AgriEngineering

MDPI

Article

An Image-Based Approach for Grape Yield Estimation During Mechanical Harvesting Using Pre-Trained Visual Embeddings

Giuliano Vitali [1,*], Federico Terzi [2], Marco Arru [1], Francesco Pelusi [2], Cristiano Fragassa [1], Matteo Gatti [2,3] and Michele Mattetti [1]

[1] Department of Agricultural and Food Sciences, Alma Mater Studiorum—University of Bologna, Viale Giuseppe Fanin 44, 40127 Bologna, Italy; m.arru@ardesiatechnologies.com (M.A.); cristiano.fragassa@unibo.it (C.F.); michele.mattetti@unibo.it (M.M.)

[2] Department of Sustainable Crop Production, Università Cattolica del Sacro Cuore, Via Emilia Parmense 84, 29122 Piacenza, Italy; federico.terzi1@unicatt.it (F.T.); francesco.pelusi1@unicatt.it (F.P.); matteo.gatti@unicatt.it (M.G.)

[3] Remote Sensing and Spatial Analysis Research Center (CRAST), Università Cattolica del Sacro Cuore, Via Emilia Parmense 84, 29122 Piacenza, Italy

* Correspondence: giuliano.vitali@unibo.it

Abstract

Accurate yield mapping is essential for precision viticulture, but direct yield measurement during mechanical harvesting remains difficult because of the heterogeneous harvested biomass. This study evaluates the use of images acquired during harvesting to estimate grape yield through features extracted from pre-trained deep learning models. Images were collected over two seasons in a commercial Merlot vineyard using action cameras mounted on a grape harvester. Five pre-trained models, including one convolutional neural network and four Vision Transformer architectures, were compared. The extracted features were related to measured yield using regularized regression models. The results show that pre-trained visual features can reliably estimate grape yield despite the presence of berries, grape clusters, leaves, shoots and motion blur. The convolutional neural network achieved prediction accuracy comparable to the Vision Transformer models, indicating that local image features remain highly informative for biomass-flow estimation. A sparse fusion approach identified more stable features, supporting improved model stability under the investigated conditions. This work demonstrates that general-purpose pre-trained visual embeddings provide an effective basis for grape yield estimation without task-specific model training. The proposed method, not depending on dedicated sensors, specific calibration or model tuning, offers a practical and scalable solution for integration into commercial grape harvesters, including aftermarket installations.

Keywords: precision viticulture; grape harvester; artificial intelligence; yield distribution; biomass flow

Check for updates

Academic Editor: Ray E. Sheriff

Received: 30 June 2026
Revised: 28 August 2026
Accepted: 14 September 2026
Published: 20 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Precision viticulture is gaining importance in vineyard management. Mapping vine vigor and soil properties provides valuable insights for managing within-field variability of yield and optimizing grape quality [1]. Multiple approaches have been used for mapping, combining remote and proximal sensing tools, including Unmanned Aerial Vehicle (UAV) and multi-spectral imagery [2,3]. Focusing on yield estimation, most of the experiments are based on the use of non-invasive digital technologies before harvest [4,5]. On the other

hand, the direct measurement of the yield at harvest is still a challenge. Using a soft-picking experimental harvester [6] contributes to maintaining grape integrity, also allowing for yield measurement [7]. Commercial solutions are also available on the market, which are based on the monitoring of mass flow by load cells mounted under the discharge conveyor belt of grape harvesters, registering the grape weights as they are being off-loaded [8].

Volume estimation of complex lumped materials is not new in computer science, and is leveraged by image analysis [9]. Such tools have already been applied to estimate the volume of loose materials in vessels with different shapes, as in handling excavation materials (e.g., [10]). Dynamic image (movie) analysis has been applied to belt-conveyed material in row material extraction [11].

The literature gives comparatively little attention to studies specifically targeting real-time image-based yield estimation during harvesting, biomass-flow sensing, and related approaches operating under dynamic harvesting conditions. This is a distinct research area characterized by rapidly changing viewpoints, motion blur, varying illumination, occlusions, and continuous crop transport, which pose challenges not typically encountered in pre-harvest imaging [12].

In the case of grapevine harvesting, the main issue still seems to be represented by the complexity of the conveyed biomass, made of grape clusters (or parts), berries, leaves (or parts of them not completely eliminated by the cleaning fan system), branches and shoots (sometimes with leaves and grapes) and grape juice. The composition of the harvested biomass mainly depends on (i) the genotype—cluster architecture [13] and a different berry retention force [14]; (ii) time of harvest—the berry retention force decreases during ripening; and (iii) mechanical aspects—the setup of mechanical harvesters, combining different systems for shaking, collecting, cleaning, transporting and storing grape berries [15].

Deep learning models have been successfully applied to object detection, tracking and image segmentation in agricultural environments. However, developing task-specific models requires large annotated datasets and substantial computational resources, and the resulting models often show limited transferability across vineyards, seasons and acquisition conditions. For this reason, the embedding technique, based on the adoption of pre-trained models, is becoming popular [16].

Embedding allows the extraction of image features, numerical values representing the characteristics of an image or part of it. Although pre-trained models learn from large, generic image datasets, they develop hierarchical visual representations that capture low-level features (e.g., edges, textures and colors), intermediate structures (e.g., shapes and spatial arrangements) and high-level semantic patterns. During mechanical harvesting, the harvested biomass exhibits characteristic visual properties, including the texture and color of berries, the morphology of grape clusters, leaves and shoots, and their spatial distribution within the harvester. These visual cues are implicitly encoded in the learned feature representations, allowing pre-trained models to provide informative descriptors that correlate with biomass flow and, consequently, grape yield, even without task-specific training.

While CNNs such as ResNet are based on local convolutions, the recent Vision Transformers (ViTs) analyze the image globally by splitting it into patches (small square matrices of pixels, also referred to as tokens). These patches are compared with one another through self-attention mechanisms, enabling the model to learn relationships related to shapes, patterns, context, and semantics across the whole image [17].

Among the most influential transformer-based approaches is DINOv2, a self-supervised model which is not trained with manual labels but by comparing large numbers of unlabeled images. Although transformer-based models generally achieve superior performance, lightweight architectures remain relevant for real-time applications or scenarios with limited computational resources (e.g., ResNet), particularly when only small datasets

are available [18]. Though these models are normally used for classification, their hidden layers can be extracted and used as image features in the so-called embedding process [19]. However, the features extracted are not guaranteed to be invariant when analyzing images coming from different populations [20]; therefore, a feature identification, comparison and validation strategy has to be adopted.

The present analysis is oriented to evaluate the possibility of estimating yield through a computer vision approach, which reformulates biomass-flow estimation as an image analysis problem. The aim is producing a proof of concept based on the use of the embedding process to analyze images captured during mechanical harvesting and to extract features from available pre-trained models. Several embedders have been compared to the ground-truth together with the persistency of identified features and values in different conditions.

To avoid ambiguity, this study distinguishes between the sensing methodology and the prediction target. Image-based sensing denotes the computer vision approach used to analyze the harvested crop flow, while biomass-flow and volume estimation represent intermediate measurements derived from the images. The final objective is real-time yield estimation during mechanical harvesting.

2. Materials and Methods

2.1. Experimental Vineyard and Harvesting System

The study was performed over two years (2024 and 2025) in a commercial vineyard planted with *Vitis vinifera* L. cv. Merlot located in Ravenna, Italy (44°26′17″ N, 12°09′27″ E). The vineyard, extending for 1.78 ha, was established in 2015 at a vine spacing of 2.8 × 1.1 m (between- and within-row, respectively), on a flat area characterized by a deep loamy-sand soil. All the vines were trained to a vertical shoot-positioned, spur-pruned cordon trellis with a bud-load of about 15 count nodes per vine distributed on 7–8 two-node spurs per vine. The cordon was set 1 m from the ground with three pairs of surmounted catch wires for a canopy wall extending 1.4 m above the cordon.

A vigor map of the hedge-trained vineyard was generated at full canopy by using the ROBO® sensor (Appleby Italiana, Roveleto di Cadeo, Italy), a tractor-mounted device allowing high-resolution mapping through a bottom-up view to exclude the vineyard floor interference. Similarly to MECS-VINE® [21], of which it represents the most recent commercial version, ROBO allows expression of vegetative vigor as the Canopy Index (CI) with a spatial resolution of 1.3 m [3]. As part of the study, CI values have been partitioned into three classes as low (LV), medium (MV) and high vigor (HV), corresponding to the 0–30, 30–70, and 70–100 percentile intervals, respectively.

The experimental layout was a randomized block design, where high (HV) and low vigor (LV) were assumed to be representative of high- and low-yield areas, respectively, as previously reported by Gatti et al. [22]. From east to west the vineyard was divided into three blocks, each including the HV and LV classes. Each vigor x block combination extended over 46.2 m and encompassed seven trellis bays (the 6.6 m space between two adjacent in-row posts), which were labeled as “B”, “D”, and “ND” depending on the different data acquisition procedures performed for agronomic ground-truthing. A buffer zone (B) corresponding to the first and last bays among the seven sections was hand-picked prior to mechanical harvest, while the second and the sixth bays were used for running destructive measurements (D). Both B and D spaces represented a fruit-free zone during mechanical operations, ensuring a gap in the grape flow along the conveyor aimed at facilitating the identification of grapes deriving from the ND vines. Non-destructive (ND) measurements were performed on the inner portion of each vigor x block combination (see Figure 1b). Per each season, destructive measurements were performed on eight vines equally split between the two sections located before and after the ND area, resulting in a

total of 48 vines per year. Yield per vine was measured by using a portable field scale, grape clusters were individually harvested and counted, and the average cluster weight was calculated accordingly. Per each tagged vine, a sub-sample of three representative clusters was collected and brought to the laboratory for fruit composition analysis. In parallel, grapes within the ND zones were left on the vines (18 plants considered for each of the six vigor x block combinations) for subsequent mechanical harvesting. Cluster number per vine was counted, and the associated total yield (Yt) was calculated based on the average cluster weight obtained from the corresponding "D" zone.

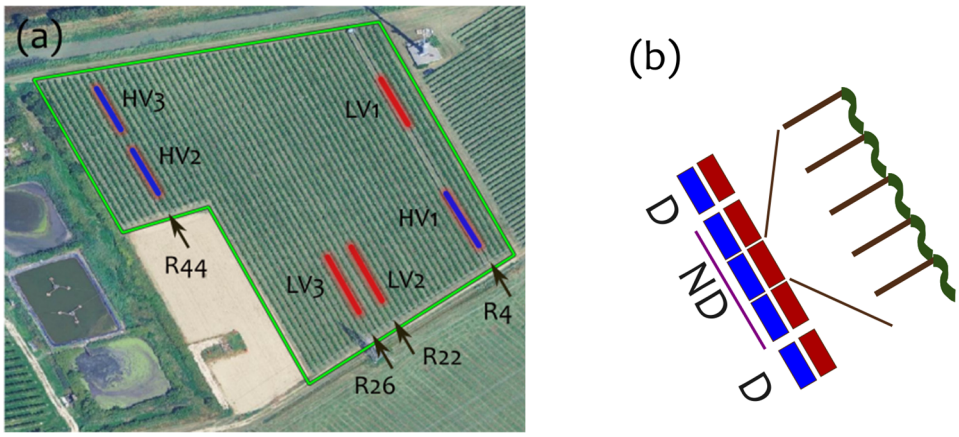

Figure 1. Aerial view of the experimental vineyard (**a**) with the path of the machine during the harvesting of the rows (R4, R22, R26, R44) and the selected 6 experimental sections, 3 replicates for HV (in blue) and LV treatments (in red), and the scheme of each experimental section (**b**), where the rectangles represent bays delimited by planting poles.

Once mechanical harvesting was completed, visible losses were quantified as (i) non-detached grapes per vine (NDGs) and (ii) grapes on the ground in the under-trellis strip (GG), as previously reported by Caprara and Pezzi [23]. Harvested yield per vine (Yh) was therefore calculated per each of the ND vines as Yh = Yt − (NDG + GG).

Tests were carried out on 11 September 2024 and 12 September 2025 using the same experimental layout in order to compare image-based yield estimation under different harvesting conditions. Because of the high yields registered during the first season, in 2025 severe cluster thinning corresponding to almost 50% fruit removal was applied prior to mechanical harvesting in the LV2 and LV3 low-vigor zones to widen the range of ground-truth values.

The mechanical harvesting was performed with a Braud 9090L grape harvester (by CNH Industrial, Turin, Italy), in which horizontal oscillating rods (shakers) detach berries—vibration frequency and amplitude are adjustable to minimize damage to vines and ensure efficient grape collection (Shaking Bow Catcher technology).

During harvesting, the machine can operate at speeds ranging from 2 to 4 km/h; in this experiment the harvester speed was set to 3 km/h. Four rows were considered, namely R4, R22, R26 and R44 (see Figure 1a), with lengths of 135, 153, 158 and 78 m, respectively, corresponding to a total harvested row length of 524 m and an overall harvesting time of ~620 s (10 min 20 s).

Grapes are collected into polyurethane baskets positioned dynamically beneath the vine canopy to minimize fruit losses. The tray funnels the dislodged grapes into two conveyor belts (rolling tapes) for both sides—on top of them, a fan is used to remove lighter impurities (leaves and debris). The biomass finally reaches a sliding sieve/grate, leaving the berries to drop down into the hopper (see Figure 2b).



(**a**) (**b**)

Figure 2. Camera installation on the grape harvester: (**a**) position of the camera in front of the conveyor belt (**b**).

2.2. Image Acquisition System

An action camera (CrossTour 9500, Crosstour, Shenzhen, China) with a field of view of 170° was used for the study; as in many commercial cameras, it is not possible to set the shutter speed; therefore, a high frame rate—120 fps—was adopted. Two cameras were mounted on the right and left sides of the harvester (Figure 2a), each looking at the output of the two conveyors (Figure 2b). The use of two opposite views of the same working area was intended to improve the spatial interpretation of the biomass flow and partially overcome the limitations of single-plane image acquisition.

Because of the high number of images generated (approximately 3.6 GB s^{-1} of uncompressed RGB image data), some preliminary trials were performed to reduce the amount of data as much as possible while preserving the relevant information. Considering that the expected linear resolution of the analysis is expected to be around 1 m (plant distance), a sampling rate of 1 fps (corresponding to 0.5 m linear resolution) was considered to be appropriate to the scope of the analysis.

Harvesting time and the total number of acquired frames are reported in Table 1. The overall image analysis workflow is shown in Figure 3.

Table 1. Characteristics of the 4 rows of the vineyard.

Row ID	Length (m)	Time (s) [1]	Frames
R4	135	160	320
R22	153	180	360
R26	158	190	380
R44	78	90	190

[1] Accuracy ± 5 s.

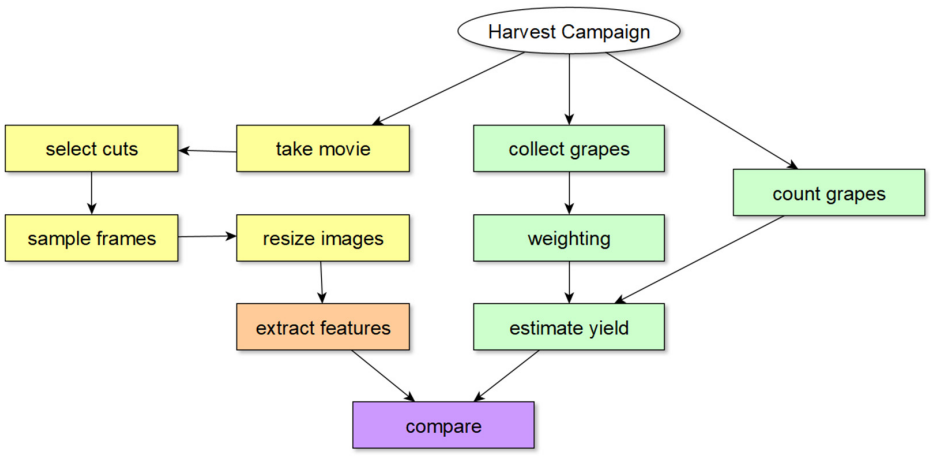

Figure 3. Pipeline used to derive the features and compare them to the ground-truth (yield).

For each experimental block a number of 14–16 images were collected. For every model, they were resized to 256 pixels (shorter side), center-cropped (224 × 224 px), and then transformed to tensor and finally normalized.

2.3. Image Embedding

Five pre-trained image embedding models were selected for comparison, representing different deep learning paradigms for visual feature extraction (Table 2). Together with ResNet, probably the most popular CNN, four Vision Transformers have been selected to cover the major learning paradigms currently used in computer vision. This selection provides a balanced comparison between supervised, distillation-based, masked self-supervised, and self-distillation approaches while avoiding unnecessary architectural variability. Consequently, the observed differences can be more confidently attributed to the representation learning strategy rather than to unrelated network design choices. Hereafter a short description is reported for each of them.

Table 2. List of models used for the analysis with some characteristics.

Embedder	Architecture	Training Dataset	Version Used	Parameters	Features
ResNet50	CNN (Residual Network)	ImageNet-1 k	50 layers	~26 M	512/1024/2048
DeiT	Vision Transformer	ImageNet-1 k	Base/patch 16	~86 M	768
BEiT-v2	Vision Transformer	ImageNet-21 K	Base/patch 14	~86 M	768
ViT	Vision Transformer	ImageNet-21 K	Base/patch 16	~86 M	768
DINOv2	Vision Transformer	142 M images	Base/patch 14	~86 M	768

ResNet50 is a CNN based on residual learning, designed to enable very deep architectures while mitigating the vanishing-gradient problem. It has become one of the most widely adopted backbone models for image recognition and feature extraction [19].

Vision Transformer (ViT) applies the transformer architecture to computer vision by splitting an image into fixed-size patches and modeling their relationships through self-attention mechanisms. Unlike CNNs, ViT captures global image context from the earliest processing stages [24].

A Data-efficient Image Transformer (DeiT) extends the Vision Transformer by introducing a knowledge distillation strategy that enables competitive performance using only ImageNet-scale datasets, substantially reducing the amount of training data required [25].

BEiT-v2 is a transformer-based model trained through masked image modeling. Instead of predicting pixels directly, it learns high-level semantic visual representations from discrete visual tokens, producing robust embeddings suitable for downstream computer vision tasks [26].

DINOv2 is a recent self-supervised Vision Transformer trained on a very large collection of unlabeled images. Without requiring manual annotations, it learns highly transferable visual representations that have shown excellent performance across a wide range of image understanding tasks [27].

Each model is available in several versions, often related to image size and number of parameters (dimensions). Every model was used in its 224 (image size) version to reduce the computation time.

Though the source images are taken from a sequence, no semantics related to the dynamics have been considered, and no fine-tuning has been applied.

In the case of ResNet50, early layers are those near the Input and can be used to select features such as borders, colors, texture and gradients; Mid layers can be used for patterns, parts and shapes; and deep layers can be used for objects, scenes and semantics.

In Vision Transformers, the Initial blocks (1–3) include features such as colors and texture, blocks 4–6 include contours, patterns, elements and sparse objects, blocks 7–9 include spatial relations, object parts, and configurations, and blocks 10–12 contain global semantics abstract concepts and scenes [28].

While in CNNs the receptive field grows gradually with depth, in transformers a global self-attention mechanism exists already from the first layers. A selection of layers and blocks have been adopted for this study: layers 2–3-4 for the CNN, blocks 5–7-911 for the VIs. Tensors extracted at intermediate levels have more dimensions; therefore, the average values have been extracted.

With the present approach there is no training dataset, nor test; as we are using pre-trained models, features extracted by embedding are directly used as estimators of the observed value (ground-truth).

2.4. Feature Identification

The extracted features were fitted to the ground-truth values (yield). Given the huge number of potential explanatory variables (features), a selection of the potentially most relevant ones is necessary.

One of most used methods for fitting with automatic selection of relevant features with many coefficients is the Least Absolute Shrinkage and Selection Operator (LASSO, [29]. The LASSO method is essentially a regression with a penalty coefficient (λ) working on the coefficients of the features (β):

$$O = \Sigma_{i=1,\mathrm{m}} (y_i - y_i')^2 + \lambda \Sigma_{j=1,n} \mid \beta_j \mid \quad (1)$$

where O is the objective function, y is the observed value and y' is the estimated one, a linear function of $x_{1,p}$, the explanatory variables (features), m are the images.

$$y_i' = \beta_o + \beta_1 x_{i,1} + \ldots + \beta_n x_{i,n} \quad (2)$$

The LASSO eliminates (zeroes) a lot of coefficients; however, when in an overfitting scenario (features >> data) many features are correlated to one another. To obtain a limited number of features, the λ (filtering) parameter may be adjusted; however, the LASSO performs a (stochastic) selection of parameters, and cannot guarantee that the same parameters are used for different data groups (tasks). In this case an extension of the LASSO has been studied that "forces" a common set of features, the Fused LASSO [30], based on a constraint combining a group penalty (λ_g) with a sparse (LASSO) penalty:

$$O = \Sigma_{i=1,\mathrm{m}} (y_i - y_i')^2 + \lambda \Sigma_{j=1,\mathrm{n};t=1,\mathrm{k}} \mid \beta_{j,t} \mid + \lambda_g \Sigma_{j=1,\mathrm{n}} \mid \beta_j \mid \quad (3)$$

where t is the dataset (task) index (in our case the two years).

The objective of this study implicitly requires identifying a compact set of features that are consistently informative across multiple datasets, in this case grape yield across the two growing seasons, rather than optimizing predictive performance. For this purpose, a further extension of the method, the Sparse Fusion LASSO (SFL) [31], has been used, which requires that the coefficients of the common features have the same values.

Alternative regularization techniques, such as Ridge regression or Elastic Net, either do not perform feature selection or do not explicitly promote coefficient agreement across multiple datasets. The SFL is particularly suited to applications where robustness and feature consistency across heterogeneous data sources are desirable.

The SFL is based on the enhanced objective function:

$$O = \Sigma_{i=1,\mathrm{m}} (y_i - y_i')^2 + \lambda \Sigma_{j=1,\mathrm{n}\ t=1,\mathrm{k}} \mid \beta_{j,t} \mid + \lambda_g \Sigma_j \mid \beta_j \mid + \lambda_f \Sigma_{\mathrm{j}=1,\mathrm{n}} \mid \beta_{1,j} - \beta_{2,j} \mid \quad (4)$$

where λ_f forces the common features of the two datasets to have similar values.

In the present analysis, a preliminary LASSO approach has been used, with the scope to see the potential of embedding without constraints. Both the LASSO and the SFL were performed on average values of features from the images taken from each block; all images contributed equally to the regression models, and no quality filtering was performed before analysis.

Given the limited number of independent harvesting blocks (n = 18), the feature-selection procedure was performed on the complete dataset, while ensuring that the number of selected features remained substantially lower than the number of available samples; this choice was also aimed at identifying the degrees of freedom in relation to the risk of over-parameterization. The selected features should therefore be regarded as exploratory and warrant validation on a larger independent dataset.

The LASSO and fusion penalty parameters were interactively adjusted to obtain a sparse subset of predictors shared across the two growing seasons while maintaining the number of selected variables below the number of independent harvesting blocks. This pragmatic criterion was preferred because the limited sample size (18 independent observations) would have resulted in highly unstable estimates of the cross-validation error.

Both LASSOs are included in Python's *sklearn* library (https://scikit-learn.org/, accessed on 1 September 2026). All the code (in Python 3.13) was developed in *VSCode* (https://code.visualstudio.com/, accessed on 1 September 2026).

3. Results

During the harvest, Merlot grapes were confirmed to be suitable for mechanical harvesting with limited juicing losses; on the other hand, the fan separator was not able to remove all leaves and shoots. Moreover, the shutter time for snapshots of the action camera was revealed to be insufficient for taking still images, as the rolling tape was discharging the mass at a high rate on the lower sliding grid. In the samples provided in Figure 4 it is possible to recognize the complexity of images where in general single and composite objects are present, consisting of single berries, grapes and grape parts, and branches with grapes and leaves, and most of those objects have motion blur—the background consists of a grid that is transversal in motion, a rolling "tapis roulant" (blue) under (variable) sun and shade and a vibrating and shaking carrier.



Figure 4. Sample of pictures taken during the harvest process from the start of the harvest (**top left**) to the end (**bottom right**).

For each harvesting block (18 each year), a slightly variable number of images was collected (14–16); for each of them the features were extracted with the five models, and regression with yields was performed on their average values.

For LASSO regression the λ value was increased till it reached the overfitting limit ($R^2 \approx 1$, [32] Hastie et al., 2009), and the resulting feature number (N_{cf}) was kept below 17. A value of λ = 0.000035 was found for ResNet50 and BEIT, 0.00004 for the other models. The number of features together with the R^2 values for the years 2024 and 2025 are reported in Table 3 (columns 3–5th).

Table 3. Features (*n*) and R^2 values with LASSO methods. Best values are shown in bold; strikethrough values indicate overfitting results.

MODEL	Layer Block	LASSO			Sparse Fusion LASSO		
		N_{cf}	R^2		N_{cfsv}	R^2	
			2024	2025		2024	2025
resnet50	L2	10	0.356	0.959	13	**0.796**	**0.995**
	L3	8	0.404	0.711	18	~~0.879~~	~~1.000~~
	L4	**15**	~~**0.879**~~	**0.951**	13	0.637	0.972
BEIT	B5	8	0.261	0.865	7	0.381	0.847
	B7	7	0.379	0.927	7	0.377	0.839
	B9	5	0.076	0.657	10	0.467	0.949
	B11	12	**0.475**	**0.947**	15	**0.639**	**0.973**
DEIT	B5	10	**0.750**	**0.915**	4	0.099	0.876
	B7	8	0.228	0.955	10	0.470	0.966
	B9	5	0.358	0.611	15	**0.794**	**0.996**
	B11	9	0.580	0.583	23	~~0.879~~	~~1.000~~
ViT	B5	13	0.816	0.956	2	0.111	0.849
	B7	7	0.359	0.885	5	0.268	0.856
	B9	12	0.834	0.827	8	**0.450**	**0.940**
	B11	13	**0.849**	**0.939**	16	~~0.879~~	~~1.000~~
dinov2	B5	6	0.232	0.855	3	0.056	0.751
	B7	14	**0.825**	**0.922**	6	0.135	0.945
	B9	10	0.398	0.939	15	**0.684**	**0.998**
	B11	6	0.381	0.723	32	~~0.879~~	~~1.000~~

The SFL method was successfully applied. A preliminary sensitivity analysis was conducted on the three coefficients, showing a low effect of λ (kept as low as 10–3 for every embedding (10^{-3}) and λ_f (=3)); a bigger effect was that of the grouping coefficient (λ_g) impacting the number of features: for the five embedders the values were, respectively, 0.5 for ResNet, 1.5 for BEIT, 1.5 for DEIT, 4 for ViT, and 0.2 for DINO. The number of common features with the same value (N_{cfsv}), together with the R^2, are reported in the last three columns of Table 3.

In Table 3 the values where overfitting occurs are strikethrough, showing that 14–15 represent the degree of freedom possessed by the data samples.

Table 3 shows how the complexity of images transfers to different model levels/blocks. Looking at the LASSO columns it is possible to observe that the best fits are obtained using semantic layers with ResNet50, BEIT and ViT, and using deeper layers/blocks with DEIT and dinoV2. The best fit observed for ResNet50, as already mentioned above, is just an effect of overfitting.

In fact, the values obtained are strongly related to the fitting parameter (λ), and there is a threshold number of features delimiting overfitting conditions; in 2024 R^2 = 0.879 represents the maximum attainable value, as does R^2 = 1 in 2025. This means that the success of ResNet50 is apparent, and that a value of 15 features is higher than the degrees of freedom in 2024.

It is possible to see that a higher R^2 also corresponds to a higher number of features, an indicator of the amount of extractable information, which has always been lower than the number of images available for each post.

As reported in Table 3, the final models retained only a limited number of predictors in each case, preserving sufficient residual degrees of freedom for the subsequent regression analysis and reducing the risk of model over-parameterization. Since the aim of the Sparse Fusion LASSO analysis was to identify predictors that are consistently informative across the two growing seasons rather than to maximize predictive accuracy, the feature-selection procedure was performed using the complete dataset. The identified predictor sets should therefore be regarded as exploratory and should be confirmed using larger independent datasets.

The relation between R^2 and the number of features is also reported in Figure 5, where the higher and lower values of the error bars refer to the values obtained for 2025 and 2024, respectively.

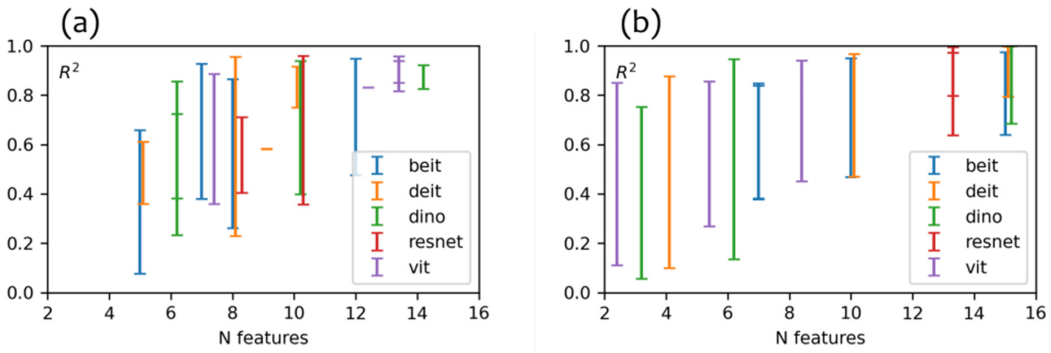

Figure 5. Barplot reporting the minimum and maximum R^2 values for every models in relation to regression features for simple LASSO (**a**) and SFL (**b**).

From the regression plots, reported in Figure 6 for LASSO and Figure 7 for SFL, emerges the importance of performing the embedding on vineyards with a high span of productivity, while they appear to not be as reliable in vineyards with a uniform yield distribution, like that in 2024.

In the same time the number of features required for the regression could be relatively low. From Table 3 it is possible to see that DINOv2, with only six features, may fit 2025's data with R^2 = 0.945, though with a low performance in 2024. The ViT embedder is also probably a good compromise, fitting with R^2 = 0.940 in 2025 and 0.450 in 2024 with just eight features.

Though amounts of 13 and 15 features allow every embedder to get top R^2 values, they are also probably too close to the overfitting limit.

Finally, RMSE and MAE values are reported (Table 4) to quantify the prediction error associated with each model. Across all experiments, both metrics remained below 1 kg m^{-1}, indicating a satisfactory level of estimation accuracy. Excluding the lowest-performing configurations (strikethrough), which are likely affected by an excessive number of selected features and the consequent over-parameterization, the best-performing models (highlighted in bold) show that ResNet50 achieves prediction errors comparable to those obtained with

the investigated Vision Transformer architectures. Specifically, RMSE values consistently range between 0.30 and 0.60 kg m^{-1}, while MAE remains within 0.25–0.35 kg m^{-1}.

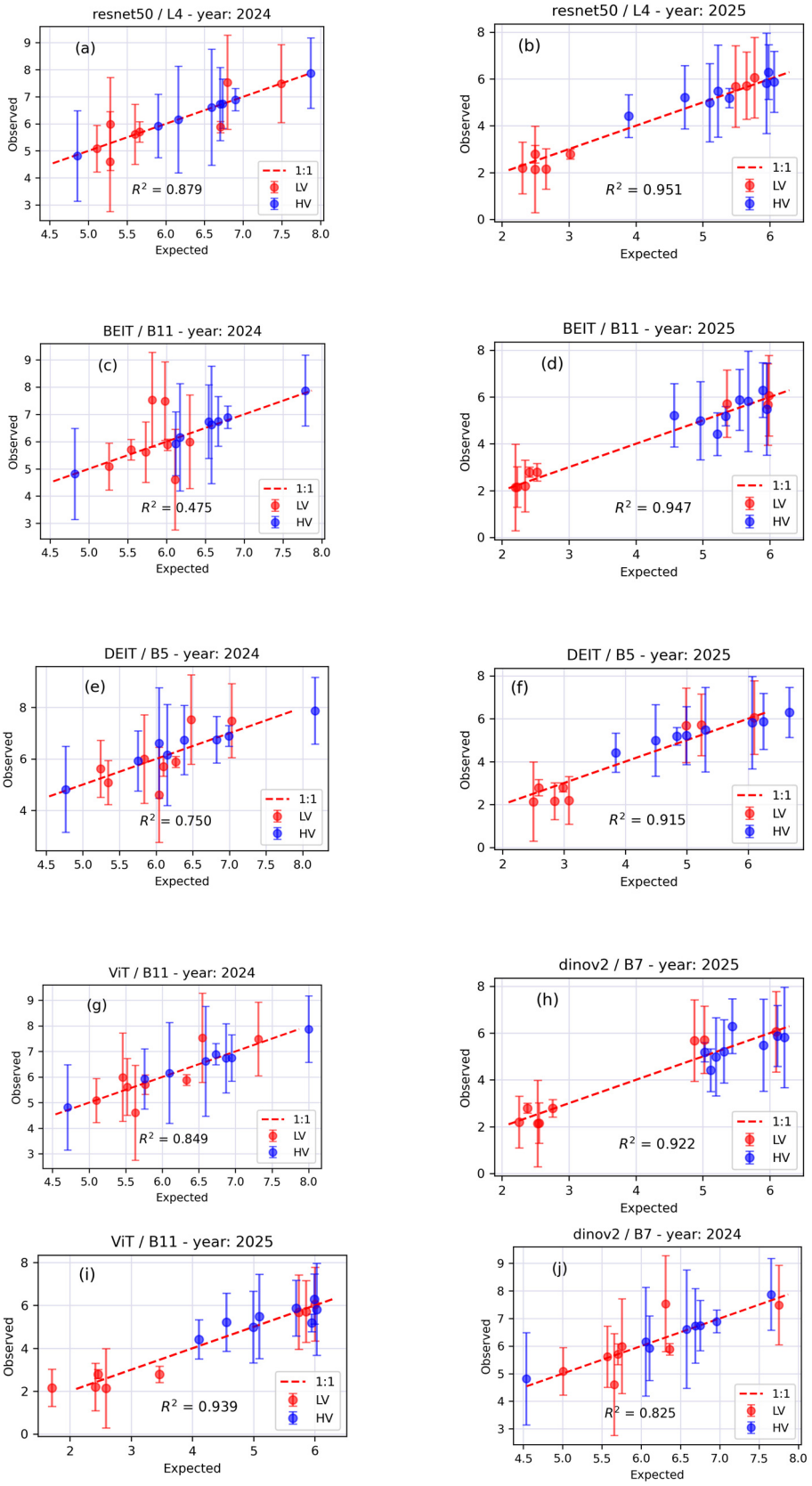

Figure 6. Scatterplots (**a**–**j**) report values of best regression obtained with simple LASSO (bold values in Table 3, columns 3–4)—LV and HV labels refer, respectively, to low-vigor and high-vigor experimental blocks.

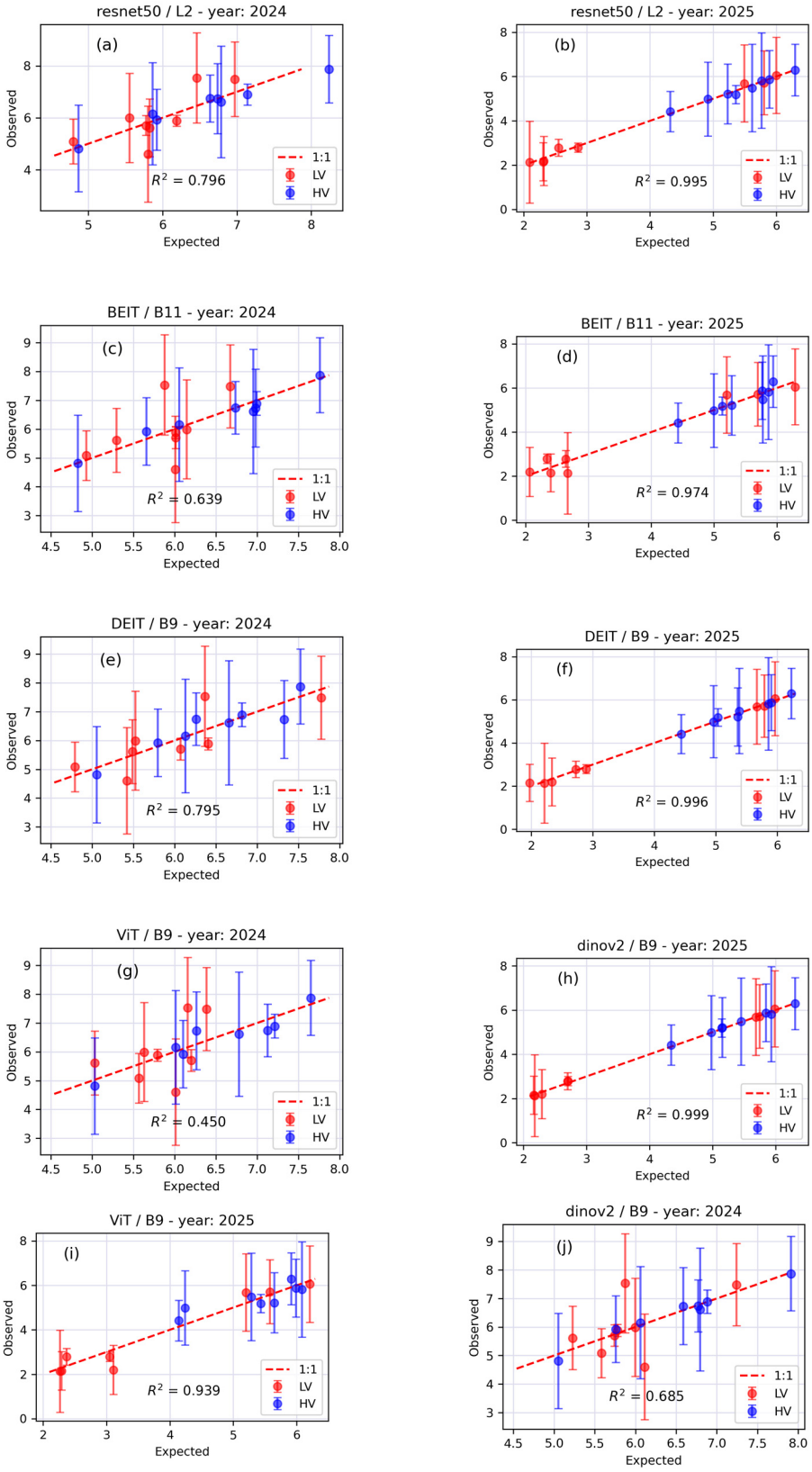

Figure 7. Scatterplots (**a**–**j**) report values corresponding to the best regressions obtained with SFL (bold values in Table 3, right hand columns) for both years—LV and HV labels refer, respectively, to low-vigor and high-vigor yields (kg m^{-1}).

Table 4. RMSE and MAE obtained from Sparse Fusion. Strikethrough values refer to over-fitting condition, while bold figures refer to best values.

MODEL	Layer Block	Sparse Fusion LASSO			
		RMSE (kg m^{-1})		MAE (kg m^{-1})	
		2024	2025	2024	2025
resnet50	L2	0.452	0.111	0.320	0.091
	L3	0.348	0.164	~~0~~	~~0~~
	L4	0.603	0.400	**0.260**	**0.206**
BEIT	B5	0.788	0.596	0.604	0.498
	B7	0.790	0.614	0.619	0.517
	B9	0.731	0.500	0.348	0.277
	B11	**0.601**	**0.387**	**0.251**	**0.189**
DEIT	B5	0.950	0.771	0.544	0.382
	B7	0.728	0.539	**0.283**	**0.233**
	B9	**0.454**	**0.351**	0.092	0.079
	B11	0.348	0.164	~~0~~	~~0~~
ViT	B5	0.944	0.741	0.599	0.519
	B7	0.856	0.665	0.587	0.457
	B9	**0.742**	**0.581**	0.380	0.313
	B11	0.348	0.164	0.011	0.009
dinov2	B5	0.973	0.752	0.771	0.635
	B7	0.931	0.756	0.362	0.251
	B9	**0.562**	**0.308**	0.058	0.048
	B11	0.348	0.164	~~0~~	~~0~~

4. Discussion

A result emerging from the present analysis is given by the fact that images collected during mechanical harvesting possess sufficient information to estimate yield without the need to identify objects (grapes or berries). The pre-trained embedders appeared to be able to extract features from complex scenes, though (to date) without an explanatory approach.

The classical CNN approach (ResNet50) remains competitive; however, a semantic interpretation of the scene is shown by blocks 9, and 11 using Vision Transformers.

However, other scientific domains are trying embedding, from medicine (radiology) to earth observation.

However, given the novelty of the approach, a comparison with previously published models based on similar methods is challenging—most image-based grape yield estimation studies focus on pre-harvest crop assessment from static vineyard images.

Mohimont et al. [33] reviewed computer vision approaches for grape yield estimation, reporting that published image-based methods achieve coefficients of determination up to approximately $R^2 \approx 0.95$, depending on the sensing modality and prediction task. Several approaches aim at simplifying the survey technique (e.g., [34]), as we did through the use of off-the-shelf cameras. In a growing number of cases, simplification is achieved through robotization while adding complexity (e.g., [35]).

All of the mentioned works are based on pre-harvest surveys; therefore, the comparison should be interpreted as a comparison of regression performance rather than a comparison between identical prediction tasks.

To the best of our knowledge, few studies have addressed image-based biomass-flow estimation during mechanical grape harvesting on a harvester conveyor under dynamic operating conditions and subsequently derived linear yield (kg m^{-1}). Also, for monitoring

systems based on load cells, pressure sensors, or volumetric measurements, the focus is on spatial variability, which reached $R^2 \approx 0.54–0.87$ in Palacios et al. [36] and $R^2 \approx 0.73$ in Diago et al. [37].

Because of the objective of the study (proof of concept), the solution adopted is currently based on off-the-shelf cameras devoted just to collecting images—no real-time processing or integration with the vehicle monitoring system (e.g., ISO-bus) has been approached. Therefore system represents a starting point for the development of an integration, aimed at representing an alternative to existing typologies of sensors.

5. Conclusions

Precision viticulture increasingly requires affordable sensing solutions providing high-resolution yield information to support site-specific vineyard management and improve both productivity and resource-use efficiency.

The present study demonstrates that image-based yield estimation during mechanical harvesting can be effectively addressed. Although the proposed methodology should currently be regarded as a proof of concept, it demonstrates that generic visual representations learned from large image collections can successfully encode information related to grape biomass flow, even without being specifically trained for viticulture.

Despite the intrinsic complexity of the harvested biomass—comprising berries, grape clusters, leaves, shoots, juice, motion blur and moving mechanical components—the extracted embedding contained sufficient information to achieve accurate yield estimation without requiring dedicated sensors, radiometric calibration, manual annotation or model fine-tuning.

The principal scientific contribution of this study is the demonstration that general-purpose visual embeddings extracted from pre-trained deep learning models provide an effective representation for grape yield estimation during mechanical harvesting under the investigated experimental conditions. Unlike conventional approaches based on task-specific network training, the proposed framework exploits pre-trained feature extractors without fine-tuning, while sparse fusion regularization allows the selection of features that remain consistent across the two monitored seasons, thereby improving model stability under the investigated conditions. These findings suggest that foundation visual models represent a promising basis for developing scalable, low-cost vision-based yield sensing systems for precision viticulture.

An important outcome of this work is that the predictive capability did not depend exclusively on the most recent transformer architectures. While Vision Transformers provided competitive results, the ResNet50 backbone achieved comparable performance, suggesting that biomass-flow estimation relies strongly on local visual characteristics, such as texture, color distribution and object density, rather than exclusively on high-level semantic representations. This observation indicates that relatively lightweight architectures may still represent an effective solution for real-time applications on commercial harvesters.

The comparison between seasons also highlighted that predictive accuracy alone is insufficient to assess the stability and reproducibility of feature selection. Features selected independently through the standard LASSO showed limited consistency between seasons, indicating that many descriptors may be dataset-specific and susceptible to overfitting. Conversely, the Sparse Fusion LASSO promoted the identification of a smaller subset of features consistently selected across seasons, providing a more reliable basis for model development under similar acquisition conditions. Features selected independently through the standard LASSO showed very limited consistency between years, indicating that many descriptors are dataset-specific and potentially affected by overfitting. Conversely, the Sparse Fusion LASSO approach identified a smaller set of stable features shared across

seasons, providing a more reliable basis for cross-season model transferability. These findings underline that controlling both the number and the stability of the selected features is essential for developing image-based yield estimation systems capable of generalizing across different harvesting conditions.

A limit of the study is that all experiments were performed in a single commercial vineyard using only one grape cultivar (Merlot), one harvesting machine and one acquisition setup. While this design is appropriate for a proof-of-concept study, it does not fully support several statements regarding the transferability and general applicability of the proposed methodology. Environmental conditions, canopy architecture, berry morphology, harvesting settings and lighting conditions can differ substantially across vineyards and cultivars.

A further methodological limitation is the lack of validation on an independent dataset. Although the proposed feature-selection strategy identified descriptors that remained consistent across the two monitored seasons, the selected feature sets should currently be regarded as exploratory. Independent validation on larger datasets collected under different vineyard, cultivar and harvesting conditions will be necessary to assess the reproducibility and practical applicability of the proposed models.

Future work should therefore include validation on independent datasets collected across multiple vineyards, grape cultivars, canopy architectures, maturity stages and harvester configurations in order to assess the reproducibility of the selected features and the broader applicability of the proposed methodology. Future research should also investigate larger vision models operating at higher image resolutions, a more systematic analysis of intermediate embedding layers, alternative sparse regression and feature-selection methods, temporal models capable of exploiting image sequences rather than individual frames, and the optimization of camera placement and acquisition parameters.

Of non-negligible importance is considering the presence of blurred components in the processed images, and the use of versions of training datasets including artificial corruptions of images, such as ImageNet-C, should certainly be taken into consideration, together with specific tuning.

Last but not least, the results are strongly dependent on the (LASSO) parameters used to identify the proper features. A future effort will be focused on making the features explainable, in order to eventually introduce fine-tuning or drive active learning algorithms. These developments will be necessary to improve both prediction accuracy and the robustness of image-based yield estimation systems for precision viticulture.

The assembly of devices can already be used for mapping vineyard yield, and in the future it may be upgraded onto smart cameras for real-time integration with harvester monitoring systems.

Author Contributions: Conceptualization, G.V., C.F. and M.G.; methodology, G.V. and M.A.; software, G.V.; validation, G.V., F.T., F.P. and M.G.; formal analysis, G.V., C.F. and M.A.; investigation, G.V., F.T., F.P. and M.G.; resources, M.M. and M.G.; data curation, G.V. and M.A.; writing—original draft preparation, G.V. and C.F.; writing—review and editing, G.V., C.F. and M.G.; visualization, G.V. and M.A.; supervision, G.V. and M.G.; project administration, M.G.; funding acquisition, M.M. and M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research is part of the project Precision Tools and Technologies for Monitoring and Sustainable Management of Vineyard Variability—SPRINT, co-funded by the Emilia-Romagna Region under the PR FESR Emilia-Romagna 2021–2027 Programme, Action 1.1.2.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are not available due to being part of ongoing research, but will be made available upon reasonable request.

Acknowledgments: The authors want to thank the "Passanti Massimo" Estate for hosting the trial and the field crew for technical assistance, Appleby Italiana for providing the vigor maps, and PRO.PA.R. (Soc. Coop. Agricola, Fornace Zarattini, RA, Italy) for performing mechanical harvesting. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare that the "Passanti Massimo" Estate, Appleby Italiana, and PRO.PA.R. Soc. Coop. Agricola were not involved in the study design, the data collection planning, the statistical analysis, the interpretation of the results, the writing of the manuscript, or the decision to submit the manuscript for publication. The authors declare no other commercial or financial relationships that could be construed as potential conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial neural network
BEiTv2	Bidirectional Encoder representation from Image Transformers
CNN	Convolutional neural network
DeiT	Data-efficient Image Transformer
DINOv2	Self-supervised Vision Transformer model
fps	Frames per second
LASSO	Least Absolute Shrinkage and Selection Operator
PCA	Principal Component Analysis
ResNet	Residual Neural Network
RGB	Red–Green–Blue
R^2	Coefficient of determination
SFL	Sparse Fusion LASSO
UAV	Unmanned Aerial Vehicle
ViT	Vision Transformer
β	Regression coefficient
λ	LASSO regularization parameter
O	Objective function
x	Explanatory variable/extracted feature
y	Observed yield value
y'	Estimated yield value

References

1. Bramley, R.G.V.; Trought, M.C.T.; Praat, J.P. Vineyard variability in Marlborough, New Zealand: Characterising variation in vineyard performance and options for the implementation of precision viticulture. *Aust. J. Grape Wine Res.* **2011**, *17*, 72–78. [CrossRef]
2. Matese, A.; Di Gennaro, S.F. Technology in precision viticulture: A state-of-the-art review. *Int. J. Wine Res.* **2015**, *7*, 69–81. [CrossRef]
3. Squeri, C.; Poni, S.; Di Gennaro, S.F.; Matese, A.; Gatti, M. Comparison and ground truthing of different remote and proximal sensing platforms to characterize variability in a hedgerow-trained vineyard. *Remote Sens.* **2021**, *13*, 2056. [CrossRef]
4. Tardaguila, J.; Stoll, M.; Gutiérrez, S.; Proffitt, T.; Diago, M.P. Smart applications and digital technologies in viticulture: A review. *Smart Agric. Technol.* **2021**, *1*, 100005. [CrossRef]
5. Draigneau, C. *Precision Viticulture: Tools to Measure and Manage Vineyard Variability*; School of Geosciences, University of Aberdeen: Aberdeen, UK, 2009.
6. PFG Australia. *Gregoire Grape Harvesters*; PFG Australia: Melbourne, Australia, 2017. Available online: https://www.pfgaustralia.com.au/gregoire/ (accessed on 18 June 2026).

7. New Holland Agriculture. *Braud High and Extra High Capacity Grape Harvesters: 9090L and 9090X*; New Holland Brand Communications: Auckland, New Zealand, 2019. Available online: https://agriquip.co.za/wp-content/uploads/2025/10/braud-high-and-extra-high-capacity-grape-harvesters-brochure-en.pdf (accessed on 18 June 2026).
8. Taylor, J.A.; Sanchez, L.; Sams, B.; Haggerty, L.; Jakubowski, R.; Djafour, S.; Bates, T.R. Evaluation of a commercial grape yield monitor for use mid-season and at harvest. *J. Int. Sci. Vigne Vin* **2016**, *50*, 57–63. [CrossRef]
9. Płaskowski, A.; Beck, M.S.; Thorn, R.; Dyakowski, T. *Imaging Industrial Flows: Applications of Electrical Process Tomography*; CRC Press: Boca Raton, FL, USA, 1995.
10. Helian, B.; Huang, X.; Yang, M.; Bian, Y.; Geimer, M. Computer vision-based excavator bucket fill estimation using depth map and faster R-CNN. *Autom. Constr.* **2024**, *166*, 105592. [CrossRef]
11. Sabih, M.; Farid, M.S.; Ejaz, M.; Husam, M.; Khan, M.H.; Farooq, U. Raw material flow rate measurement on belt conveyor system using visual data. *Appl. Syst. Innov.* **2023**, *6*, 88. [CrossRef]
12. Laurent, C.; Oger, B.; James Arnold Taylor, J.A.; Scholasch, T.; Metay, A.; Tisseyre, B. A review of the issues, methods and perspectives for yield estimation, prediction and forecasting in viticulture. *Eur. J. Agron.* **2021**, *130*, 126339. [CrossRef]
13. Asadullayev, R.; Abasova, K.; Mammadova, K.; Guliyeva, A. Assessment of grape cultivar traits and environmental adaptation in the Absheron Peninsula, Azerbaijan. *DYSONA–Appl. Sci.* **2025**, *6*, 51–59. [CrossRef]
14. Gatti, M.; Zamboni, M.; Merli, M.C.; Civardi, S.; Poni, S. Characterization of the new cultivar Ervi as a function of mechanical harvesting. *Acta Hortic.* **2013**, *978*, 229–236. [CrossRef]
15. Intrieri, C. Research and innovation for full mechanization of Italian vineyards at Bologna University. *Acta Hortic.* **2013**, *978*, 151–168. [CrossRef]
16. Archana, R.; Jeevaraj, P.S.E. Deep learning models for digital image processing: A review. *Artif. Intell. Rev.* **2024**, *57*, 11. [CrossRef]
17. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in vision: A survey. *ACM Comput. Surv.* **2022**, *54*, 1–41. [CrossRef]
18. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021*; IEEE: New York, NY, USA, 2021; pp. 9650–9660.
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016*; IEEE: New York, NY, USA, 2016; pp. 770–778.
20. Schneider, S.; Rusak, E.; Eck, L.; Bringmann, O.; Brendel, W.; Bethge, M. Improving robustness against common corruptions by covariate shift adaptation. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 11539–11551.
21. Gatti, M.; Dosso, P.; Maurino, M.; Merli, M.C.; Bernizzoni, F.; Pérez, J.; Platè, B.; Bertuzzi, G.C.; Poni, S. MECS-VINE®: A new proximal sensor for segmented mapping of vigor and yield parameters on vineyard rows. *Sensors* **2016**, *16*, 2009. [CrossRef] [PubMed]
22. Gatti, M.; Rahal, K.; DAmbrosio, R.; Pelusi, F.; Bonini, P.; Poni, S. Seasonal interactions between spatial and temporal variability in a steep Barbera vineyard. *Precis. Agric.* **2026**, *27*, 16. [CrossRef]
23. Caprara, C.; Pezzi, F. Measuring the stresses transmitted during mechanical grape harvesting. *Biosyst. Eng.* **2011**, *110*, 97–105. [CrossRef]
24. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
25. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers distillation through attention. In *Proceedings of the International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021*; PMLR: New York, NY, USA, 2021; pp. 10347–10357.
26. Peng, Z.; Dong, L.; Bao, H.; Ye, Q.; Wei, F. BEiT v2: Masked image modeling with vector-quantized visual tokenizers. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.
27. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning Robust Visual Features without Supervision. *Trans. Mach. Learn. Res.* **2024**. [CrossRef]
28. Chefer, H.; Gur, S.; Wolf, L. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021*; IEEE: New York, NY, USA, 2021; pp. 782–791.
29. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58*, 267–288. [CrossRef]
30. Simon, N.; Friedman, J.; Hastie, T.; Tibshirani, R. A sparse-group lasso. *J. Comput. Graph. Stat.* **2013**, *22*, 231–245. [CrossRef]
31. Tibshirani, R.; Saunders, M.A.; Rosset, S.; Zhu, J.; Knight, K. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2005**, *67*, 91–108. [CrossRef]
32. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009.

33. Mohimont, L.; Alin, F.; Rondeau, M.; Gaveau, N.; Steffenel, L.A. Computer Vision and Deep Learning for Precision Viticulture. *Agronomy* **2022**, *12*, 2463. [CrossRef]
34. Parr, B.; Legg, M.; Alam, F. Grape yield estimation with a smartphone's colour and depth cameras using machine learning and computer vision techniques. *Comput. Electron. Agric.* **2023**, *213*, 108174. [CrossRef]
35. Devanna, R.P.; Romeo, L.; Reina, G.; Milella, A. Yield estimation in precision viticulture by combining deep segmentation and depth-based clustering. *Comput. Electron. Agric.* **2025**, *232*, 110025. [CrossRef]
36. Palacios, F.; Diago, M.P.; Melo-Pinto, P.; Tardaguila, J. Early yield prediction in different grapevine varieties using computer vision and machine learning. *Precis. Agric.* **2023**, *24*, 407–435. [CrossRef]
37. Diago, M.-P.; Correa, C.; Millán, B.; Barreiro, P.; Valero, C.; Tardaguila, J. Grapevine Yield and Leaf Area Estimation Using Supervised Classification Methodology on RGB Images Taken under Field Conditions. *Sensors* **2012**, *12*, 16988–17006. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.