IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 21, 2021, accepted November 2, 2021, date of publication November 10, 2021,
date of current version November 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3127453

Using Machine Learning Algorithms to Forecast the Sap Flow of Cherry Tomatoes in a Greenhouse

AMORA AMIR[1], **MARYA BUTT**[2], **AND OLAF VAN KOOTEN**[1,3]
[1]Research and Innovation Centre Techniek, Ontwerpen en Informatica, Inholland University of Applied Sciences, 2628 AL Delft, The Netherlands
[2]Research and Innovation Centre Techniek, Ontwerpen en Informatica, Inholland University of Applied Sciences, 2015 CE Haarlem, The Netherlands
[3]Hortical Production Chains Group, Department of Plant Sciences, Wageningen University & Research, 6708 PB Wageningen, The Netherlands

Corresponding author: Amora Amir (amora.amir@inholland.nl)

This work was supported in part by the European Funds for Regionale Ontwikkeling (EFRO) EVERGREEN Greenport Noord-Holland Noord under Project 3355 EFRO Evergreen, in part by Kansen voor West under Project KVW-00052, and in part by the EFRO Fieldlab Vertical Farming under Project A1411.

ABSTRACT The sap flow of plants directly indicates their water requirements and provides farmers with a good understanding of a plant's water consumption. Water management can be improved based on this information. This study focuses on forecasting tomato sap flow in relation to various climate and irrigation variables. The proposed study utilizes different machine learning (ML) techniques, including linear regression (LR), least absolute shrinkage and selection operator (LASSO), elastic net regression (ENR), support vector regression (SVR), random forest (RF), gradient boosting (GB) and decision tree (DT). The forecasting performance of different ML techniques is evaluated. The results show that RF offers the best performance in predicting sap flow. SVR performs poorly in this study. Given water/m^2, room temperature, given water EC, humidity and plant temperature are the best predictors of sap flow. The data are obtained from the Ideal Lab greenhouse, in the Netherlands, in the framework of the European Funds for Regionale Ontwikkeling (EFRO) EVERGREEN Greenport Noord Holland Noord project (2018-2020).

INDEX TERMS Sap flow, tomato, future forecasting, machine learning, feature importance, hyperparameters, adjusted R^2.

I. INTRODUCTION

In alignment with artificial intelligence (AI) and big data technology, machine learning (ML) introduces new opportunities to unravel, measure, mine and understand the hidden patterns of data processes in dynamic and static environments [1]. ML is defined as the scientific field of statistical techniques that confers machines with the ability to learn from a series of input and output examples. ML is applied in many scientific fields, for example, bioinformatics, medicine, finance and economic sciences, robotics and vision engineering, sentiment analysis of social media, agriculture, climatology and food security. One important use of ML is predicting possible factors that influence crop management, specifically yield forecasting, crop growth forecasting, health prediction, decision making and crop mapping [2]. ML has potential to address existing and future challenges in agriculture by means of massive volumes of data containing a wide variety of indicators that can be captured, analyzed, processed and used for decision making. It is essential to gather data from various sources when making predictive decisions e.g., preventing crop loss and increasing yield while minimizing the use of resources. Many studies have investigated applications of ML in agriculture. Kaul *et al.* [3] applied artificial neural networks for highly accurate corn and soybean yield prediction. Logan *et al.* [4] applied generalized linear model (GLM), Bayesian additive regression tree (BART), and classification and regression tree (CART) methods together to utilize the high predictive power to achieve efficacy in the decision-making process with respect to Royal Gala apples. In another study, Delgado *et al.* [5] adapted a fuzzy logic information network and a decision-support system to address imprecision and inaccuracy for effective decision making in olive cultivation. Furthermore, Utkarsha *et al.* [6] focused on clustering ML for crop growth prediction, while Jing-Xian *et al.* [7] performed regression supervised learning to forecast sugarcane yield.

This study aims to predict sap flow in cherry tomatoes. Currently, automated irrigation systems are commonly used in greenhouses. In contrast to the manual effort required to

The associate editor coordinating the review of this manuscript and approving it for publication was Jad Nasreddine.

water plants, farmers need only one person to control the computer. The amount of the water to be given is determined by solar radiation. However, with the use of new technologies, the current irrigation strategy does not meet the required accuracy for greenhouse applications. According to Sutton and Barto [8], the water given to plants based on solar radiation might be wasted by the low water storage capacity substrate (Rockwool), which contradicts the energy-saving strategy and might cause declines in production and quality. Sap flow sensors make the water requirements of plants more obvious, accurate and direct. In contrast to the mass-balance technique to check plant water uptake, the sap flow sensor provides real-time data [9] and shows precise changes in water use in response to different environmental conditions. For commercial purposes, such sensors can help farmers to improve or adjust their water management strategy, as according to Gimenez *et al.* [10], sap flow can be used as an indicator of a plant's water status. Sap flow sensors are commonly used in forestry and vine production for research and commercial purposes and have even been adopted in the orchid industry. However, there is a gap in the use of sensors for edible herbaceous plants. Some features of cherry tomato plants make them good research objects. Tomato plants are perennials that are grown throughout the entire year in greenhouses, which provides a long time for research. Moreover, tomato plants have strong and thick stems, which simplify sensor installation [8]. Another motivation of this study is to avoid water drainage or loss, which is directly proportional to the amount of given water. We can prevent excessive drainage or loss by providing only the required amount of water to the plant.

To contribute to the application of sap flow information, this study attempts to predict tomato sap flow based on multiple variables using ML algorithms. Three categories of variables are considered in this study: climate data, irrigation data, and sap flow data. These data were collected from the Ideal Lab greenhouse [11] in Naaldwijk, the Netherlands. To achieve optimum monitoring, various sensors with actuators are installed on and around the tomato plants in our experimental greenhouse lab. The sensors on the plant provide information about sap flow. In addition, sensors are installed in the greenhouse to obtain (big) data about the conditions within the greenhouse: for example, climate sensors for measuring temperature, humidity, sunlight and irrigation water supply. Moreover, sensors in the substrate mat continuously measure mat weight and the pH and electrical conductance (EC) of water. The forecasting problem in this study is considered as a regression problem, and ML regression models, such as linear regression (LR), least absolute shrinkage and selection operator (LASSO), elastic net regression (ENR), support vector regression (SVR), random forest (RF), gradient boosting (GB) and decision tree (DT), are used to predict sap flow. The models are trained using the climate, irrigation and sap flow datasets provided by Ideal Lab greenhouse in Naaldwijk, the Netherlands. The dataset is divided into a training set, including 80% of the records, and a test set, with the remaining 20%, and the R-squared score (R^2), adjusted R-squared score (adjusted R^2), mean square error (MSE), root mean square error (RMSE) and mean absolute error (MAE) are used as performance metrics.

The key findings of this study are listed as follows:

- RF offers the best sap flow prediction capability with the highest R^2 value of 81% (approx.) and an MSE of 0.25.
- Given water/m^2, room temperature, given water EC, humidity and plant temperature are the best predictors.

This paper is divided into five sections. Section I describes the background and goals of this study; in Section II, the materials and methods are described. The methodology is presented in Section III. Section IV presents the results and discussion of the results, including figures and tables. Finally, Section V concludes this paper.

II. MATERIALS AND METHODS

A. DATASET

A cherry tomato variety is used in this project. Tomato plants were grown in the Ideal Lab greenhouse (length: 12.50 m, width: 6.50 m; height: 6 m) located at the World Horti Center, Naaldwijk, the Netherlands [11]. The seedlings (16 cm) were provided by Axia Vegetable Seeds company [12] and were grafted onto rootstocks (Maxifor, provided by Rijk Zwaan) on 8th November 2018 [13]. Rockwool slabs were provided by Grodan' GT Master [14]. Artificial light was applied between 7:00 am and 6:00 pm each day, the average day temperature inside the greenhouse was 23 °C, the average night temperature was 17 °C, the CO_2 application was 533 ppm, and the irrigation system applied water at a rate of 0.5 L/m^2 on average. These amounts can be adjusted to meet the Dutch cultivation strategy depending on the weather outside the greenhouse. A total of 364 records from 12 samples of cherry tomato plants were considered.

1) CLIMATE AND IRRIGATION DATASETS

The climate dataset includes the room temperature, air humidity, carbon dioxide (CO_2), outside radiation, air density, outside temperature, outside air humidity, outside air density, wind speed and plant temperature. The irrigation dataset consists of given water EC, given water pH, given water/m^2, drained water EC, drained water amount, and absorbed water amount. All these data were monitored and recorded automatically via the Priva [15] climate system on a daily basis.

2) SAP FLOW DATASET

The sap flow dataset was recorded using Dynagage SF sensors provided by 2GROW [16]. The sap flow rate was recorded every 2.5 seconds. One sensor was installed on tomato plant, and data were monitored and recorded automatically for the entire project period. The data are presented visually using the Phythosense software package of 2GROW [17].

TABLE 1. **Sap flow data sample.**

Date	29-01-2019	30-01-2019	…	22-09-2019
Sap Flow	82.7723	271.6120	…	955.4850

TABLE 2. **Climate data sample.**

Date	29-01-2019	30-01-2019	…	22-09-2019
Room Temperature	14.10	14.20	…	17.70
Humidity	77.99	75.98	…	87.00
CO2	465.07	476.02	…	433.18
Radiation	0.00	0.00	…	0.91
Plant Temperature	14.70	14.20	…	17.30

TABLE 3. **Irrigation data sample.**

Date	29-01-2019	30-01-2019	…	22-09-2019
Given water EC	3.8	3.5	…	3.0
Given water pH	5.8	5.8	…	4.1
Drain water EC	2.9	2.9	…	3.2
Drain water pH	6.8	6.8	…	6.7
Drain water	0.000000	0.412167	…	1.116000
Given water/m^2	0.170	0.341	…	3.600

Data samples from each dataset are shown in Tables 1, 2, and 3. The variables selected for inclusion in the study are as follows:

- Room temperature
- Air humidity
- Carbon dioxide (CO_2)
- Plant temperature
- Given water EC
- Given water pH
- Given water/m^2
- Drained water amount
- Sap flow

B. SUPERVISED MACHINE LEARNING MODELS

The purpose of this study is to construct models to predict sap flow based on input predictors: seven supervised ML methods are considered in this study.

1) LINEAR REGRESSION

LR is a supervised ML algorithm [18] based on independent and dependent variables. According to the number of variables, LR can be categorized as simple LR or multiple LR, as shown in the following equations. The goal of LR is to identify the best combination of weight (w) and bias (b) that leads to the lowest cost (J).

$$f(x) = wx + b \tag{1}$$

Or

$$J = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2 \tag{2}$$

J is the cost function, $f(x_i)$ is the predicted value, and y_i is the actual value.

2) LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR

LASSO is an LR regression technique [19] that performs well in cases of high multicollinearity and sparse models [20]. In contrast to normal multiple LR, LASSO performs automatic selection among the predictors. The goal of LASSO is to minimize a coefficient, i.e., to minimize (sum of squared residuals + λ*|slope|). The equation is shown below.

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \tag{3}$$

λ is the shrinkage. When λ equals 0, the estimate is the same as that from LR.

3) ELASTIC NET REGRESSION

ENR is a regularized regression algorithm that combines LASSO and ridge regression [21]. The estimates for ENR are the minima (sum of the squared residuals + λ*|slope | + λ*slope^2). ENR addresses the disadvantage of LASSO by removing the limitation on the number of selected variables. The equation is shown below.

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j| + \lambda\sum_{j=1}^{p}|\beta_j|^2 \tag{4}$$

4) SUPPORT VECTOR REGRESSION

SVR is widely used in classification problems in ML [22]. A line, also called a hyperplane, is constructed to separate the training data in N dimensions. Multiple hyperplanes can be used to classify the data. The hyperplane with the best performance is the one that achieves the largest separation. Subsequently, the regression is performed based on the hyperplanes. The equation of SVR is as follows:

$$f(x) = x'\beta + b \tag{5}$$

5) RANDOM FOREST

RF is a supervised ML algorithm that is used for classification and regression tasks [23]. RF is an ensemble of multiple regressions, where multiple DT regressions are performed in a parallel manner. The result aggregates many DTs into a single ensemble regression via voting or by taking the mean value of different DTs. The goal of RF is to perform forecasting based on the regression trees.

6) GRADIENT BOOSTING

GB converts weak learners into strong learners [24], typically starting with a DT model. GB builds upon a previous model by adding another DT. If the new DT does not correlate with the previous forecasting system, it will be selected out. The final prediction is the weighted sum of the previous predictions.

7) DECISION TREE

DT builds regression models in a tree structure and uses a set of binary rules to calculate a target value [25]. The model can be trained to fit any historical data and to learn any relationships between data and variables.

C. EVALUATION PARAMETERS

The performance of each model was evaluated in terms of the R-squared (R^2) score, adjusted R-squared (adjusted R^2) score, mean square error (MSE), root mean square error (RMSE) and mean absolute error (MAE).

1) R-SQUARED SCORE

The R-squared score represents the performance of the regression model [26]. When R^2 is less than 0, the model has no value. When it is equal to 0, the predicted value is equal to the mean value of the dependent variable. When it is equal to 1, the model performs the best. The value of R^2 score $\in (0, 1)$; the higher the R^2 score is, the better the performance of the model.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \tag{6}$$

2) ADJUSTED R-SQUARED SCORE

Adjusted R^2 can avoid the over-feeding data problem, which leads to a continuously increasing R^2 score [27]. When useless variables are added to the model, the adjusted R^2 decreases.

$$R^2_{adj} = 1 - [\frac{(1 - R^2)(n - 1)}{n - k - 1}] \tag{7}$$

3) MEAN SQUARE ERROR (MSE)

Mean square error is the average of the squared error [28]. The smaller the MSE value is, the better the performance of the model.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y})^2 \tag{8}$$

4) ROOT MEAN SQUARE ERROR (RMSE)

RMSE is the square root of the mean square error [29]. The equation is shown below.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{1} (y_i - \hat{y})^2} \tag{9}$$

5) MEAN ABSOLUTE ERROR (MAE)

MAE is the average of all the absolute errors between the predicted values and actual values [30]. The smaller the absolute error is, the lower is the MAE, and lower values indicate better model performance.

$$MAE = \frac{1}{N} \sum_{i=1}^{n} |y_i - \hat{y}| \tag{10}$$

D. HYPERPARAMETERS

The parameters of the model that must be set before running the model, in contrast to parameters that are learned during the training process, are referred to as hyperparameters [31]. To optimize the performance criteria, these parameters should be carefully tuned, as using excessively large or small values may result in poor model performance [32]. Hyperparameter tuning is, therefore, the process of finding good values of parameters for a specific dataset [31]. Sometimes, default values of the hyperparameters are defined by the packages being used; for example, in Python, if a value for a certain hyperparameter is not provided by the user for a particular ML algorithm, the default value is applied for training. The following hyperparameters are used.

1) ALPHA

The Scikit library in Python provides GridSearchCV to find the optimum value of alpha. In this study, alpha is a hyperparameter used for LASSO and ENR, and the chosen alpha values for LASSO and ENR are 0.05 and 0.06, respectively.

2) KERNEL

SVR uses linear and nonlinear kernels to map low-dimensional data to high-dimensional data. This study uses a linear kernel that supports listing feature importance, which is not possible when using other kernels, as data are transformed to another space via the kernel method.

3) N_ESTIMATORS

n_estimators represent the number of trees to be built for making average predictions. Higher values make the model stronger and more stable, but the code becomes slower. Therefore, the highest value that a processor could handle can be chosen for best results. e_estimators is the hyperparameter used in RF and GB, with a value of 1,500.

4) MAX_DEPTH

In DT, the dataset is partitioned into different subsets. Partitioning starts with a binary split and continues until no further splitting is possible. The max_depth refers to the depth of each tree in the forest, where deeper trees are expected to capture more information about the data. In the study, max_depth was set to 3, as higher values resulted in poor model accuracy.

III. METHODOLOGY

This study focuses on sap flow prediction in tomato plants using multiple predictors, such as climate variables (room temperature, humidity, CO_2) and irrigation variables (given water, drainage water, given water pH), which are daily data. Three main steps were used to construct the forecasting system, as shown in Fig. 1. The initial dataset was processed into Table 4. Standard scaling in Python was used to obtain a dataset close to a normal distribution, which benefits the performance of many ML algorithms, such as LR and SVR.

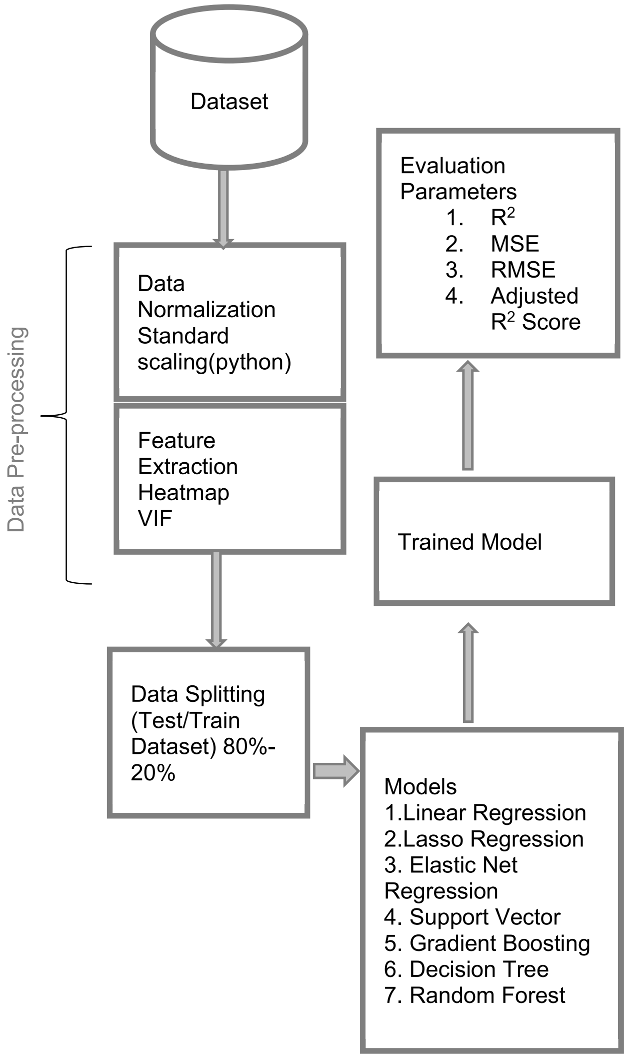

FIGURE 1. Workflow.

The output dataset after application of standard scaling is shown in Table 5.

The forecasting was performed using seven ML techniques. The potential important features were pre-selected based on the variance inflation factor (VIF), which accelerated the modeling process, and the correlations between predictors and sap flow are shown in Fig. 2.

Then, the dataset was split into training and test sets based on a parameter between 0 and 1, expressed as a percentage. The dataset was split to prevent look-ahead bias, overfitting and underfitting. Common split percentages include 80%, 67% and 50% [33]. The value of 80% indicates that 80% of the data are included in the training set, while 20% of the data are included in the test set. These values were adopted in this study. Commonly, there is no optimal split percentage. A chosen split percentage needs to meet the project's objectives with respect to the following consideration [33]:

- Computational cost in training the model.
- Computational cost in evaluating the model.
- Training set representativeness.
- Test set representativeness.

The ML models LR, LASSO, ENR, SVR, RF, GB and DT were implemented, and the performance was evaluated in terms of R^2, adjusted R^2, MSE, RMSE and MAE.

IV. RESULTS AND DISCUSSION

A. RESEARCH METHODOLOGY

Sap flow directly represents the water requirements of a plant and provides an opportunity to understand the plant's hydraulic function and plant growth in a given environment [34]. The movement of sap illustrates the connection between a plant and its surroundings [34], and sap flow sensors are applied broadly in the forestry sector for water management and research purposes [35]. However, such sensors are rarely used for herbaceous plants. In this research, the tomato, an herbaceous plant, was chosen as the research object to contribute to the sap flow database. The relationships between sap flow and climate factors were analyzed, and a predictive model of sap flow was constructed and tested. This model can be used to enhance greenhouse automation management, to improve water use efficiency and to reduce waste during production. In previous studies, sap flow was generally studied in reference to solar radiation, vapor pressure deficit, relative humidity, and air temperature [36]. By contrast, this study includes more measured variables as compared with vapor pressure deficit, which is calculated based on measured data, and more potential variables, such as plant temperature and CO_2, are included. Moreover, since most of the sensors are developed for woody plants [37], this research may contribute to sap flow sensor innovation.

B. SAP FLOW FORECASTING

Models for predicting tomato plant sap flow were developed and tested in this study. The performance evaluation results are presented in Table 6. According to the results, RF shows the highest correlation between the predicted values and actual values. RF is followed by LR, ENR, SVR and LASSO, which have similar R^2 values of approximately 0.790. GB shows the worst correlation, with an R^2 value of 0.663.

Figs. 3, 4, 5, 6, 7, 8, and 9 show the predicted values and actual values of sap flow for the different models. Figs. 10, 11, 12, 13, 14, 15, and 16 show the performances of different ML algorithms for sap flow prediction. According to the results, the sap flow data change frequently. Most of the forecasted values are accurate at the lowest peak; however, predictions of the highest peak are unstable. SVR shows good predictive ability for the highest peak of sap flow; however, the predicted values are not highly correlated with the actual values and the mean square error is relatively high, which reduces the reliability of SVR.

LR, LASSO and ENR show similar patterns with respect to the trend of sap flow data. LR shows the highest correlation between the predicted data and actual data (0.792)

TABLE 4. Unnormalized dataset view.

Sap flow	Humidity	CO$_2$	Outside temperature	Given Water EC	Given Water pH	Given Water per/m^2	Drained Water EC	Drained Water amount	Absorbed Water amount	Radiation Morning
3948.530	4.134644	534.991186	6.169085	3.600000	5.780000	1.776400	4.820000	0.047400	1.729000	62.036431
4287.320	4.296558	525.989956	7.742307	3.528571	5.942857	2.600000	4.457143	0.080571	2.519429	64.225337
4465.725	4.076610	539.001600	8.608055	3.257143	6.085714	2.674000	3.871429	0.105286	2.568714	58.735539
4644.130	4.191465	539.763186	7.221450	3.228571	5.800000	2.959571	3.757143	0.873571	2.086000	58.233885

TABLE 5. Normalized dataset view.

Sap flow	Humidity	CO$_2$	Outside temperature	Given Water EC	Given Water pH	Given Water per/m^2	Drained Water EC	Drained Water amount	Absorbed Water amount	Radiation Morning
-0.703	0.546	0.981	-1.443	2.873	0.721	-2.742	0.443	-3.011	-1.948	0.007
-0.665	0.742	0.757	-1.105	2.465	1.063	-1.822	-0.022	-2.921	-0.810	0.743
-0.645	0.475	1.081	-0.919	0.911	1.363	-1.740	-0.773	-2.853	-0.739	-1.105
-0.625	0.614	1.100	-1.217	0.748	0.763	-1.421	-0.919	-0.755	-1.434	-1.273

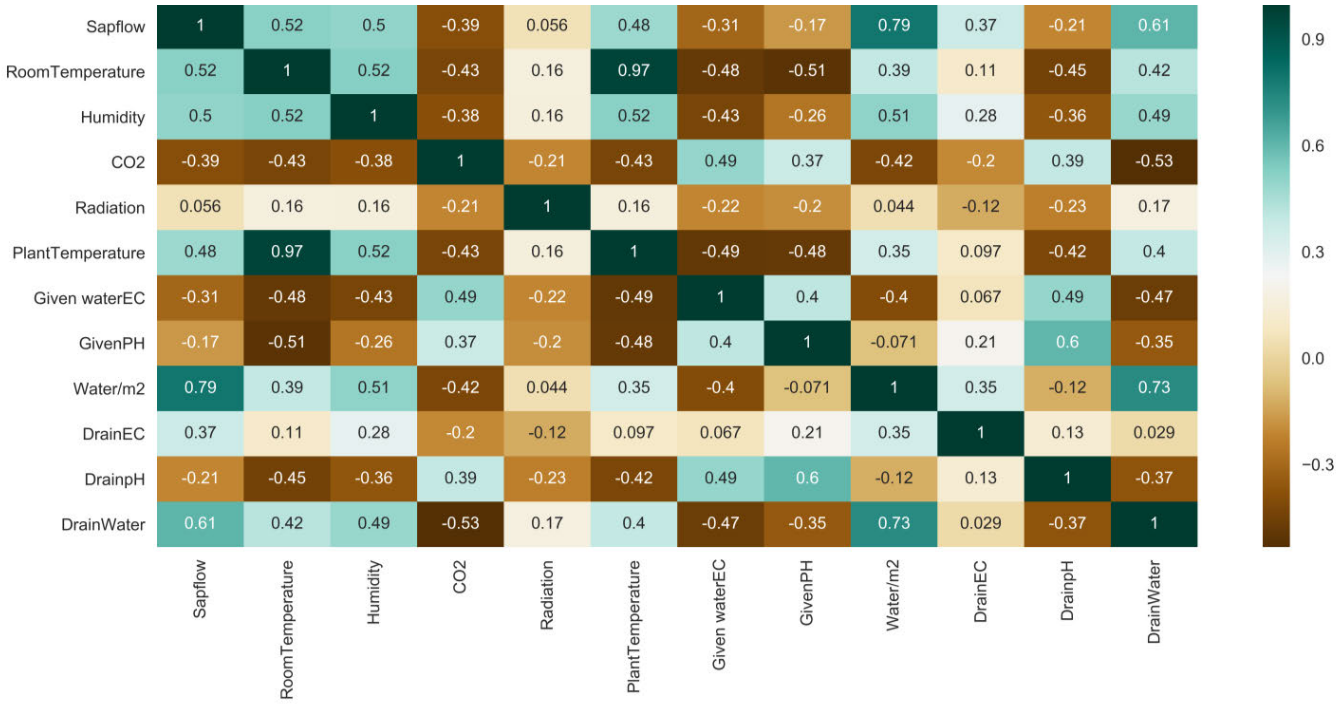

FIGURE 2. Correlation heatmap.

and is closely followed by ENR (0.788). In terms of the MSE and MAE, LR achieves the lowest values; therefore, LR performed the best in the linear regression technique group.

RF performed best in this study; it achieved the best sap flow prediction with the highest correlation and lowest error for peaks. In previous research, RF was described as a simple and diverse supervised learning algorithm, as it can be easily

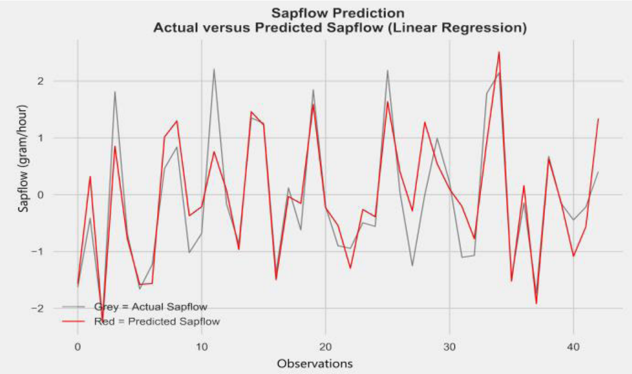

FIGURE 3. Prediction value and actual value of sap flow by linear regression.

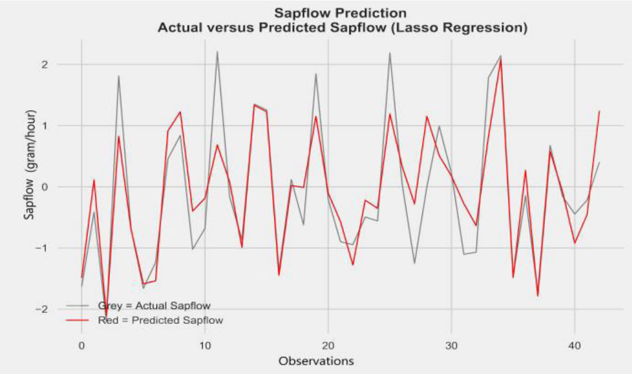

FIGURE 4. The prediction value and actual value of sap flow by Lasso regression.

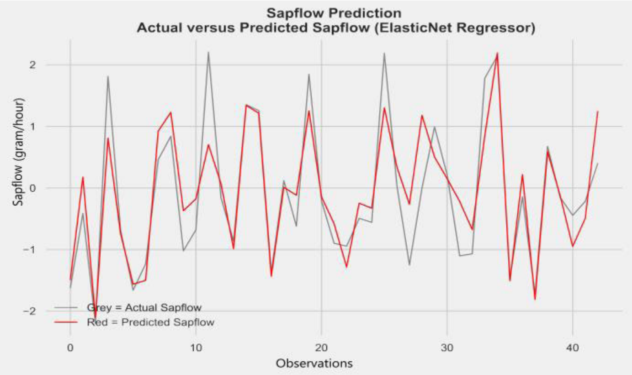

FIGURE 5. The prediction value and actual value of sap flow by elastic net regression.

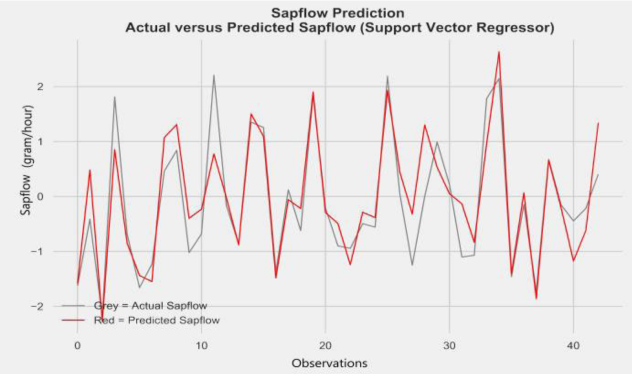

FIGURE 6. The prediction value and actual value of sap flow by support vector regressor.

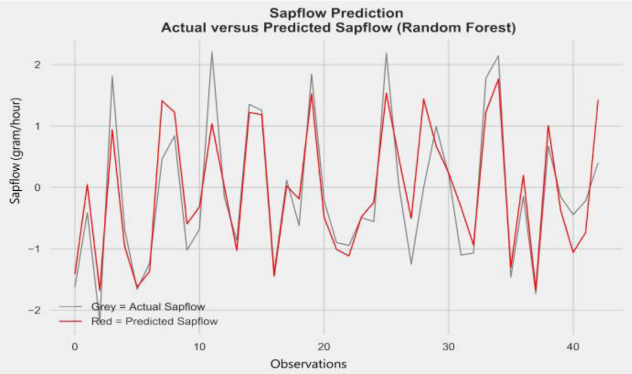

FIGURE 7. The prediction value and actual value of sap flow by random forest.

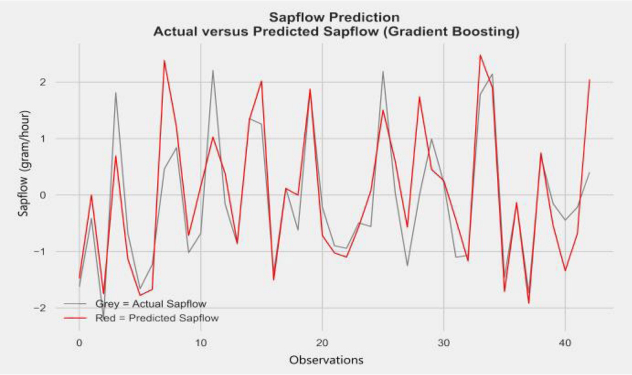

FIGURE 8. The prediction value and actual value of sap flow by gradient boosting.

used for both classification and regression. Niklas illustrated that RF is likely to achieve better performance than other approaches because of its high tree diversity [38]. When splitting a node, the best feature among a random subset of features, rather than the most important feature overall, is selected [38]. Additional advantages are reported by Julia: RF works well with high-dimensional data and unstable data [39]. RF achieves a lower variance than DT, as the variance of each DT is averaged in RF [39]. Moreover, RF does not suffer from excessive overfitting [40] and includes a rapid training process [39]. GB and DT show lower correlation and higher error than other algorithms. Therefore, GB and DT did not perform well with respect to sap flow prediction in this study.

C. FEATURE IMPORTANCE

To further improve the sap flow prediction performance, the feature importance in LR, ENR, SVR, RF, GB and DT was

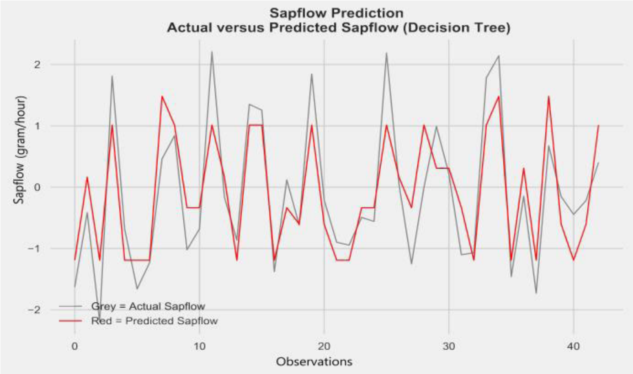

FIGURE 9. The prediction value and actual value of sap flow by decision tree.

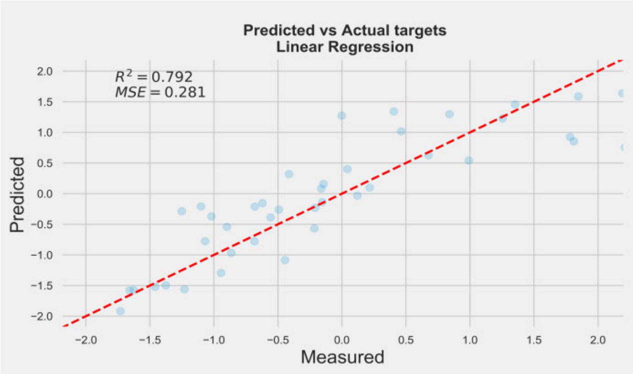

FIGURE 10. The model performance of linear regression.

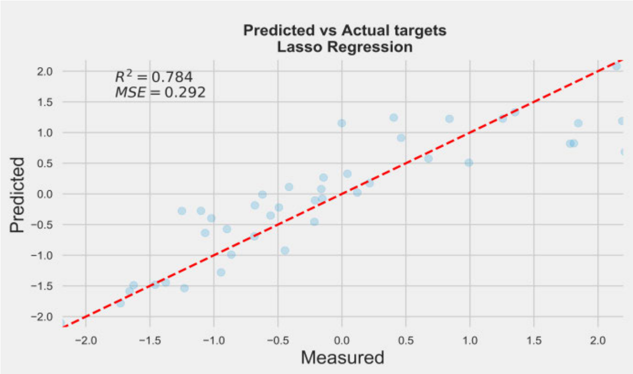

FIGURE 11. The model performance of Lasso regression.

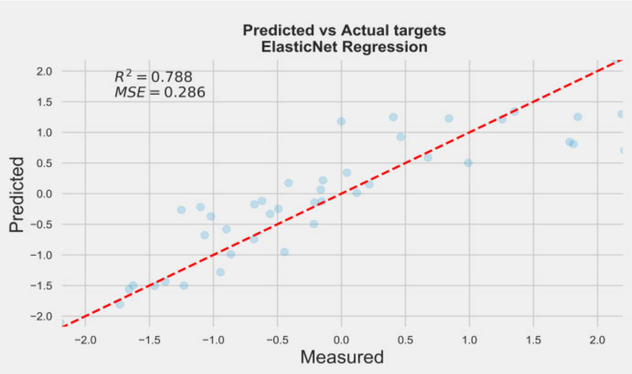

FIGURE 12. The model performance of elastic net regressor.

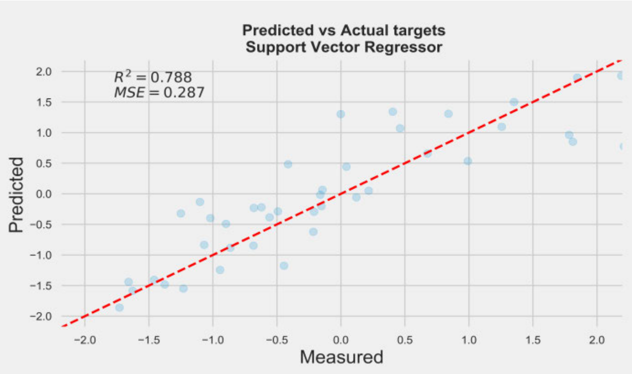

FIGURE 13. The model performance of support vector regressor.

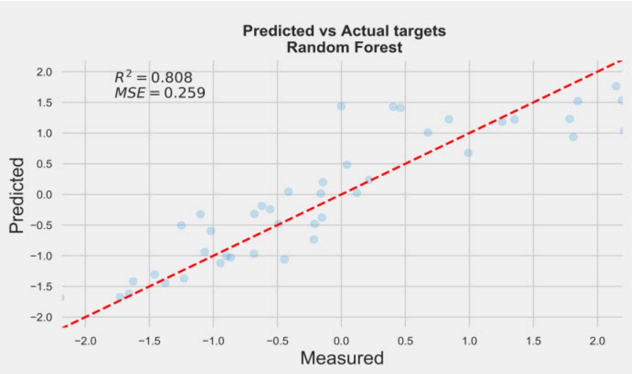

FIGURE 14. The model performance of random forest.

analyzed. LASSO was excluded from this process, as the features are automatically selected in LASSO.

The feature importance results are presented in Table 7 as the feature importance score (FI score). Four features contributed to the prediction by LR: given water/m^2, given water EC, room temperature and humidity. Features such as CO_2, plant temperature, given water pH and drained water have negative values and should be removed from the LR sap flow prediction model. The prediction of sap flow by ENR is based on given water/m^2, room temperature, humidity, given water EC and plant temperature; thus, removing CO_2, given water pH and drained water might improve the performance. The most important features for SVR sap flow prediction are given water/m^2, room temperature, given water EC and humidity. By contrast, RF and GB rely on all 7 features: given water/m^2 has the highest FI score, followed by room temperature. The remaining features have similar FI scores (greater than 0 and less than 0.1). The most important features for DT are water amount, room temperature, plant temperature and humidity.

Given water/m^2, room temperature, given water EC and plant temperature contribute the most to the sap flow predictions of the different models. Given water/m^2 has previously

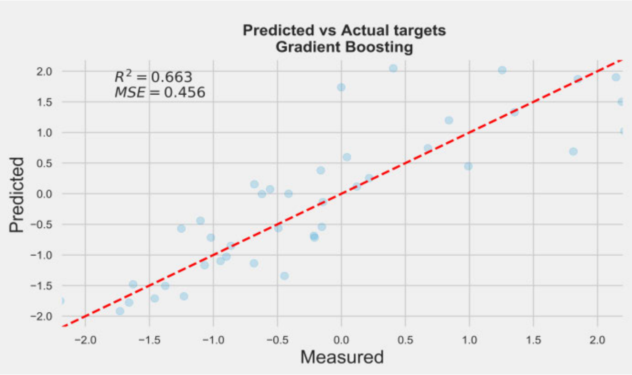

FIGURE 15. The model performance of gradient boosting.

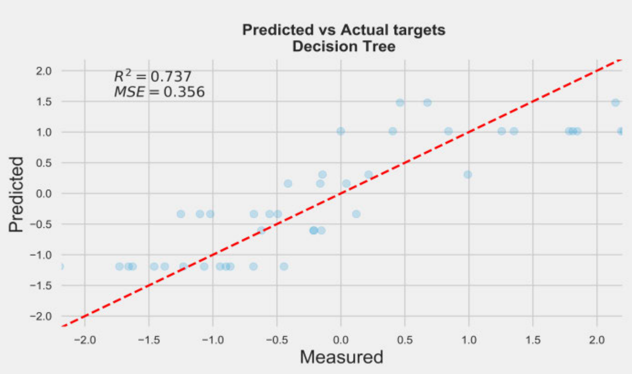

FIGURE 16. The model performance of decision tree.

TABLE 6. Performance evaluation.

MODEL	R^2	Adjusted R^2	MSE	RMSE	MAE
Linear Regression	0.792	0.74	0.281	0.53	0.40
Lasso Regression	0.784	0.73	0.292	0.54	0.40
Elastic Net Regression	0.788	0.74	0.286	0.535	0.40
Support Vector	0.788	0.74	0.287	0.536	0.40
Random Forest	0.808	0.76	0.259	0.509	0.39
Gradient Boosting	0.663	0.58	0.456	0.675	0.49
Decision Tree	0.737	0.67	0.356	0.597	0.51

TABLE 7. Correlation between features and predicted value by different algorithms.

Features	LR	ENR	SVR	RF	GB	DT
Room Temperature	0.281	0.230	0.385	0.133	0.177	0.150
Humidity	0.057	0.046	0.083	0.065	0.070	0.009
CO_2	-0.008	-0.000	-0.050	0.074	0.078	0.000
Plant Temperature	-0.002	0.008	-0.079	0.066	0.071	0.076
Given Water EC	0.125	0.035	0.227	0.027	0.024	0.000
Given Water pH	-0.028	-0.000	-0.032	0.027	0.032	0.000
Water/m^2	0.724	0.6421	0.74610	0.511	0.455	0.764
Drained Water	-0.022	0.000	-0.036	0.095	0.093	0.000

TABLE 8. Hyperparameters of each ML technique.

Model	Hyperparameter
Lasso Regression	α=0.055
Elastic Net Regression	α=0.06
Support Vector Regressor	Kernel= linear
Random Forest	n = 1500 min samples split = 8 min impurity decrease = 0.0
Gradient Boosting	n= 1500, min samples split = 12 min impurity decrease = 0.0 max features = auto max depth= 36 criterion=mae
Decision Tree	max depth=3

been identified as an important feature in the prediction of plant sap flow [41]. The relationship between room temperature and sap flow is also consistent with previous research [42]. Given water EC has previously been found to negatively influence sap flow [43]. Moreover, plant temperature, an indicator of sap flow in this research, has not been reported in previous research. Plant temperature represents stomatal conductance, which is linked to transpiration and plant growth [44]. Furthermore, transpiration is the main driver of sap flow; therefore, theoretically, plant temperature and sap flow may be related. Given water EC indicates the degree of difficulty for plants to absorb water [45]. CO_2 is strongly related to plant photosynthesis and does not show strong correlation with sap flow in this research. However, Remy *et al.* showed that CO_2 concentration exerts a significant negative influence on sap flow [46]. The relationship might be very minimal, which would require an accurate measurement methodology to support it. Given water pH contributes to nutrient uptake, which exhibits no relationship with sap flow. As shown in Table 6, RF offers the best sap flow prediction performance. Moreover, on the basis of Table 7, the predictors CO_2, given water pH and drained water amount should be removed to improve the prediction performance of RF.

V. CONCLUSION

The use of sap flow information can improve water management. Such information allows farmers to easily adapt irrigation strategies, which may help to minimize the waste of resources. In this study, an ML-based prediction system was used to predict sap flow, and the results show that RF performed best. Moreover, the literature has previously shown that RF has high tree diversity, low bias, moderate variance, and minimal problems with overfitting, which contributes to good predictions. LR and ENR also show good performance. Given water/m^2, room temperature, given water EC, humidity and plant temperature were identified as the most important features for sap flow predictions. Among these features, given water/m^2 was the most important variable for RF, and plant temperature was newly identified as an indicator for plant sap flow. A reliable prediction model (with higher R^2 value) for sap flow may contribute to better decision making during the irrigation process. This study will be enhanced in the future, and the dataset will be updated with additional records, including growth parameters such as stem growth, head thickness, and stem thickness.

ACKNOWLEDGMENT

The author Amora Amir thanks all the growers for providing insight into data and the Research Assistant Wenjing Zhao for providing paperwork help and valuable feedback.

REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[2] K. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, ‘‘Machine learning in agriculture: A review,’’ *Sensors*, vol. 18, no. 8, p. 2674, Aug. 2018.

[3] M. Kaul, R. L. Hill, and C. Walthall, ‘‘Artificial neural networks for corn and soybean yield prediction,’’ *Agricult. Syst.*, vol. 85, no. 1, pp. 1–18, Jul. 2005.

[4] T. M. Logan, S. McLeod, and S. Guikema, ‘‘Predictive models in horticulture: A case study with royal gala apples,’’ *Scientia Horticulturae*, vol. 209, pp. 201–213, Sep. 2016.

[5] G. Delgado, V. Aranda, J. Calero, M. Sanchez-Maranon, J. M. Serrano, D. Sanchez, and M. A. Vila, ‘‘Building a fuzzy logic information network and a decision-support system for olive cultivation in Andalusia,’’ *Spanish J. Agricult. Res.*, vol. 6, no. 2, pp. 252–263, 2008.

[6] P. Utkarsha, N. Narkhede, and K. P. Adhiya, ‘‘Evaluation of modified K-means clustering algorithm in crop prediction,’’ *Int. J. Adv. Comput. Res.*, vol. 4, no. 3, p. 1, 2014.

[7] J.-X. Xu, J. Ma, Y.-N. Tang, W.-X. Wu, J.-H. Shao, W.-B. Wu, S.-Y. Wei, Y.-F. Liu, Y.-C. Wang, and H.-Q. Guo, ‘‘Estimation of sugarcane yield using a machine learning approach based on UAV-LiDAR data,’’ *Remote Sens.*, vol. 12, no. 17, p. 2823, Aug. 2020.

[8] T. De Swaef, K. Verbist, W. Cornelis, and K. Steppe, ‘‘Tomato sap flow, stem and fruit growth in relation to water availability in rockwool growing medium,’’ *Plant Soil*, vol. 350, pp. 237–252, Jan. 2011.

[9] K. Vermeulen, K. Steppe, K. Janssen, P. Bleyaert, J. Dekock, J.-M. Aerts, D. Berckmans, and R. Lemeur, ‘‘Solutions to overcome pitfalls of two automated systems for direct measurement of greenhouse tomato water uptake,’’ *HortTechnology*, vol. 17, no. 2, pp. 220–226, Apr. 2007.

[10] C. Giménez, M. Gallardo, and R. B. Thompson, ‘‘Plant-water relations,’’ in *Reference Module in Earth Systems and Environmental Sciences*. Amsterdam, The Netherlands: Elsevier, 2013, doi: 10.1016/B978-0-12-409548-9.05257-X.

[11] Ideal Lab Greenhouse. *Providing Research Facilities and Data for This Study*. Accessed: Feb. 3, 2021. [Online]. Available: https://www.inholland.nl/onderzoek/onderzoeksprojecten/ideal-research-greenhouse-lab

[12] Axia Vegetable Seeds. *Providing the Seedlings for This Study*. Accessed: Feb. 3, 2021. [Online]. Available: https://www.axiaseeds.com

[13] Rijk Zwaan. *Providing the Rootstocks for This Study*. Accessed: Feb. 3, 2021. [Online]. Available: https://www.rijkzwaan.nl

[14] Grodan. *Providing the Rockwool Grodan GT Master for This Study*. Accessed: Feb. 3, 2021. [Online]. Available: https://www.grodan.com

[15] Priva. *Providing the Climate Monitoring System for this Study*. Accessed: Feb. 3, 2021. [Online]. Available: https://www.priva.com

[16] 2GROW. *Providing the Sap Flow Sensors for This Study*. Accessed: Feb. 3, 2021. [Online]. Available: http://2grow.earth/en

[17] Phythosense. *The Software Provided by 2GROW to Access the Sap Flow Data*. Accessed: Feb. 3, 2021. [Online]. Available: http://2grow.earth

[18] O. Obulesu, M. Mahendra, and M. ThrilokReddy, ‘‘Machine learning techniques and tools: A survey,’’ in *Proc. Int. Conf. Inventive Res. Comput. Appl. (ICIRCA)*, Jul. 2018, pp. 605–611.

[19] S. S. Roy, D. Mittal, A. Basu, and A. Abraham, ‘‘Stock market forecasting using LASSO linear regression model,’’ in *Proc. Afro-Eur. Conf. Ind. Advancement*. Cham, Switzerland: Springer, 2015, pp. 371–381.

[20] T. Sirimongkolkasem and R. Drikvandi, ‘‘On regularisation methods for analysis of high dimensional data,’’ *Ann. Data Sci.*, vol. 6, no. 4, pp. 737–763, 2019.

[21] S. Reid and G. Grudic, ‘‘Regularized linear models in stacked generalization,’’ in *Proc. Int. Workshop Multiple Classifier Syst.* Berlin, Germany: Springer, 2009, pp. 112–121.

[22] A. Karatzoglou, D. Meyer, and K. Hornik, ‘‘Support vector machines in R,’’ *J. Stat. Softw.*, vol. 15, no. 9, pp. 1–28, 2006.

[23] A. B. Shaik and S. Srinivasan, ‘‘A brief survey on random forest ensembles in classification model,’’ in *Proc. Int. Conf. Innov. Comput. Commun.* Singapore: Springer, 2019, pp. 253–260.

[24] J. Fan, X. Wang, L. Wu, H. Zhou, F. Zhang, X. Yu, X. Lu, and Y. Xiang, ‘‘Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China,’’ *Energy Convers. Manage.*, vol. 164, pp. 11–102, May 2018.

[25] Y. Y. Song and Y. Lu, ‘‘Decision tree methods: Applications for classification and prediction,’’ *Shanghai Arch. Psychiatry*, vol. 27, no. 2, p. 130, Apr. 2015.

[26] S. Menard, *Applied Logistic Regression Analysis*, vol. 106. Newbury Park, CA, USA: Sage, 2002.

[27] L. F. Leach and R. K. Henson, ‘‘The use and impact of adjusted R2 effects in published regression research,’’ *Multiple Linear Regression Viewpoints*, vol. 33, no. 1, pp. 1–11, 2007.

[28] F. Akdeniz and H. Erol, ‘‘Mean squared error matrix comparisons of some biased estimators in linear regression,’’ *Commun. Statist.-Theory Methods*, vol. 32, no. 12, pp. 2389–2413, Jan. 2003.

[29] C. J. Willmott and K. Matsuura, ‘‘Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance,’’ *Climate Res.*, vol. 30, no. 1, pp. 79–82, Dec. 2005.

[30] T. Chai and R. R. Draxier, ‘‘Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature,’’ *Geoscientific Model Develop.*, vol. 7, no. 3, pp. 1247–1250, 2014.

[31] A. E. Eiben and S. K. Smit, ‘‘Parameter tuning for configuring and analyzing evolutionary algorithms,’’ *Swarm Evol. Comput.*, vol. 1, no. 1, pp. 19–31, Mar. 2011.

[32] P. Probst, B. Bischl, and A.-L. Boulesteix, ‘‘Tunability: Importance of hyperparameters of machine learning algorithms,’’ 2018, *arXiv:1802.09596*.

[33] J. Brownlee. (2020). *Train-Test Split for Evaluating Machine Learning Algorithms*. Machine Learning Mastery. [Online]. Available: https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/

[34] K. Steppe, M. W. Vandegehuchte, R. Tognetti, and M. Mencuccini, ‘‘Sap flow as a key trait in the understanding of plant hydraulic functioning,’’ *Tree Physiol.*, vol. 35, no. 4, pp. 341–345, Apr. 2015, doi: 10.1093/treephys/tpv033.

[35] J. Čermák, J. Kučera, and N. Nadezhdina, ‘‘Sap flow measurements with some thermodynamic methods, flow integration within trees and scaling up from sample trees to entire forest stands,’’ *Trees*, vol. 18, no. 5, pp. 529–546, Sep. 2004.

[36] J. C. Suárez, F. Casanoves, M. A. N. Bieng, L. M. Melgarejo, J. A. Di Rienzo, and C. Armas, ‘‘Prediction model for sap flow in cacao trees under different radiation intensities in the western Colombian Amazon,’’ *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, 2021.

[37] *Measuring Sap Flow in Small Stems, Branches and Petioles*. Accessed: May 5, 2021. [Online]. Available: https://edaphic.com.au/sap-flow-in-small-stems

[38] N. Donges. (2019). A complete guide to the random forest algorithm. Build In. Accessed: May 5, 2021. [Online]. Available: https://builtin.com/data-science/random-forest-algorithm

[39] J. Kho. (2018). *Why Random Forest is My Favorite Machine Learning Model*. Towards Data Science. [Online]. Available: https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706

[40] P. Płoński. (2019). *Does Random Forest Overfit*. Mljar. [Online]. Available: https://mljar.com/blog/random-forest-overfitting

[41] R. Qiu, T. Du, S. Kang, R. Chen, and L. Wu, "Influence of water and nitrogen stress on stem sap flow of tomato grown in a solar greenhouse," *J. Amer. Soc. Horticultural Sci.*, vol. 140, no. 2, pp. 111–119, Mar. 2015.

[42] P. G. Oguntunde and N. van de Giesen, "Water flux measurement and prediction in young cashew trees using sap flow data," *Hydrol. Processes*, vol. 19, no. 16, pp. 3235–3248, Oct. 2005.

[43] E. Jeon, S. Baek, S. Choi, K. S. Park, and J. Lee, "Real-time monitoring of electroconductivity in plants with microscale needle probes," *Environ. Control Biol.*, vol. 56, no. 4, pp. 131–135, Oct. 2018.

[44] J. Cai and M. Cespedes, "Plant temperature measurement and analysis from infrared images," in *Proc. 27th Conf. Image Vis. Comput.*, 2012, pp. 406–411.

[45] E. Van Os, C. Blok, W. Voogt, and L. Waked. (2016). *Water Quality and Salinity Aspects in Hydroponic Cultivation*. [Online]. Available: https://edepot.wur.nl/403810

[46] R. Manderscheid, M. Erbs, S. Burkart, K.-P. Wittich, F.-J. Löpmeier, and H.-J. Weigel, "Effects of free-air carbon dioxide enrichment on sap flow and canopy microclimate of maize grown under different water supply," *J. Agronomy Crop Sci.*, vol. 202, no. 4, pp. 255–268, Aug. 2016.

AMORA AMIR was born in Baghdad, Iraq, in 1979. She received the B.S. degree in computer science from the University of Baghdad, in 2002, and the M.Sc. degree in computer engineering from the Delft University of Technology, The Netherlands, in 2012. From 2015 to 2017, she was a Lecturer Expert on big data with the Computer Sciences Department, Inholland University of Applied Sciences. Since 2018, she has been a Senior Researcher with the Data Driven Smart Society Research Group, Inholland University of Applied Sciences, where she has also been working with the Agriculture, Horticulture and Business Department. She is the developer of the Ideal Research Greenhouse Laboratory, which is part of the program of smart farming at the Inholland University of Applied Sciences. Her research interests include big data and artificial intelligence, machine learning and applications in horticulture, specifically greenhouses, and studying plants through the data from shared value for business and vertical farming.

MARYA BUTT was born in Rawalpindi, Pakistan, in 1981. She received the B.S. degree in software engineering, in 2004, the master's degree (Hons.) in governance and organizational sciences, in 2007, and the Ph.D. degree in the emerging field of electronic governance from Utrecht University, The Netherlands, in 2015. She wrote four research articles as a part of her Ph.D. research. Since 2019, she has been working with the Faculty of Engineering, Design and Computing Inholland University of Applied Sciences, as a Lecturer/Researcher, where she is also associated with the Data Driven Smart Society and the Robotics Research Group. Her research interests include image localization and segmentation, multi-class classification, and multi-label classification using deep learning and computer vision in smart farming and health.

OLAF VAN KOOTEN was born in The Netherlands, in 1952. He studied at the Escola Americana do Rio de Janeiro, from 1964 to 1967. He received the M.Sc. degree in experimental physics from Vrije Universiteit Amsterdam, in 1980, and the Ph.D. degree in biophysics in plant physiology from Wageningen University & Research, in 1988. From 1989 to 1999, he was the Department Head of Postharvest Physiology and Technology at Wageningen University & Research, where he became a Professor of horticultural production chains and produce quality control throughout supply chain, in 1999. He was a Representative of The Netherlands on the ISHS Council, from 2000 to 2014. Since 2011, he has been lecturing at the Inholland University of Applied Sciences, Delft. His research interests include the transition from pre-harvest to post-harvest in product physiology and product quality.