biology and life sciences forum

MDPI

Proceeding Paper

Performance of Gradient Boosting Learning Algorithm for Crop Stress Identification in the Greenhouse Cultivation †

Angeliki Elvanidi and Nikolaos Katsoulas *

University of Thessaly, Department of Agriculture Crop Production and Rural Environment, Fytokou Str., 38446 8 Volos, Greece; elaggeliki@gmail.com

* Correspondence: nkatsoul@uth.gr; Tel.: +30-24210-93249

† Presented at the 1st International Electronic Conference on Horticulturae, 16–30 April 2022; Available online: https://iecho2022.sciforum.net/.

Abstract: Greenhouse cultivation is one of the most crucial circular economy systems in agriculture that allows the maximum production in less cultivation area, with minimum inputs and low environmental impact. The data generated in high-tech and sophisticated greenhouse operations are provided by a variety of different sensors that enable a better understanding of the operational environment. In this study a learning algorithm namely Gradient Boosting Machine was tested using the generated data-base in order to estimate different type of stress in tomato crop. The examined model perform qualitative classification of the data, depending on the type of stress (such as no stress, water stress and cold stress). For the comparison was selected 10-fold cross validation strategy on the 10,763 samples from the training set. The dataset was divided in two parts, one for training-validation 80% (8610) and a second one for testing 20% (2152). The cross-validation process was repeated 50 times. Among the data entries was used to build the model, the leaf temperature was one of the highest in the feature importance with ratio 0.51. According to the results, the Gradient Boosting algorithm defined all the cases with high accuracy. Particularly, the model find correct all the 372 samples of the cold stress plants, the 1305 out of 1321 samples of the no stress plants and the 431 out of 452 samples of the water stress plants. In these results, the model preserved accuracy of 98% in the testing performance and more than 98% in the validation performance. This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the project "Reinforcement of Postdoctoral Researchers—2nd Cycle" (MIS-5033021), implemented by the State Scholarships Foundation (ΙΚΥ).

Keywords: remote sensing; photochemical reflectance index; decision tree; stress detection; real time

Citation: Elvanidi, A.; Katsoulas, N. Performance of Gradient Boosting Learning Algorithm for Crop Stress Identification in the Greenhouse Cultivation. *Biol. Life Sci. Forum* **2022**, 2, x. https://doi.org/10.3390/xxxxx

Academic Editor(s):

Published: 16 April 2022

1. Introduction

The most important call for a sustainable future in the food sector is to produce more food per hectare without expanding agricultural land in order to feed the rapid growth of the world population. To achieve this, there is a need to increase the productivity in Greenhouse hydroponic cultivation by redesigning their operation control system [1].

The development of a machine learning model (ML) that will combine climatic and crop physiology data for detecting different type of stress will result to the improvement of the greenhouse operation.

Up to now, it wasn't feasible to incorporate in a ML model crop physiology data, since the most agronomy factors are measured using labor and time-consuming protocols [2]. Leaf temperature is one of the few indicators that can be measured in a time-series protocol producing a large volume database required in order to build a machine learning model. However leaf or crop temperature is an unstable factor that can present an intense

variation according to the climatic and abiotic conditions and cannot be used on its own to estimate different types of crop stress [3]. The combination of leaf temperature with the photosynthesis (P_s) could improve the methodology of defining the type of stress performed in a vegetable cultivation.

Recently, the Photochemical Reflectance Index (PRI) that is correlated with rapid changes in de-epoxidation of the xanthophylls cycle and photosynthesis efficiency (P_s) with very good results, is able to be measured using *soft-sensor* (i.e., mathematical models using real-time sensor data) performing a time-series database.

In this research the methodology of developing a Gradient Boosting algorithm is presented in order to build ML model that will combine climatic data with leaf temperature and photosynthesis rate. In this sense, a multisensory tower placed within the greenhouse was used to record how the physiology status of the tomato plants was changing according to their surrounding microclimate. The plants were cultivated under extreme conditions of air temperature and water in the root zone. The resulted database was used to train and test the model.

2. Material & Methods

The measurements were carried out from May to December of 2019 in one of the five compartments of a multi-tunnel greenhouse with a total ground area of 1500 m^2 (250 m^2 each compartment). The establishments were located at the facilities of the University of Thessaly, Velestino, Volos (Latitude 39° 22′, longitude 22° 44′ and altitude 85 m), in the continental area of eastern Greece.

The tomato plants (*Solanum lycopersicum*cv. Elpida) were cultivated in slabs filled with perlite slabs (ISOCON Perloflor Hydro 1, ISOCON S.A., Athens, Greece). The plants were fertigated with fresh nutrient solution with set-points of electrical conductivity (EC) around 2 dS m^{-1} and pH 5.8. The nutrient solution supplied to the crop was a standard nutrient solution for tomato grown in open hydroponic systems adapted to Mediterranean climatic conditions. The nutrient solution was supplied via a drip system and was controlled by a time-program irrigation controller (8 irrigation events per day).

In order to record the physiological response of the plants to their surrounding microclimate, tomato plants were imposed to different types of stress. Specifically, the plants were cultivated under (i) low air temperature around 15 °C (LTS treatment), (ii) low-water concentration in the root zone with dose 30 mL per plant (LWS treatment). Additionally, measurements of (iii) no stressed (NoS treatment) plants were also recorded.

In order to build the data-base of crop physiology and environment microclimate under the mentioned extreme conditions, a multisensory tower was build consisted by by air temperature sensor (Thygro SDI-12, Symmetron, Greece) relative humidity (Thygro SDI-12, Symmetron, Greece), solar radiation (SP-SS, Apogee instruments, USA), leaf temperature sensor (Thermocouples, type T), leaf wetness (PS-0061-AD, Netsense, Italia) and PRI sensor (type SRS-PRI, Meter group, USA). The multisensory was placed within the greenhouse in parallel with vertical axis of tomato main stem. The measurements started 10 days after the day of each treatment was applied and lasted for 25 days.

Totally, 9 features: air temperature (T_a, °C), relative humidity (RH, %), solar radiation (SR, W m^{-2}), leaf temperature (T_L, °C), leaf wetness in young leaves (Lw_{up}, %), leaf wetness in mature leaves (Lw_{dn}, %), photochemical reflectance index (PRI), photosynthesis rate (P_s, μmol m^{-2} s^{-1}) and crop water stress index (CWSI) were added to the model in order to exist three outputs (LTS, LWS and NoS).

In the current research, the CWSI developed by Jackson et al. [4] was calculated. The methodology followed in the current research is described in Baille et al. [5]. The calibration procedure of remote PRI sensor and how P_s is calculated was presented in Elvanidi & Katsoulas [6]. The resulted data-sample was of 10,763 values.

In order to obtain high performance in greenhouse data, a series of ML algorithms like gradient boosting (GB), multilayer perceptron (MLP) and other artificial neural network algorithms were examined. Among the algorithms, the GB technique corresponded

more sufficient in the studied tested sample where the measurable parameters were defined as distinct and not as time-series. The GB modeling part of the ensemble learning algorithms that rely on a collective decision from inefficient prediction models is called decision trees.

In the model, a list of hyperparameters were used (Learning rate, Number of estimators, Max tree depth, Max features). The cross-validation process was repeated 50 times. The methodology was followed in the current research is described in Friedman et al. [7], Khan et al. [8] and Karamoutsou [9].

The dataset was divided in two parts, one for training-validation 80% (8610) and a second one for testing 20% (2152). All steps, learning and classification were written in Python. For machine learning, the Python ML Scikit-learn [10] library and the Spyder environment were used.

The statistical criteria concern the Accuracy (1), Positive predictive values (PPV or Precision) (2), Sensitivity (or Recall) (3) and F1 (F1-score) (4) (where P is the number of real positive cases in data and N the number of real negative cases in data) were used:

$$\text{Accuracy} = TP + TN / (P + N) \quad (1)$$

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

$$\text{Sensitivity} = TP / (TP + FN) \quad (3)$$

$$F1 = 2\,(\text{Precision} * \text{Sensitivity}) / (\text{Precision} + \text{Sensitivity}) \quad (4)$$

3. Results

During the training procedure, it was defined the optimum rates of each hyperparameter.

The range values of learning rate was 0–1, with the most common values being 0.001–0.3. Smaller values made the model robust to the specific characteristics of each individual tree and reduced the possibility of overfitting. However, the low values increase the risk of not reaching the optimum with a fixed number of trees. For the development of the current GB-based classifier, the optimum value that have been chosen among the above learning values (0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 1) was 0.5.

The optimum number of estimators in which the total number of sequential trees was defined have been chosen among the values (10, 20, 30, 40, 50, 60, 70, 80, 90, 100) and was 70.

In the Max tree depth indicator in which the depth of the individual tress was controlled, the optimum value has been chosen among the values (1, 2, 3, 4, 5, 6, 7, 8, 9, 10) and was 9.

The optimum Max features indicator that defines the number of features that will be used for a best split was chosen among the values (1, 2, 3, 4, 5, 6, 7, 8) and was 7.

The combination of the optimum hyperparameters developed the current GB algorithm for detecting the three specific types of stress. According to the training process, the number of features that will be imposed to the model were defined to 9.

Figure 1 shows in histograms the feature importance values obtained from GB approach. It is observed that out of the 9 features, two features improve the present models to classify the three types of stress, namely a) T_L and b) T_a. The other characteristics complement the forecasting process by further improving the model. Therefore, in the current algorithm, the more variables were performed as an input, the higher was the predictor accuracy. For decreasing the number of inputs there is a need to increase the testing sample since the greenhouse system is considered as a non-linear system, where the lack of datasets produces a very complex dynamic relation between the climatic factors and the crop physiology response difficult to predict.

Table 1 presents the statistical criteria performed in GB algorithm during training and testing process. According to the data, the GB algorithm performed high criteria in

the training set where the Accuracy, Precision, Sensitivity and F1-score was 100%. GB belongs to the family of models that can handle even features with low predictive power. In addition, the GB model was found to have high performance in the test set with 98% Accuracy, 98% Precision, 98% Sensitivity and 98% F1. Comparison of the metrics between the training and testing phase shows that overfitting was avoided.

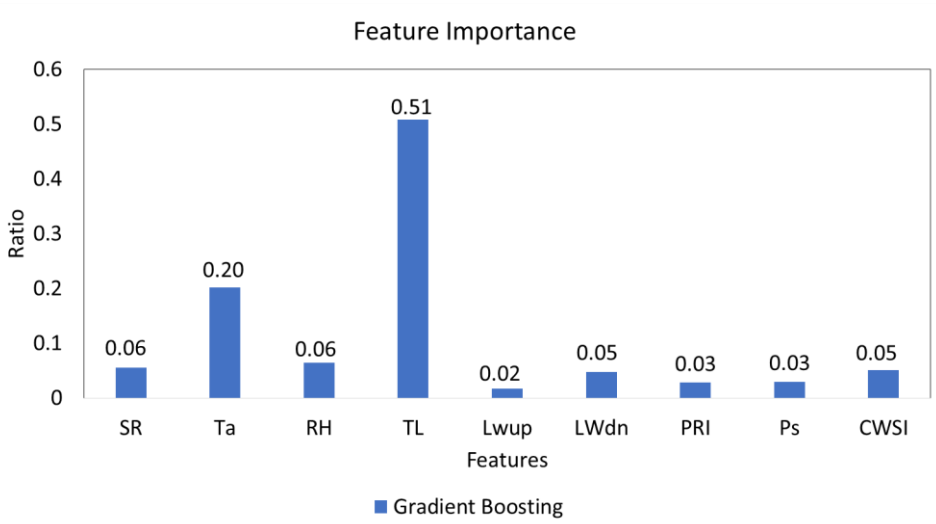

Figure 1. Feature importance of the measured factors in the set-up of GB algorithm.

Table 1. Statistical criteria resulted from (a) the validation (training sample 8610) and (b) the performance (testing sample 2152) of GB algorithm.

Performance and Validation				
GB Algorithm	**Accuracy**	**Precision**	**Recall**	**F1**
Trainning set	100%	100%	100%	100%
Testing set	98%	98%	98%	98%

Figure 2 shows the performance distribution for the GB model according to the three types of stress. More specifically, the GB model correctly "understood" all cases presented as LTS, it "confused" 16 NoS cases as LWS, 21 LWS cases as NoS, and only 1 LWS case as LTS.

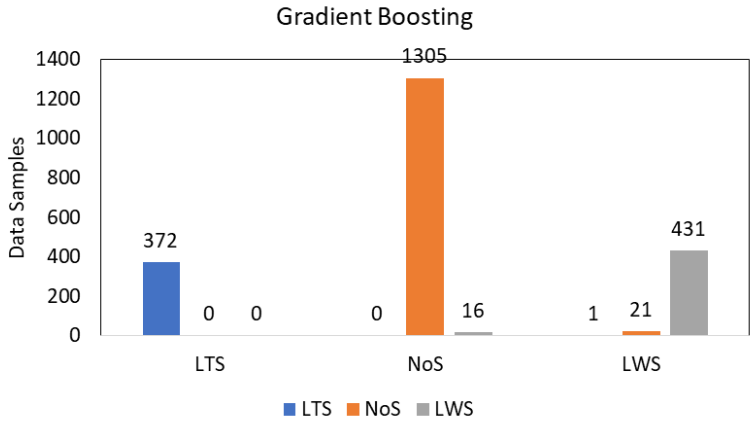

Figure 2. The predicted category of the samples of each treatment according to the type stress for GB algorithm in the testing process (testing sample 2152).

4. Discussion

Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. Gradient boosting algorithm can be used for predicting not only continuous target variable (as a Regressor) but also categorical target variable (as a Classifier). In the current research, quality and quantitative data are involved in the process of building ML model. Additionally, GB can build highly efficient, more accurate and high quality of ML model in a less time. GB performs well under small weak size of datasets and unbalanced data like real time data management [11,12]. Ravi and Baranidharan [13] and Cai et al. [14] sustain that GB is faster from all over machine learning algorithms.

In the current research, the GB algorithm was performed for the first time ever to classify qualitative and quantitative data under greenhouse conditions with very good statistical results. The developed model can be applicable in other greenhouse systems of Mediterranean region that cultivate tomato crop in hydroponics.

The next step of the current research is to improve the model was developed by GB algorithm by decreasing the number of inputs in order to define more type of stress like the stress is occurred in the plants due to high air temperature and low nutrient performance.

5. Conclusions

The current research presented the development of Gradient Boosting algorithm to predict three types of stress under greenhouse conditions. The model was made for tomato crop, while the training and the testing of the models was performed in a sample of 10,763 datasets. In the model, 9 features inputs were adjusted for predicting 3 outputs. The developed GB model presented high statistical criteria more than 98% accuracy, performing high sustainability in greenhouse data able to be connecting with the operation systems already used. Future perspective of the current research is to extend the model in order to predict more than three type of stress. Application of the current model in greenhouse cultivation allows more efficient and precise farming with less human manpower with high quality production contributing to the further reduction of the resource's inputs, energy and environmental footprint.

Author Contributions: Conceptualization, N.K.; methodology, N.K. and A.E.; formal analysis, A.E.; investigation, A.E.; resources, N.K. and A.E.; data curation, A.E.; writing—original draft preparation, A.E.; writing—review and editing, N.K. and A.E.; supervision, N.K.; project administration, N.K.; funding acquisition, N.K. and A.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research is co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the project "Reinforcement of Postdoctoral Researchers—2nd Cycle" (MIS-5033021), implemented by the State Scholarships Foundation (ΙΚΥ).

Institutional Review Board Statement:

Informed Consent Statement:

Data Availability Statement:

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Elvanidi, A.; Benitez Reascos, C.M.; Gourzoulidou, E.; Kunze, A.; Max, J.F.J.; Katsoulas, N. Implementation of the circular economy concept in greenhouse hydroponics for ultimate use of water and nutrients. *Horticulturae* **2020**, *6*, 83.
2. Katsoulas, N.; Elvanidi, A.; Ferentinos, K.P.; Kacira, M.; Bartzans, T.; Kittas, C. Crop reflectance monitoring as a tool for water stress detection in greenhouses: A review. *Biosyst. Eng.* **2016**, *151*, 374–398.
3. Katsoulas, N.; Savas, D.; Tsirogiannis, I.; Merkouris, O.; Kittas, C. Response of an eggplant crop grown under Mediterranean summer conditions to greenhouse fog cooling. *Sci. Hortic.* **2009**, *123*, 90–98.

4. Jackson, R.D.; Idso, S.B.; Reginato, R.J.; Pinter, P.J. Canopy temperature as a crop water stress indicator. *Water Resour. Res.* **1981**, *171*, 133–138.
5. Baille, A.; Kittas, V.; Katsoulas, N. Influence of whitening on greenhouse microclimate and crop energy. *Agric. For. Meteorol.* **2009**, *107*, 293–306.
6. Elvanidi, A.; Katsoulas, N. Calibration methodology of remote PRI sensor for plant photosynthesis rate status assessment in greenhouse. In Proceedings of the 1st International Electronic Conference on Agronomy, 2021. https://doi.org/10.3390/IECAG2021-10018.
7. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001,** *29*, 1189–1232. https://doi.org/10.1214/aos/1013203451.
8. Khan, R.; Mishra, P.; Baranidharan, B. Crop Yield Prediction using Gradient Boosting Regression. *Int. J. Innov. Technol. Explor. Eng. (IJITEE) ISSN* **2020**, *9,* 2278–3075.
9. Karamoutsou, L. Investigation of the Water Quality Parameters of Lake Kastoria from Time-Series Monitoring Data Using Machine Learning Techniques for Simulation and Prediction. PhD Thesis, University of Thessaly, 2020.
10. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B. Scikit-learn: Machine learning in Python*, J. Mach. Learn. Res.* **2011**, *Volume*, page number.
11. Puligudla, P.; Karthik, K.S.; Kumar, K.V.N.; Thirugnanam, M. Prediction of crop yield using gradient boosting. *J. Xi'an Univ. Archit. Technol.* **2020**, *XII*, 2020. ISSN 1006-7930.
12. Shyamala, K.; Rajeshwar, I. Enhanced gradient boosting regression tree for crop yield prediction. *Int. J. Sci. Technol. Res.* **2020**, *9*, page number.
13. Ravi, R.; Baranidharan, B. Crop yield Prediction using XG Boost algorithm. *Int. J. Recent Technol. Eng. (IJRTE) ISSN* **2020**, *8*, 2277–3878.
14. Cai, W.; Wei, R.; Xu, L.; Ding, X. A method for modelling greenhouse temperature using gradient boost decision tree. *Inf. Process. Agric.* **2021**, *Volume*, page number. https://doi.org/10.1016/j.inpa.2021.08.004.