligence*, pp. 4278–4284. Palo Alto, CA, USA: AAAI Press.
- Tan M and Le Q** (2019) EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6105–6114. International Machine Learning Society.
- Terral JF, Tabard E, Bouby L, Ivorra S, Pastor T, Figueiral I, Picq S, Chevance JB, Jung C, Fabre L, Tardy C, Compan M, Bacilieri R, Lacombe T and This P** (2010) Evolution and history of grapevine (*Vitis vinifera*) under domestication: new morphometric perspectives to understand seed domestication syndrome and reveal origins of ancient European cultivars. *Annals of Botany* 105, 443–455.
- This P, Jung A, Boccacci P, Borrego J, Botta R, Costantini L, Crespan M, Dangi G, Eisenheld C, Ferreira-Monteiro F, Grando S, Ibáñez J, Lacombe T, Laucou V, Magalhães R, Meredith CP, Milani N, Peterlunger E, Regner F, Zulini L and Maul E** (2004) Development of a standard set of microsatellite reference alleles for identification of grape cultivars. *Theoretical and Applied Genetics* 109, 1448–1458.
- This P, Lacombe T and Thomas MR** (2006) Historical origins and genetic diversity of wine grapes. *Trends in Genetics* 22, 511–519.
- Thomas M and Cain P and Scott NS** (1994) DNA typing of grapevines: a universal methodology and database for describing cultivars and evaluating genetic relatedness. *Plant Molecular Biology* 25, 939–949.
- UPOV** (1991) *International Convention for the Protection of New Varieties of Plants*, Vol. 221. Geneva, CH: The International Union for the Protection of New Varieties of Plants.
- Vezzulli S, Leonardelli L, Malossini U, Stefanini M, Velasco R and Moser C** (2012) Pinot blanc and Pinot gris arose as independent somatic mutations of Pinot noir. *Journal of Experimental Botany* 63, 6359–6369.
- Vlah M** (2021) Grapevine leaves. Available at <https://www.kaggle.com/ds/1248678> (accessed 08 November 2022).
- WFO** (2022) *Vitis L.* Available at <http://www.worldfloraonline.org/taxon/wfo-4000040377> (accessed 08 November 2022).
- Wortsman M, Ilharco G, Gadre SY, Roelofs R, Gontijo-Lopes R, Morcos AS, Namkoong H, Farhadi A, Carmon Y, Kornblith S and Schmidt L** (2022). Model soups: averaging weights of multiple finetuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning* 162, pp. 23965–23998. Baltimore, Maryland, USA: Proceedings of Machine Learning Research.
- Yang B and Xu Y** (2021) Applications of deep-learning approaches in horticultural research: a review. *Horticulture Research* 8, 123.
- Zeng X and Martinez TR** (2000) Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence* 12, 1–12.
- Zhang J, Yanne P and Li H** (2010) Identification of grape varieties via digital leaf image processing by computer. In *33rd World Congress of Vine and Wine. 8th General Assembly of the OIV*. Tbilisi, GE: International Organisation of Vine and Wine.
- Zohary D** (1995) Domestication of the grapevine *Vitis vinifera L.* in the near east. In McGovern P, Fleming SJ and Katz SHE (eds), *The Origins and Ancient History of Wine*. New York, USA: Gordon and Breach, pp. 23–30.