JOURNAL OF WATER AND LAND DEVELOPMENT

e-ISSN 2083-4535

Polish Academy of Sciences (PAN) Institute of Technology and Life Sciences – National Research Institute (ITP – PIB)

JOURNAL OF WATER AND LAND DEVELOPMENT
DOI: 10.24425/jwld.2026.157835
2026, No. 68 (I–III): 196–206

Architecture-based machine learning models for peach yield prediction before bloom

Abderrahim Zegoumou*[1)], Mohammed Ibriz[1)], Zakariae El Housni[2)], Badr Bounsir[1)], Ayoub Ba-ichou[3)], Chaymaa Lamini[3)], Abdelaziz Ait Elkassia[4)], Reda Meziani[5)], Hicham Bouzelmate[6)], Mustapha Fagroud[7)]

[1)] Ibn Tofail University, Faculty of Sciences, Laboratory of Vegetal, Animal and Agro Productions Industry, University campus, P.O. Box 133, 14000, Kenitra, Morocco

[2)] Moulay Ismail University, Meknes Faculty of Sciences, Department of Biology, Laboratory of Biotechnology and Molecular Biology, P.O. Box 11201, Zitoune, 50000, Meknes, Morocco

[3)] Moulay Ismail University, Meknes Faulty of Sciences, Department of Computer Science, Lab TSI, P.O. Box 11201, Zitoune, 50000, Meknes, Morocco

[4)] ENSAM of Rabat, Mohamed V University, Ave des Forces Armées Royales, P.O. Box 6207, 10100, Rabat, Morocco

[5)] National Institute for Agronomic Research, CRRA, km 10, Haj Kaddour Rd, P.O. Box 578, 50000, Meknes, Morocco

[6)] Moulay Ismail University, Faculty of Sciences and Technology, Biodiversity, Environment and Plant Protection Team, National Rd No. 13 (RN13), 52000, Errachidia, Morocco

[7)] National School of Agriculture of Meknes, Department of Sciences and Techniques in Plant Production, km 10, Haj Kaddour Rd, P.O. Box S/40, 50001, Meknès, Morocco

* Corresponding author

RECEIVED 21.06.2025 ACCEPTED 29.09.2025 AVAILABLE ONLINE 18.03.2026

Abstract: For fruit crop management to be optimised, early yield prediction is essential. In order to predict peach yields ('N48-52') across four tree levels using non-destructive, pre-bloom architectural measurements, this study assesses three machine learning (ML) models: random forest (RF), extreme gradient boosting (XGBoost), and support vector machines (SVM). Structural dimensions, fruit count, and weight were among the data gathered in 2019, 2021, and 2022. In order to train the models, 80% of the dataset was used, and the remaining 20% was used for validation. According to the results, SVM performed best for the 3rd level (coefficient of determination (R^2) = 0.91), while RF was the most accurate for the 1st, 2nd, and 4th levels (R^2 = 0.79, 0.91, and 0.93, respectively). To further enhance the accuracy of the proposed models, additional data points were randomly collected from different trees in 2023. These data included measurements of the complete path to a given level along with its fruit production, allowing for verification of the precision and stability of the proposed models. In 2023, the models maintained an accuracy of R^2 = 0.9054, 0.7684, 0.8768, and 0.7964, respectively, for RF (1st level; 2nd level; 4th level) and SVM (3rd level). A comparison between the estimated production from the trees and the actual production showed a statistically similar result (accepted statistical error for the analysis of variance statistical test (α = 0.05).

Keywords: before bloom, machine learning, peach, tree architecture, yield prediction

INTRODUCTION

Data utilisation in agriculture, particularly in early yield estimation, presents several challenges. Researchers aim to identify the most robust yield prediction model using a minimal number of input features, raising the complex issue of data dimensionality reduction to enhance modelling efficiency while saving data collection costs (He *et al*., 2022). Yield estimation through fruit counting presents a major obstacle due to the need for fruit formation, while estimation is more valuable for improving size, managing labour, estimating cold storage space requirements (Wang *et al*., 2013), and selling or exporting produce (Nuske *et al*., 2014). Moreover, the sampling procedure has limitations, compromising the accuracy of yield estimation for the entire orchard (Payne *et al*., 2013). Additionally, spring frosts in fruit-growing regions pose problems, necessitating damage estimation by comparing the amount of fruit remaining on the tree with what would have been produced under normal conditions (Jiménez and Díaz, 2003).

The trend for early yield estimation has primarily relied on image processing through fruit recognition using RGB cameras for ease of use and reduced investment costs (Gongal *et al*., 2015; Kamilaris and Prenafeta-Boldú, 2018). However, machine learning has recently become more prevalent in the agricultural sector, including yield estimation, weed detection and classification, quality and disease estimation (Liakos *et al*., 2018). It is considered to provide more accurate results than previous image processing techniques (Kamilaris and Prenafeta-Boldú, 2018).

Over time, the development of early yield estimation methods in agriculture, particularly direct methods, has seen significant advancements. Studies have shown that canopy structure and tree vegetation index are correlated with final yield, providing indicators for predicting orchard yields (Matese and Di Gennaro, 2021). Remote sensing in the visible and near-infrared spectrum has been widely used to obtain multiple vegetation indices, enabling fruit growth monitoring and yield estimation (Serrano, González-Flor and Gorchs, 2012; Hacking *et al*., 2019). An innovative approach involves directly detecting and counting fruits on trees using image processing methods and classifiers based on machine learning algorithms that recognise fruits based on their colour, shape, and texture characteristics (Liu *et al*., 2018; Xu *et al*., 2019). Computer vision technologies have been extensively explored for automatic yield mapping of fruit and vegetable crops (Darwin *et al*., 2021). Recent studies highlight that vision systems dominate current research in orchard yield assessment, with models and artificial vision technologies constantly evolving (Anderson, Walsh and Wulfsohn, 2021). These advancements have led to a diversification of application platforms for direct yield estimation, including manual ground-based approaches, vehicle-mounted platforms (robots, tractors, quads, etc.), and unmanned aerial vehicle platforms, each tailored to specific applications (He *et al*., 2022).

The development of indirect methods for early yield estimation in agriculture has been marked by a diversity of approaches. Studies have established close relationships between tree size and trunk cross-sectional area (TCA). The TCA is used to compare different plots based on parameters such as vigour, yield, and flower load (Dennis, Masabni and Ketchie, 1996; Caruso *et al*., 1999; Strong and Azarenko, 2000). Flower density, expressed as a function of the number of flowers per TCA, has been shown to be an accurate estimate of bud load, influencing the productivity of fruit trees (Chang, Iezzoni and Flore, 1987; Kappel, 1990). The correlation between shoot length and flower density has been favoured in the case of stone fruit trees (Marini and Sowers, 2000). Studies have also revealed that stem TCA is closely related to the number of flowers and leaves on a tree, providing a more manageable alternative to controlling the number of individual flowers or leaves (Murray, 1927). Models have been developed to predict harvest density and fruit weight, using factors such as TCA·ha^{-1}, estimated total shoot length per trunk cross-sectional area (SLT, m of shoots per cm^2 TCA), and days between full bloom and harvest for the peach (BHP) (Jiménez and Díaz, 2003). The mechanical and conductivity properties of shoots have been found to have a direct relationship with branch thickness and also an impact on production (Cannell and Morgan, 1987; Upadhyaya, Cooke and Rand, 1987; Tyree *et al*., 1991). Linear, multiple, and partial least squares regression methods have been used to establish linear relationships between different input data and predicted orchard yields (Beek van *et al*., 2015; Underwood *et al*., 2016; Anderson *et al*., 2019; Wu *et al*., 2021). Additionally, the integration of machine learning techniques, such as artificial neural networks (ANN) and support vector machines (SVM), has been adopted to improve model accuracy, offering increased capacity to analyse complex and non-linear data. This shift towards more sophisticated methods highlights the diversity of approaches for early yield estimation, combining traditional agronomic knowledge with modern technological advancements (He *et al*., 2022).

Direct and indirect methods for early yield estimation have various imperfections. For direct methods, the complexity of the original datasets, often rich in environmental and plant information, can lead to the presence of redundant or irrelevant features to the target output (Maimaitiyiming *et al*., 2019; Ballesteros *et al*., 2020). Selecting relevant features becomes crucial, involving correlation analysis between each feature and yield change, often requiring intervention from agronomic experts (He *et al*., 2022). Moreover, fruit occlusion by elements such as branches or other fruits presents a frequent challenge in direct estimation, limiting the effectiveness of correction factors (Apolo-Apolo *et al*., 2020).

On the other hand, methods based on spectral information obtained through remote sensing of orchards face difficulties related to the variability of climatic conditions and plant growth (Beek Van *et al*., 2015; Sirsat *et al*., 2019). The correlation between vegetation indices (VI) and orchard production varies over growth stages, requiring careful selection of the spectral data observation time for acceptable indirect prediction (Beek van *et al*., 2015; Sirsat *et al*., 2019). Furthermore, estimation platforms based on vision systems are susceptible to complex environmental conditions, such as variations in light intensity, introducing a potential source of error (He *et al*., 2022).

In summary, direct methods may face challenges related to data complexity and fruit occlusion, while indirect methods must contend with the temporal variability of spectral information and sensitivity to environmental conditions. Combining these methods, while considering agronomic expertise and specific conditions, can offer a more robust approach for early yield estimation.

The primary objective of this work is to achieve a reliable early yield estimation based on machine learning techniques fed by non-destructive measurements (length and diameter) of different tree structures before flowering.

MATERIALS AND METHODS

DATA COLLECTION

To achieve the objective of characterising the structure and architecture of peach trees based on various architectural parameters, measurements were carried out in the Louata agricultural domain (Sefrou). The GPS coordinates of the plot are: 33°54'21.8"N 4°39'59.8"W. The site is characterised by a semi-arid Mediterranean climate, hot and dry in summer and rainy in winter. The area is subject to adverse climatic events for agricultural production, such as hail, frost, and "chergui" (hot wind) events. The hills are mainly affected by hail at the beginning of winter (October) and spring (March), with risks of frost in winter (January). Annual rainfall varies according to the Figure 1b.

During the experimental seasons (2018/2019, 2020/2021, 2021/2022), the average daily per month temperatures recorded at the site varied according to the graphs (Fig. 1a). The season 2019/2020 wasn't considered for the restrictions due to COVID-19 disease.

The irrigation water has an *EC* that is characterised by good quality ($S \cdot m^{-1}$). Regarding fertilisation, soil, water, and leaf analyses are carried out annually after entry into production to determine the status of the soil (correction and maintenance). The distribution of fertiliser applications is based on the phenological stages of the tree.

The experiment was conducted on a homogenous plot of seasonal peach trees ('N48-52' grafted on rootstock GF677), planted in 2009 with a density of 556 trees per ha (3 × 6 m) using a standard goblet training system with four scaffolds. For monitoring the trial, trees were selected randomly (randomised complete block design) after excluding border trees (the first row is 35 m from the borders and the first tree is 21 m from the borders), also, other criteria were used, like visual health, vigour

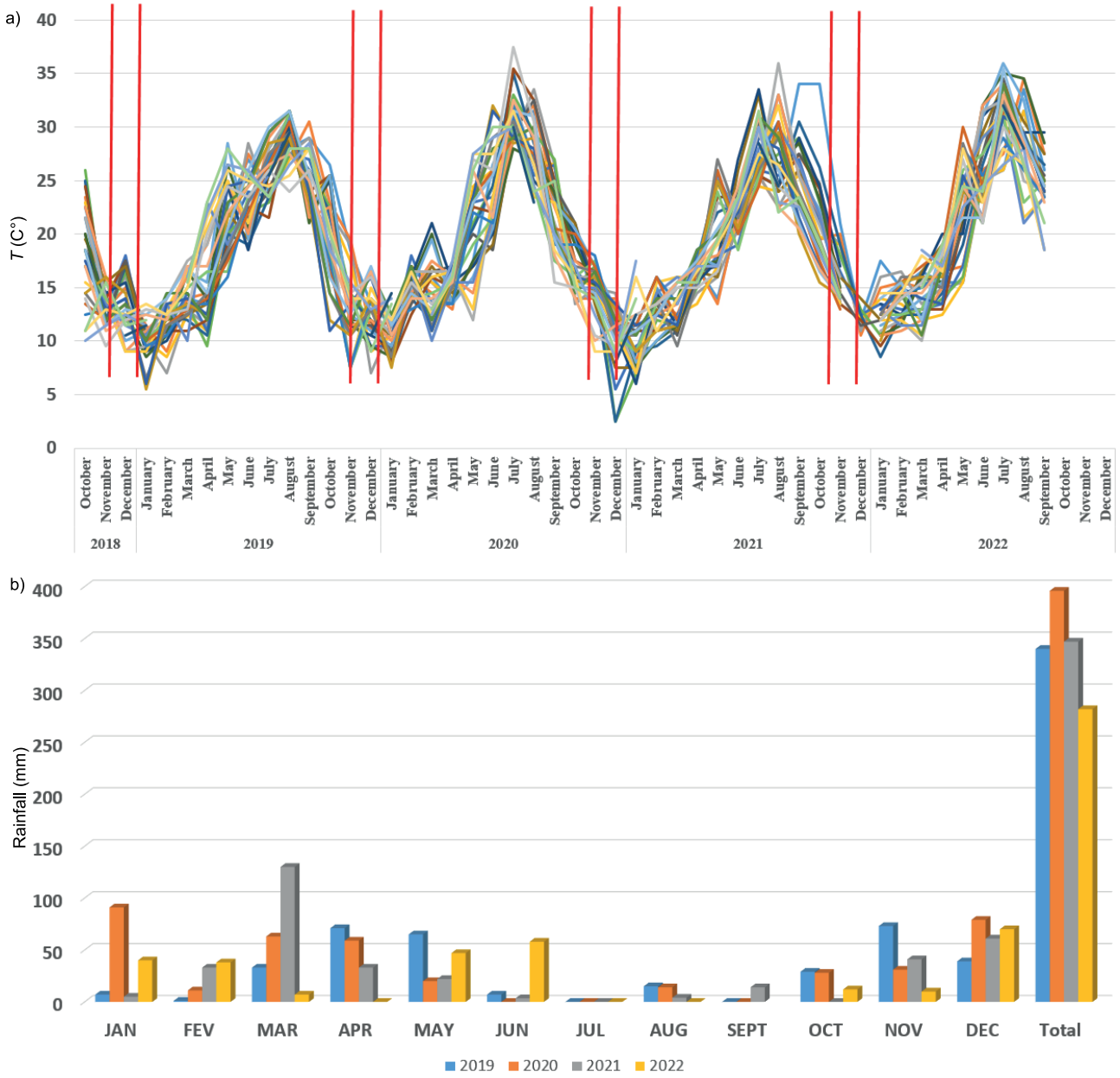

Fig. 1. Trial field evolution of: a) daily average temperature, b) rainfall; the period between two red vertical lines marks the measurement period in the trial field; source: own elaboration

homogeneity to choose representative individuals. The trees are marked with white adhesive tape at the base of the trunk to facilitate their identification.

Measurements were taken immediately after winter pruning at the same stage (according to the BBCH[1] = 0) using a digital calliper (error = 10^{-3} mm). Three parameters were measured each time: trunk circumference, diameter at the base of all organs and branches of the tree, their lengths (with a tape measure), and the distances of their insertions. All the pruning and thinning practices were the same for all the trees in the experimental plot. The fruit thinning was performed 46 days after full bloom.

In a recent study Laurent *et al.* (2021) noted that the timing of the observation of characteristics used for yield prediction modelling should be determined by the growth status of the fruit trees, rather than by a fixed date for each year.

These measurements were taken in the following order of branching: starting with the trunk circumference and its length, then the diameter at the base of the scaffold and its length up to the insertion of the first sub-scaffold or a branch, and so on until the last branch, for which the base diameter and length were also measured. Then, we move to the second scaffold located on the left (clockwise direction) and so on until all four scaffolds of the tree are completed.

All branches, including mixed branches identified for the trees, were labelled and numbered with black, resistant adhesive tape (Photo 1). The total number of structures is shown in the Table 1.

While other studies have been conducted with reduced total number of branches, Jiménez and Díaz (2003) marked 30 trees but only measured 20 shoots. Similarly, Planchon, Claustriaux, and Crabbé (2003) reduced the number of trees in their study from 70 one-year-old trees to just 5 five-year-old trees during the fifth year of the study.

At harvest, the number of fruits and their weight were recorded for each level of the marked trees to establish the relationship between the different variables measured and tree production.

MACHINE LEARNING MODELS

Using the data collected from different tree structures, the study aims to utilise various machine learning models to predict the production of each tree level with the highest possible accuracy. The chosen tree levels are levels 1 to 4, excluding the base level (trunk) and the 5th (5th sub-scaffold) and 6th level (mixed branches).

Data preparation and processing

The dataset used in this study concerned peach 'N48-52' measurements, such as length and the circumference of various tree structures, from the trunk to the mixed branches, and the corresponding number and weight of fruits for each level.

We have a dataset containing measurements taken on all tree structures over three seasons (2018/2019, 2020/2021, 2021/2022).

Photo 1. Mixed branches: a) labelled, b) diameter measurement with digital calliper (phot.: *A. Zegoumou*)

Table 1. Number of tree structure measured for the entire field trial

Year	Tree No.	Scaffold	1st level	2nd level	3rd level	4th level	5th level	MB
2019	5	20	102	75	21	4	0	491
2021	10	38	315	324	196	73	14	1353
2022	10	39	339	449	282	110	25	1328
2023	10	40	40	37	22	11	6	40
Total	35	137	796	885	521	198	45	3212

Explanations: MB = mixed branches.
Source: own study.

This data was represented in CSV format and structured hierarchically, with one file per year and per structural level of the tree. For the season 2022/2023, 10 randomly selected trees different from the other trees chosen later were used as a basis for

[1] BBCH-scale – De.: Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie – standard scale designed to identify the phenological stages of plant development

collecting 40 complete data paths (length and diameter) of the different tree levels (from the trunk to the mixed branches) coupled with the fruit yield in grams to provide additional data for verifying the robustness and stability of the prediction models obtained at the end of data processing for the three earlier seasons.

The values collected in each record are annotated according to the tree measurements. Each row of these records presents a set of values, ranging from the base of the tree trunk to the fruit, as a complete measurement path.

Data preprocessing

Data pre-processing has a fundamental role in developing high-performance predictive models from datasets. It encompasses a process of cleaning, transforming, and normalising raw data to adapt it to the context of machine learning.

First, we will undertake uniform data structuring. From the datasets presented previously, we will generate new sets that group data of the same nature, starting with fruit weight data related to the 1st level (data collected on length and diameter from the trunk to the first sub-scaffold), then fruit weight data from the 2nd level (data collected on length and diameter from the trunk to the second sub-scaffold), and so on until the 4th level. The 5th level was not included due to the limited data available at this level.

In the case of our dataset, the initial pre-processing step involves obtaining a unique result per data path provided, i.e., the data feeding the machine should have a unique total fruit weight result present at that specific level. This approach aims to focus exclusively on the data relevant to our predictive model. Regarding fruit weight data, namely ['PFr1', 'PFr2', 'PFr3', 'PFr4', 'PFr5', 'PFr6'], the objective is to consolidate this data into a single data point called total fruit weight (PTF – Fr.: Poids total des fruits).

The importance of handling missing values in data cannot be overstated in the field of data analysis and machine learning. The presence of missing values can lead to biases, decreased model accuracy, and loss of potentially valuable information. Therefore, choosing the right imputation technique to handle these missing values is crucial to ensure reliable and accurate results in data analysis (Rakićević, Savić and Bulajić, 2016). To address the problem of missing data for fruit weight, several imputation techniques were tested, including SimpleImputer using "mean" and "median" replacement strategies, "KNNImputer" using the nearest neighbour approach, and "IterativeImputer" using a multivariate method that utilises missing values as a function of other predicted values via round-robin regression. The performance evaluation of the imputation methods was based on the accuracy achieved by the machine learning model with each method.

Machine learning models

All machine learning models were built using the Python programming language, incorporating libraries such as Scikit-learn, extreme gradient boosting (XGBoost), TensorFlow, and Keras. The use of a graphics processing unit (GPU) was essential for model training. The model hyperparameters were meticulously determined through a comprehensive grid search process. To achieve this, the dataset (3,183 complete paths) was divided into two distinct subsets. The first subset, constituting 80% of the data and randomly selected (2,546 complete paths), was designated as the training subset and used for model training. The remaining 20% (637 paths) were allocated to the validation subset, which played a crucial role in hyperparameter optimisation. Consequently, hyperparameter values were selected with the primary goal of minimising errors within the validation subset (Yu and Zhu, 2020). In the final phase, the model underwent rigorous testing using data from the designated test set, in accordance with the chosen strategies.

Random forest (RF) is a decision tree-based algorithm originally introduced by Breiman (2001). This model aggregates numerous decision trees, each finely tuned to different subsets of the training data. Individually, each tree can be considered a modest or weak learner, but when merged as an ensemble, they result in a single model with robust predictive capabilities, as explained by Gao *et al.* (2021). Another strength of RF lies in its ability to assess the importance of each input variable. Regarding the training procedure, RF generally requires minimal hyperparameter tuning. In the context of this particular study, the main hyperparameters examined include the number of trees (n_estimators), the number of features considered for splitting at each leaf node (max_features), and the tree depth (max_depth). The values obtained were subjected to hyperparameter testing: for the aforementioned hyperparameters (100, 200, 400, and 500 trees), (all available features), and (4, 5, 6, 7, 8) were explored, respectively (extreme gradient boosting (XGBoost)).

XGBoost, a relative newcomer to the field, was introduced by Chen and Guestrin (Chen and Guestrin, 2015) and has since gained considerable attention in machine learning. Like random forest (RF), this algorithm relies on decision trees, but it differs in how it assembles its tree ensemble. The XGBoost capitalises on established concepts of gradient boosting and builds each tree by leveraging information derived from previously created trees. The XGBoost, similar to the RF model, has the power to evaluate the importance of input variables, which improves computational efficiency and enhances its ability to combat overfitting. During the hyperparameter optimisation process, adjustments were made to the number of trees (n_estimators) and the maximum tree depth (max_depth). The range of values explored for the number of trees included 100, 200, 400, and 500, while the maximum tree depth was examined over a spectrum of values including 3, 5, 6, 7, 8, 9, and 10 (Chapelle *et al.*, 2002).

Support vector machine (SVM) is a supervised machine learning algorithm that finds utility in both regression and classification tasks. In the realm of SVM for regression, often referred to as support vector regression (SVR) in academic literature, the relationship between the dependent variable (y) and a set of independent variables (x) is discerned through the function $f(x)$. This function is equivalent to linear regression in a high-dimensional feature space, which is diligently crafted through a non-linear mapping of the input space. To facilitate this mapping, kernel functions come into play, notably linear (Lin), polynomial (Poly), and radial basis functions (RBF), all of which serve to transform the input data into the aforementioned high-dimensional feature space. In the context of this particular study, the "linear", "poly", and "rbf" kernels were utilised. Additionally, different variations of the C, ε, and γ parameters were systematically explored to identify the optimal settings within their respective ranges: "linear", "poly", "rbf"; 1.1, 5.4, 170, 1001; 0.1, 0.0003, 0.007, 0.0109, 0.019; 0.7001, 0.008, 0.001 (Hsu, Chang and Lin, 2003; Cherkassky and Ma, 2004).

METRICS

Model accuracy was calculated using the coefficient of determination (R^2) (Eq. 1), mean squared error (*MSE*) (Eq. 2), root mean squared error (*RMSE*) (Eq. 3), and mean absolute error (*MAE*) (Eq. 4). The equations are as follows:

$$R^2 = 1 - \frac{\sum (\hat{y}_i - y_i)^2}{\sum (y_i - \bar{y})^2} \quad (1)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \quad (2)$$

$$RMSE = \sqrt{MSE} \quad (3)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i) \quad (4)$$

where: $\hat{y}_i$ = model predicted value, y_i = observed value, $\bar{y}$ = average of the observed values.

If the *MSE* score value is smaller, it means you are very close to determining the best-fit line, which also depends on the data you are working with, so it is sometimes not possible to obtain a small value of the *MSE* score (Chapelle *et al*., 2002; Thornton, 2014).

The result of the prediction model allows for approximating the total weight of fruit present at a specific level by exploiting the input data (diameter and length of the different structures up to the level in question). Indeed, the training of the machine learning (ML) models involved 80% of the total available information for the three randomly chosen agricultural seasons, while the remaining 20% was used to test the validity of the models. In addition to the 20% of the data used to test the accuracy of the models, specific data points for each level per tree were collected during the 2022/2023 season for a second confirmation step.

STATISTICAL ANALYSIS

A statistical analysis was performed between the numbers of fruits at different levels for each tree and between years to test the effect of tree and agricultural season on the number of fruits per level. Another analysis focused on the effect of the year on production per tree (in g). Since the number of trees differed between years, the analysis involved the minimum number of trees per year, which was 5 trees in 2019. In case of a significant effect, a least significant difference (*LSD*) mean comparison test ($\alpha = 0.05$) was used to compare the different groups. The XLSTAT pro 2016 software (Addinsoft 2016.1) was used for statistical analysis.

RESULTS AND DISCUSSION

RESULTS

With the complete dataset, we can obtain a representation of the average values for all different tree structures (without specifying orientation), ranging from the trunk to the mixed branches (Fig. 2). This figure provides a general overview of the tree shape with the average values of the different structures.

The 1st, 2nd and 3rd tree levels support the essential of the mixed branches (83.46%) that bear fruit (Fig. 3). The 1st and the 2nd levels combined present 64.25% of all mixed branches.

Production evolution

The number of fruits and weight per tree per level and per year, as well as the percentage contribution of each structural level of the tree in relation to the average number and weight of fruits of the trees, are presented in Table S1. The estimated plot production in Mg·ha^{-1}, derived from the production of the selected trees per plot, was tested using analysis of variance (ANOVA) for the five stable trees over the three agricultural seasons (2019/2021/2022), namely PT1 to PT5 (total fruit weight for each tree from 1 to 5).

$$TFWT \cdot NTHa = EP \quad (5)$$

where: *TFWT* = total fruit weight per tree, *NTHa* = number of trees per ha, *EP* = estimated production (Mg·ha^{-1}).

The effect was not significant ($p = 0.884$; $F = 0.332$), indicating that the estimated production from the tree unit is not statistically different from the actual harvested production and that the chosen trees do not produce an outlier production that deviates significantly from the normal production per ha of the studied plot.

The effect of the agricultural season on fruit weight per tree was significant ($p < 0.008$; $F = 3.064$) as well as for the number of fruits per level at the 2nd, 3rd, and 4th levels ($p < 0.055$; $F = 3.722$ / $p < 0.003$; $F = 9.494$ / $p < 0.008$; $F = 7.352$). However, for the 1st level, the change in the agricultural season did not affect the number of fruits borne by this level. The comparison of means via Fisher's least significant difference (*LSD*) test is reported in Table S1.

From Table S1, it can be seen that the estimation of the actual plot production from the production of the sample trees is underestimated by 27% and 18% for the 2018/2019 and 2021/2022 seasons, respectively, while for the 2020/2021 season, the average estimate is overestimated by 10. Based on the average actual production of the three seasons, the production estimation from the tree production becomes more accurate with an underestimation of 24%, 3%, and 13% for 2018/2019, 2020/2021, and 2021/2022, respectively.

Machine learning model results

Accuracy according to missing value estimation method. Information on the accuracy obtained with different methods for estimating missing values for three machine learning models (random forest (RF), extreme gradient boosting (XGBoost), support vector machines (SVM)) are provided in Table 2.

The KNNImputer method was the most accurate across all machine learning (ML) models used. Therefore, the KNNImputer method was adopted to estimate missing values for all machine learning models of the four selected tree levels. The KNNImputer method was more accurate, suggesting that missing values are better estimated by the architectural "proximity" of similar branches, implying strong structural auto-correlation within the tree.

Prediction model results

The accuracy of the different prediction models, as determined by the ML method used, is presented in Table 3, based on 100% of the data.

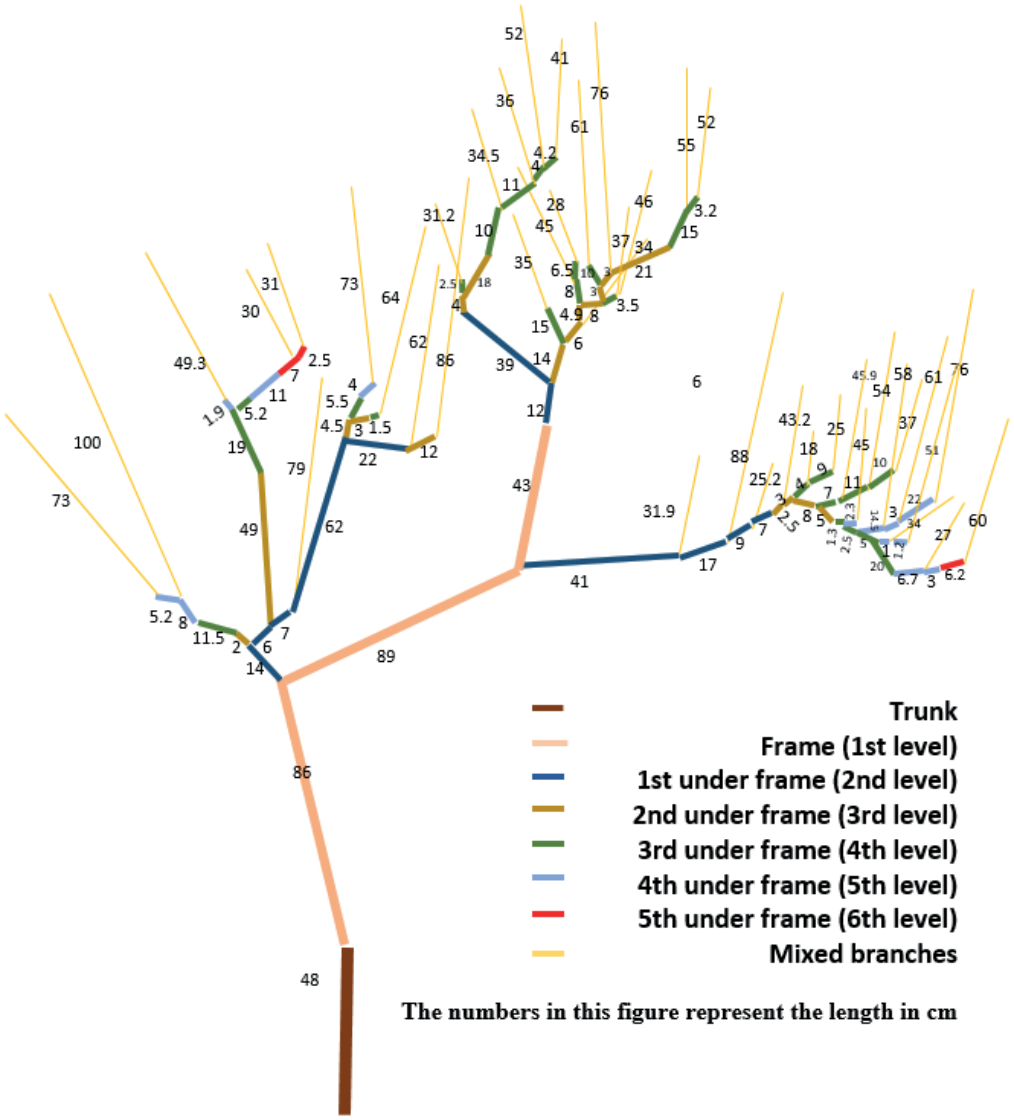

Fig. 2. Length evolution of all tree structure for 'N48-52' peach grafted on rootstock GF677; source: own study

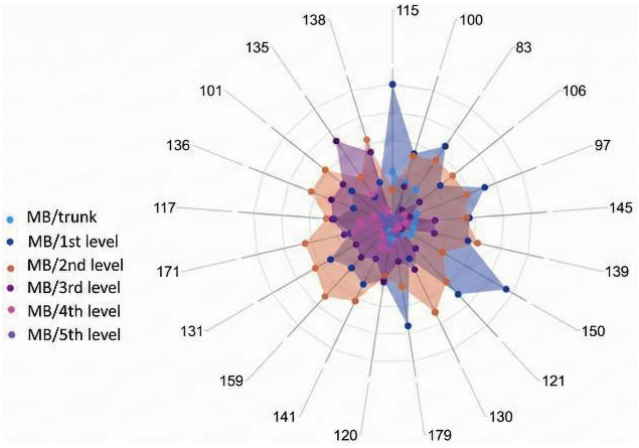

Fig. 3. Average distribution of the mixed branches (MB) on different tree level for the three growing seasons 2019/2020/2021; source: own study

Based on the summarised results in Table 3, it is observed that the RF model exhibits the best accuracy for the 1st, 2nd, and 4th levels with R^2 values of 78.90, 90.77, and 91.16%, respectively. However, for the 3rd level, the SVM model provides the best accuracy with R^2 = 92.64%.

Table 2. Accuracy according to the method used to estimate missing values coupled with different machine learning model

Model	Number of training underframe	R^2
RF	SimpleImputer	0.9004
	KNNImputer	**0.9116**
	IterativeImputer	0.9054
XGboost	SimpleImputer	0.8630
	KNNImputer	**0.8939**
	IterativeImputer	0.7049
SVM	SimpleImputer	0.8569
	KNNImputer	**0.8941**
	IterativeImputer	0.8922

Explanations: R^2 = coefficient of determination, RF = random forest, XGBoost = extreme gradient boosting, SVM = support vector machines, values in bold = the highest values for each machine learning models under different messing values estimation methods.
Source: own study.

Table 3. Random forest (RF), extreme gradient boosting (XGBoost), support vector machines (SVM) machine learning model prediction accuracy per tree level

Model	Training tree level	R^2	*MAE*	*MSE*	*RMSE*
			mm·d^{-1}		
RF	4th	**0.9116**	0.3415	0.1839	0.4288
	3rd	0.9181	0.3323	0.1703	0.4127
	2nd	**0.9077**	0.3183	0.1920	0.4382
	1st	**0.7890**	0.4956	0.4392	0.6627
XGboost	4th	0.8939	0.3807	0.2208	0.4699
	3rd	0.9049	0.3733	0.1978	0.4448
	2nd	0.8923	0.3521	0.2241	0.4734
	1st	0.7867	0.5007	0.4439	0.6662
SVM	4th	0.8941	0.3581	0.2204	0.4695
	3rd	**0.9264**	0.2923	0.1530	0.3912
	2nd	0.8887	0.3499	0.2315	0.4812
	1st	0.7833	0.5029	0.4510	0.6715

Explanations: R^2 = coefficient of determination, *MAE* = mean absolute error, *MSE* = mean squared error, *RMSE* = root mean squared error, values in bold = the highest coefficient of determination between different machine learning models for each tree level.
Source: own study.

Model validation with specific data from the 2022 season

To verify the robustness of the generated models, 40 complete paths (length and diameter) per level with their respective production values were collected for 10 trees during the 2022/2023 agricultural season and fed into the selected ML models.

In Table 4, the results indicate that the models for levels 2, 3, and 4 lost some of their accuracy, with a maximum of 15.3% for the 2nd level. However, the 1st level gained in accuracy with an increase of 14.8%.

Table 4. Random forest (RF) and support vector machines (SVM) machine learning model accuracy per tree level with only 40 complete architectural data path from 10 different trees and the gap with the complete data set from 2019/2021/2022

Model	Tree level	R^2		Accuracy variation	% of variation
		2023	2019/ 2021/ 2022		
RF	4th	0.8768	**0.9116**	–0.0348	–3.8%
RF	2nd	0.7684	**0.9077**	–0.1393	–15.3%
RF	1st	**0.9054**	0.7890	0.1164	14.8%
SVM	3rd	0.7964	**0.9264**	–0.13	–14.0%
Average	–	0.8367	0.8836	–	–

Explanations: R^2 = coefficient of determination, values in bold = the highest coefficient of determination between the two data set.
Source: own study.

By compiling the results obtained by the different models according to levels, we can obtain the following equation:

$$ETP = \mathrm{RF}(\mathrm{L1} + \mathrm{L2} + \mathrm{L4}) + \mathrm{SVM}(\mathrm{L3}) \quad (6)$$

where: *ETP* = estimated tree production, RF = random forest, SVM = support vector machines, L1–L4 = 1st–4th levels.

This approximation will always result in an underestimation of tree production since the equation approximates the production of the 4 structural levels of the tree, which contribute to 91% of the tree's production, and neglects the production of the remaining 9% of other structural levels.

The SVM's better performance for the 3rd level, while RF excels for others, suggests probable difference growth dynamics or allometric relationships at this specific architectural level. It is possible that the 3rd level represents a transition between main structural branches and fruit-bearing shoots, where non-linear relationships captured by the SVM kernel are more decisive.

DISCUSSION

The prediction method is based on two basic parameters known for their high correlations with production, namely the diameter of structures (trunk cross-sectional area (TCA)) (Jiménez and Díaz, 2003) and the length of branches (Pérez González, 1993; Marini and Sowers, 1994). Additionally, dividing the tree structures into several levels allowed for greater precision in choosing machine learning models based on the evolution of data within each level, compared to other studies that rely on a single model for the entire dataset.

In Table S1, the results remind us that production for the same peach orchard can vary significantly between years in terms of the number and weight of fruits. Estimating plot production from the individual production of selected trees within the plot proved to be accurate to 88.3% compared to the average production of three agricultural seasons. The overestimation by 10% for the years 2020/2021 is probably due to the quantity of rain received (396 mm) that exceeded the years 2018/2019 (366 mm), and the same for 2021/2022, which received an amount equal to 347 mm. Additionally, the years of 2020/2021 cumulated more chilling hours between the two months December and January (300 h) in comparison with the two years 2018/2019 (282 h) and 2021/2022 (127 h) (unpublished data). To exploit this production from individual trees, estimating their production using machine-learning models based on non-destructive measurements before flowering is very promising. Indeed, with two different machine learning (ML) models (random forest (RF) and support vector machines (SVM)) fed with data (diameter and length) from only four structural levels of the trees, we can achieve an average total accuracy of 88.37% of the production of the four levels of the tree. Moreover, with only 40 paths chosen from the 10 sampled trees, the model combining RF and SVM generates an average accuracy of level per tree production of the four levels at 83.67% (Tab. 5) with an average loss of accuracy of 4.7% for a data reduction of 98.7%. This provides insight into the stability and robustness of the model generated by ML by combining RF and SVM for different levels of the tree.

The proposed prediction model in this study offers several advantages. Firstly, the number of trees is reduced (10 trees)

compared to previous studies (28 trees in the study by Underwood *et al.* (2016)). Secondly, the measurements are taken before flowering. Most prediction techniques are based on estimating flower or leaf development canopy (Sarron *et al.*, 2018; Matese and Di Gennaro, 2021) or vegetation index (Rahman, Robson and Bristow, 2018; Maimaitiyiming *et al.*, 2019; Sun *et al.*, 2020).

The proposed prediction method is more suitable for stone fruit production, especially peaches, because canopy formation and vegetation index occur later in the botanical development of the trees, after the formation of the 3rd level. Additionally, for stone fruit trees like peaches, measuring the canopy is relatively late as the foliage appears after flower production.

The measurements are accurate compared to other assessment methods, such as canopy measurement, which is difficult to measure and requires the use of image analysers that are expensive and time-consuming (Robinson and Lakso, 1991; Giuliani *et al.*, 2000).

The model remains stable concerning production changes related to variations in the agricultural season. Moreover, the technical and labour requirements for taking the necessary measurements are low, with minimal user-related bias.

CONCLUSIONS

Yield prediction has become a necessity for assessing the economic potential of production as well as managing logistical resources such as labour for harvesting and transportation. The proposed yield prediction method relies on structural measurements of the tree architecture early in the peach production season, well before flowering and canopy formation. The stability of the prediction model, based on the processing of main architectural levels of the tree by two types of machine learning (ML) models (random forest (RF) and support vector machines (SVM)), in the face of production variation linked to agricultural practices, is a major asset for the adoption of a robust and accurate prediction model.

Although the proposed model is closely linked to the cultivar, locality, and technical management adopted in the farm hosting the study trial, the promising results of the model make it possible to consider testing it for other cultivars in the same or different locations. After the study of the production distribution per level for the target stone cultivar, it is also interesting to consider significantly reducing the number of measurements per tree and focus interest on the 2nd to 4th productive levels according to the percentage of contribution to individual tree production and replacing it with an average correction factor.

SUPPLEMENTARY MATERIAL

Supplementary material to this article can be found online at: https://www.jwld.pl/files/Supplementary_material_68_Zegoumou.pdf.

CONFLICT OF INTERESTS

All authors declare that they have no conflict of interests.

REFERENCES

Anderson, N.T. *et al.* (2019) "Estimation of fruit load in mango orchards: tree sampling considerations and use of machine vision and satellite imagery," *Precision Agriculture*, 20, pp. 823–839. Available at: https://doi.org/10.1007/s11119-018-9614-1.

Anderson, N.T., Walsh, K.B. and Wulfsohn, D. (2021) "Technologies for forecasting tree fruit load and harvest timing – from ground, sky and time," *Agronomy*, 11(7), 1409. Available at: https://doi.org/10.3390/agronomy11071409.

Apolo-Apolo, O.E. *et al.* (2020) "A cloud-based environment for generating yield estimation maps from apple orchards using UAV imagery and a deep learning technique," *Frontiers in Plant Science*, 11, 1086. Available at: https://doi.org/10.3389/fpls.2020.01086.

Ballesteros, R. *et al.* (2020) "Vineyard yield estimation by combining remote sensing, computer vision and artificial neural network techniques," *Precision Agriculture*, 21, pp. 1242–1262. Available at: https://doi.org/10.1007/s11119-020-09717-3.

Beek van, J. *et al.* (2015) "Temporal dependency of yield and quality estimation through spectral vegetation indices in pear orchards," *Remote Sensing*, 7(8), pp. 9886–9903. Available at: https://doi.org/10.3390/rs70809886.

Breiman, L. (2001) "Random forests," *Machine learning*, 45(1), pp. 5–32. Available at: https://doi.org/10.1023/A:1010933404324.

Cannell, M.G.R. and Morgan, J. (1987) "Young's modulus of sections of living branches and tree trunks," *Tree Physiology*, 3(4), pp. 355–364. Available at: https://doi.org/10.1093/treephys/3.4.355.

Caruso, T. *et al.* (1999) "Effect of planting system on productivity, dry-matter partitioning and carbohydrate content in above-ground components of 'Flordaprince' peach trees," *Journal-American Society for Horticultural Science*, 124, pp. 39–45. Available at: https://doi.org/10.21273/JASHS.124.1.39.

Chang, L.S., Iezzoni, A.F. and Flore, J.A. (1987) "Yield components in 'Montmorency' and 'Meteor' sour cherry," *Journal of the American Society for Horticultural Science*, 112(2), pp. 247–251. Available at: https://doi.org/10.21273/JASHS.112.2.247.

Chapelle, O. *et al.* (2002) "Choosing multiple parameters for support vector machines," *Machine Learning*, 46, pp. 131–159. Available at: https://doi.org/10.1023/A:1012450327387.

Chen, T. and Guestrin, C. (2015) "XGBoost: A scalable tree boosting system," in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. San Francisco, CA, USA, 13–17 Aug 2016. New York: Association for Computing Machinery. Available at: https://doi.org/10.1145/2939672.2939785.

Cherkassky, V. and Ma, Y. (2004) "Practical selection of SVM parameters and noise estimation for SVM regression," *Neural Networks*, 17(1), pp. 113–126. Available at: https://doi.org/10.1016/S0893-6080(03)00169-2.

Darwin, B. *et al.* (2021) "Recognition of bloom/yield in crop images using deep learning models for smart agriculture: A review," *Agronomy*, 11(4), 646. Available at: https://doi.org/10.3390/agronomy11040646.

Dennis, F.J., Masabni, J.G. and Ketchie, D.O. (1996) "Evaluating twenty-eight strains of 'Delicious' apple in Michigan," *Journal of the American Society for Horticultural Science*, 121(6). Available at: https://doi.org/10.21273/JASHS.121.6.988.

Gao, Y. *et al.* (2021) "Prediction model of random forest for the risk of hyperuricemia in a Chinese basic health checkup test," *Bioscience Reports*, 41(4). Available at: https://doi.org/10.1042/BSR20203859.

Giuliani, R. *et al.* (2000) "Ground monitoring the light–shadow windows of a tree canopy to yield canopy light interception and morphological traits," *Plant, Cell & Environment*, 23(8), pp. 783–796. Available at: https://doi.org/10.1046/j.1365-3040.2000.00600.x.

Gongal, A. *et al.* (2015) "Sensors and systems for fruit detection and localization: A review," *Computers and Electronics in Agriculture*, 116, pp. 8–19. Available at: https://doi.org/10.1016/j.compag.2015.05.021.

Hacking, C. *et al.* (2019) "Investigating 2-D and 3-D proximal remote sensing techniques for vineyard yield estimation," *Sensors*, 19(17), 3652. Available at: https://doi.org/10.3390/s19173652.

He, L. *et al.* (2022) "Fruit yield prediction and estimation in orchards: A state-of-the-art comprehensive review for both direct and indirect methods," *Computers and Electronics in Agriculture*, 195, 106812. Available at: https://doi.org/10.1016/j.compag.2022.106812.

Hsu, C.W., Chang, C.C. and Lin, C.J. (2003) *A practical guide to support vector classification*. National Taiwan University, Department of Computer Science. Available at: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf (Accessed: April 11, 2025).

Jiménez, C.M. and Díaz, J.B.R. (2003) "A statistical model to estimate potential yields in peach before bloom," *Journal of the American Society for Horticultural Science*, 128(3), pp. 297–301. Available at: https://doi.org/10.21273/JASHS.128.3.297.

Kamilaris, A. and Prenafeta-Boldú, F.X. (2018) "Deep learning in agriculture: A survey," *Computers and Electronics in Agriculture*, 147, pp. 70–90. Available at: https://doi.org/10.1016/j.compag.2018.02.016.

Kappel, F. (1990) "Yield component analysis of 'Harrow Delight', 'Kieffer', and 'Harvest Queen' pear," *Journal of the American Society for Horticultural Science*, 115(1), pp. 25–29.

Laurent, C. *et al.* (2021) "A review of the issues, methods and perspectives for yield estimation, prediction and forecasting in viticulture," *European Journal of Agronomy*, 130, 126339. Available at: https://doi.org/10.1016/j.eja.2021.126339.

Liakos, K.G. *et al.* (2018) "Machine learning in agriculture: A review," *Sensors*, 18(8), 2674. Available at: https://doi.org/10.3390/s18082674.

Liu, T.H. *et al.* (2018) "Detection of citrus fruit and tree trunks in natural environments using a multi-elliptical boundary model," *Computers in Industry*, 99, pp. 9–16. Available at: https://doi.org/10.1016/j.compind.2018.03.007.

Maimaitiyiming, M. *et al.* (2019) "Dual activation function-based extreme learning machine (ELM) for estimating grapevine berry yield and quality," *Remote Sensing*, 11(7), 740. Available at: https://doi.org/10.3390/rs11070740.

Marini, R.P. and Sowers, D.L. (1994) "Peach fruit weight is influenced by crop density and fruiting shoot length but not position on the shoot," *Journal of the American Society for Horticultural Science*, 119(2), pp. 180–184. Available at: https://doi.org/10.3390/rs11070740.

Marini, R.P. and Sowers, D.S. (2000) "Peach tree growth, yield, and profitability as influenced by tree form and tree density," *HortScience*, 35(5), pp. 837–842. Available at: https://doi.org/10.21273/HORTSCI.35.5.837.

Matese, A. and Di Gennaro, S.F. (2021) "Beyond the traditional NDVI index as a key factor to mainstream the use of UAV in precision viticulture," *Scientific Reports*, 11(1), 2721. Available at: https://doi.org/10.1038/s41598-021-81652-3.

Murray, C.D. (1927) "A relationship between circumference and weight in trees and its bearing on branching angles," *The Journal of General Physiology*, 10(5), 725. Available at: https://doi.org/10.1085/jgp.10.5.725.

Nuske, S. *et al.* (2014) "Automated visual yield estimation in vineyards," *Journal of Field Robotics*, 31(5), pp. 837–860. Available at: https://doi.org/10.1002/rob.21541.

Payne, A.B. *et al.* (2013) "Estimation of mango crop yield using image analysis – Segmentation method," *Computers and Electronics in Agriculture*, 91, pp. 57–64. Available at: https://doi.org/10.1016/j.compag.2012.11.009.

Peréz-González, S. (1993) "Bud distribution and yield potential in peach," *Fruit Varieties Journal*, 47(1). Available at: https://doi.org/10.71318/apom.1993.47.1.18.

Planchon, V., Claustriaux, J.J. and Crabbé, J. (2003) "Description et modélisation de la croissance et du développement du pommier (*Malus* x *domestica* Borkh.): II. Caractéristiques et distribution spatiale et temporelle des sites de floraison [Description and modeling of apple tree (*Malus* x *domestica* Borkh.) growth and development: II. Characteristics and spatial and temporal distribution of flowering sites]," *Biotechnologie, Agronomie, Société et Environnement*, 7(2), pp. 99–110.

Rahman, M.M., Robson, A. and Bristow, M. (2018) "Exploring the potential of high resolution worldview-3 Imagery for estimating yield of mango," *Remote Sensing*, 10(12), 1866. Available at: https://doi.org/10.3390/rs10121866.

Rakićević, J., Savić, G. and Bulajić, M. (2016) "Selecting an appropriate method for missing data imputation: A case of countries ranking," in O. Jaško and S. Marinković (eds.) *Proceedings of XV International Symposium Symorg 2016: Reshaping the Future through Sustainable Business Development and Entrepreneurship*, pp. 91–99. Zlatibor, Serbia, 10–13 Jun 2016. Belgrade: University of Belgrade, Faculty of Organizational Sciences.

Robinson, T.L. and Lakso, A.N. (1991) "Bases of yield and production efficiency in apple orchard systems," *Journal of the American Society for Horticultural Science*, 116(2), pp. 188–194. Available at: https://doi.org/10.21273/JASHS.116.2.188.

Sarron, J. *et al.* (2018) "Mango yield mapping at the orchard scale based on tree structure and land cover assessed by UAV," *Remote Sensing*, 10(12), 1900. Available at: https://doi.org/10.3390/rs10121900.

Serrano, L., González-Flor, C. and Gorchs, G. (2012) "Assessment of grape yield and composition using the reflectance based Water Index in Mediterranean rainfed vineyards," *Remote Sensing of Environment*, 118, pp. 249–258. Available at: https://doi.org/10.1016/j.rse.2011.11.021.

Sirsat, M.S. *et al.* (2019) "Machine learning predictive model of grapevine yield based on agroclimatic patterns," *Engineering in Agriculture, Environment and Food*, 12(4), pp. 443–450. Available at: https://doi.org/10.1016/j.eaef.2019.07.003.

Strong, D. and Azarenko, A.N. (2000) "Relationship between trunk cross-sectional area, harvest index, total tree dry weight and yield components of 'Starkspur Supreme Delicious' apple trees," *Journal of American Pomological Society*, 54(1), pp. 22–27. Available at: https://doi.org/10.71318/apom.2000.54.1.22.

Sun, G. *et al.* (2020) "A canopy information measurement method for modern standardized apple orchards based on UAV multimodal information," *Sensors*, 20(10), 2985. Available at: https://doi.org/10.3390/s20102985.

Thornton, C. (2014) *Auto-WEKA: combined selection and hyperparameter optimization of supervised machine learning algorithms*. PhD Thesis. University of British Columbia.

Tyree, M.T. *et al.* (1991) "Water relations and hydraulic architecture of a tropical tree (*Schefflera morototoni*) data, models, and a comparison with two temperate species (*Acer saccharum* and

Thuja occidentalis)," *Plant Physiology*, 96(4), pp. 1105–1113. Available at: https://doi.org/10.1104/pp.96.4.1105.

Underwood, J.P. *et al.* (2016) "Mapping almond orchard canopy volume, flowers, fruit and yield using lidar and vision sensors," *Computers and Electronics in Agriculture*, 130, pp. 83–96. Available at: https://doi.org/10.1016/j.compag.2016.09.014.

Upadhyaya, S.K., Cooke, J.R. and Rand, R.H. (1987) "Variation in Young's modulus along apple limbs," *Transactions of the ASABE*, 30(5), pp. 1501–1505. Available at: https://doi.org/10.13031/2013.30593.

Wang, Q. *et al.* (2013) "Automated crop yield estimation for apple orchards," in J.P. Desai *et al.* (eds.) *Experimental Robotics. Proceedings of the 13th International Symposium on Experimental Robotics*, pp. 745–758. Québec City, Canada, 18–21 Jun 2012. Cham: Springer.

Wu, Z. *et al.* (2021) "Segmentation of abnormal leaves of hydroponic lettuce based on DeepLabV3+ for robotic sorting," *Computers and Electronics in Agriculture*, 190, 106443. Available at: https://doi.org/10.1016/j.compag.2021.106443.

Xu, W. *et al.* (2019) "Shadow detection and removal in apple image segmentation under natural light conditions using an ultrametric contour map," *Biosystems Engineering*, 184, pp. 142–154. Available at: https://doi.org/10.1016/j.biosystemseng.2019.06.016.

Yu, T. and Zhu, H. (2020) "Hyper-parameter optimization: A review of algorithms and applications," *arXiv*, 2003.05689v1. Available at: https://doi.org/10.48550/arXiv.2003.05689.