Machine Learning Models Comparison for Water Stress Detection Based on Stem Electrical Impedance Measurements

Machine Learning Models Comparison for Water Stress Detection based on Stem Electrical Impedance Measurements

Federico Cum*, Stefano Calvo*, Danilo Demarchi* and Umberto Garlando*
*Department of Electronics and Telecommunications (DET), Politecnico di Torino, Torino, Italy

Email: umberto.garlando@polito.it

***Abstract*—Smart agriculture is a promising solution to improve food production and reduce waste of resources. The idea is to adopt electronics and sensors to monitor key parameters of the crops and integrate these data with farmer knowledge. Sensors monitor both the environment and the plant itself, generating a huge amount of data. Data processing is a key aspect of smart agriculture, and machine learning can help to understand the data and extract the needed feature. In this paper, we present a performance comparison of several machine learning models trained to detect the water stress condition of plants. The dataset used for this study includes the stem electrical impedance, a novel parameter directly measured on the plants. The machine learning models are compared based on three different metrics, and the average accuracy is higher than 85%. The effect of removing the stem electrical impedance results in worse performance of the models, indicating its impact in the application.**

***Index Terms*—Machine learning, support vector machines, decision tree, random forest, smart agriculture, food security**

I. Introduction

In recent years, the world has been facing the effects of global warming. The progressive rise in global temperatures, primarily triggered by the release of greenhouse gases into the atmosphere, is exacerbating the desertification phenomenon, placing an increasing number of lands at risk and reducing the available grounds for cultivation [1]. Moreover, the global population is constantly growing, and the projections suggest it is expected to surpass 10 Billion by 2050 [2]. Global food security is a big concern as climate change, and the subsequent lack of water affect crop production's potential yield [3]. Smart agriculture, also known as precision agriculture or digital farming, offers advanced technologies and data-driven approaches in agricultural practices to optimize and enhance various aspects of crop production and livestock management. It leverages innovative technologies such as the Internet of Things (IoT), artificial intelligence (AI), drones, sensors, and big data analytics to monitor, collect, and analyze real-time data about soil conditions, weather patterns, crop growth, and livestock health [4]. One of the main aspects to consider when building an efficient and automatic monitoring system for smart agriculture is assessing plants' health. Monitoring plants involves detecting and diagnosing various factors that can affect their growth, such as nutrient deficiencies, diseases, pests, and environmental stressors. Modern IoT systems with nodes connected inside a Wireless Sensor Network (WSN) can help farmers by inspecting the surrounding environmental conditions providing valuable information for efficient cultivation management. On the other hand, it is also essential that monitoring is not limited only to the plant's surroundings but to one or more parameters directly extracted from it.

In [5] *Garlando et al.* propose a novel approach to directly evaluate plants' water stress status by measuring the value of stem impedance. This approach is explored in our work, and environmental and stem impedance measurements are used as features to build binary classifiers able to identify water stress conditions. This study evaluates and compares the results obtained using different learning algorithms such as Support Vector Machine (SVM), Decision Tree, Random Forest, and Artificial Neural Networks to asses water stress in tobacco plants. This paper is organized as follows: Section II summarizes the existing literature on machine learning applied in smart agriculture. Section III presents how data in this study are collected and how machine learning models are used to evaluate tobacco plants' water stress. Then, in section IV, the results are presented. Finally, section V draws conclusions and future perspectives.

II. Related Work

In recent years, many machine learning techniques have been applied to various aspects of smart agriculture, such as crop monitoring, yield prediction, disease detection, and resource management [6] [7]. Regarding disease and pest detection, computer vision is widely used: for instance, in [8], the authors focus on utilizing Convolutional Neural Networks (CNN) to identify various diseases in potatoes. Similarly, in [9], a hybrid approach that combines CNN and SVM is explored to detect diseases in rice leaf images.

Kerkech et al. [10] used visible and infrared images collected by an Unmanned Aerial Vehicle (UAV) to build CNN based model able to detect vine diseases both at grapevine and at leaf level.

Much effort is put into optimizing the irrigation procedure; for example, authors in [11] employ a neural network to build an irrigation prediction model based on light intensity, external temperature, soil conductivity, and soil moisture.

In their study, Bettelli et al. [12] conducted direct plant monitoring by employing a sensor named "bioristor." They effectively utilized the gathered data to classify and forecast water stress in tomato plants by applying decision trees, random forests, and recurrent neural networks.

In [13], authors exploited neural networks to highlight the importance of impedance as an indicator of plants' well-being, and this approach is extended in this research by adding different machine learning algorithms, namely support vector machine, decision tree, and random forest.

III. Methodology

The first step in this research is data collection from tobacco plants. Following that, it is crucial to preprocess the data, employing operations such as standardization and division into train and test portions. This preprocessing ensures that the measurements are appropriate for feeding into the learning algorithm and facilitates the evaluation of the models. The models are built using Python 3.9.15 and the SciKit-Learn framework [14], and training is run on a laptop with the following characteristics: Intel Core i7-8750H 8th generation processor, Graphical Processing Unit (GPU) NVIDIA GeForce GTX 1050, 16 GB of RAM and Microsoft Windows 11 as the operating system.

Models' performance is evaluated by splitting the dataset using a 5-fold cross-validation technique. Three binary classification metrics are calculated: accuracy, F1 score, and Area Under the ROC Curve (AUC) [15]. A block diagram showing the methodology employed in this work is depicted in figure 1.

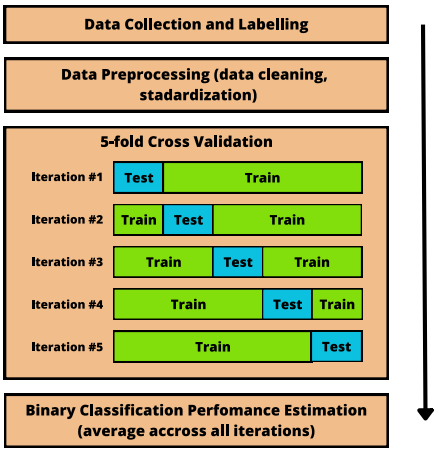

Fig. 1. Methodology block diagram

A. Data collection

Data analyzed in this paper is obtained using the setup described in [5] and [16]. Impedance modulus and phase measurements are performed using a *Keysight 4294a* bench analyzer with a four-point probe technique. The operating frequency at which impedance is measured is set to 10.145 KHz, where the best trade-off between sensitivity and noise mitigation is available [16]. Environmental measurements (temperature, air humidity, soil water potential, and ambient light) and impedance value (modulus and phase) are sampled every hour; moreover, a plant picture is taken to allow visual inspection and subsequent labeling of plant status. Up to 4 plants can be connected and monitored by the system simultaneously. Plants must be subjected to different conditions to create a meaningful dataset: initially, each plant is kept regularly watered; then, drought stress is induced by water deprivation. Pictures taken at each measurement are analyzed, and a label "water stress" is assigned to a sample where the plant shows signs of drought (such as low leaves turgidity) and a label "water ok" when there are no apparent signs of water stress. Overall data is collected from 24th March 2021 to 28th July 2021 for 6306 samples in total. Figure 2*a* shows a pie chart representing the distribution of the labels inside the dataset, showing that both classes are adequately represented with a slight prevalence of "Water stress" labels (52.7 %) with respect to "Water ok (47.3%).

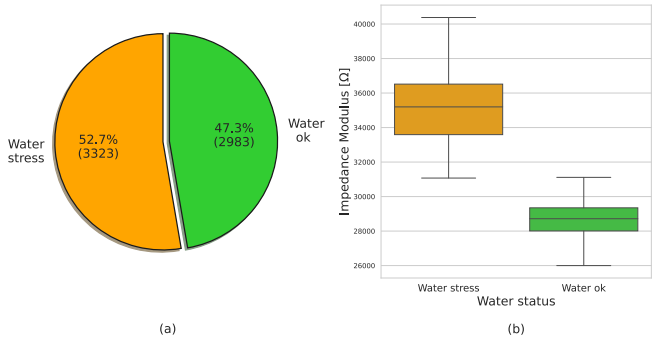

Fig. 2. (a) Pie chart showing the distribution of labels in the dataset. (b) Box plot of impedance modulus distribution.

As an initial analysis, comparing the stem impedance module values in presence of water stress is worthwhile since it can help gain interpretability on the trained models. Each plant shows a different value ranges for impedance modulus. However, the emerging trend is that when the plant is subject to drought conditions, this value tends to rise compared to regular watering conditions, as demonstrated by authors in [5].

Figure 2*b* illustrates a boxplot showcasing the distribution of impedance module values under the two conditions highlighted in this paper for a single plant contained in the dataset. Machine learning models can automatically use this information and calculate a sophisticated decision boundary when combined with environmental variables, providing detection capability of the plant's hydration condition.

B. Data pre-processing

This preparatory step is imperative due to the varying ranges of features, which often exhibit different orders of magnitude. If data preprocessing is not performed, it can result in the dominance of larger features within the model. Many solutions to this problem exist, but normalization and standardization are

the most common. Normalization rescales the features in the $[0-1]$ range, while standardization rescales the features to have zero mean and unit variance.

In this paper, the standardization technique is employed, and this is achieved by applying the following formula:

$$x_{scaled,stand} = \frac{x - \mu_x}{\sigma_x} \tag{1}$$

In this context, x represents the actual feature value, μ_x denotes the mean value across the entire dataset, and σ_x represents the standard deviation. It is crucial to emphasize that the mean and standard deviation should only be computed using the training portion of the dataset. Subsequently, these calculated values must be applied consistently to the training and test portions. Failing to do so would result in information leakage into the test dataset, which could compromise the reliability of the quality evaluation for the model.

C. Machine learning models

In this work, the analyzed models include support vector machine, decision tree, random forest, and neural network. Performance evaluation is conducted using a 5-fold cross-validation technique, wherein the dataset is divided into 5 folds. Iteratively, 4 folds are employed as the training set, while the remaining fold serves as the test set. This process is repeated 5 times, each altering the combination of the portions used for training and testing. The final metrics to assess the models are the average values across all fold combinations.

The considered evaluation metrics are accuracy, F1-score, and AUC.

Accuracy is calculated by the ratio between the total correct predictions over the total number of samples considered.

F1-score is the harmonic mean between precision and recall, and it is evaluated as

$$F_1 = 2 \cdot \frac{prec \cdot rec}{prec + rec}$$

where precision is calculated with

$$prec = \frac{true\ positives}{true\ positives + false\ positives}$$

and recall

$$rec = \frac{true\ positives}{true\ positives + false\ negatives}$$

F1 score is helpful when the dataset is not well balanced between the number of instances for every class. In those cases, accuracy can be misleading due to the possible high number of correct predictions for the majority class.

AUC stands for Area Under the Receiver Operating Characteristic (ROC) Curve, a plot that illustrates the performance of a binary classifier varying the discrimination threshold. AUC can be interpreted as the probability that a random sample belonging to the negative class has a lower probability of being classified as positive than the one of a random example belonging to the positive class [17].

Initially, all features, including environmental and impedance ones, are considered. Subsequently, the impedance modulus and phase are alternately excluded, leading to three distinct feature subsets. This process ensures that at least one feature directly extracted from the plant is retained to prevent classifiers from relying solely on the plant's surroundings. In this way, it is possible to evaluate the importance of impedance for detecting water stress.

The models evaluated are described here.

a) Support Vector Machine (SVM): Support vector machine (SVM) is a supervised machine learning that can be used for regression and classification tasks. In particular, given a binary classification problem, an SVM aims to find a hyperplane or decision boundary that can separate data points in classes. If data are not linearly separable, a hyperplane can still be found by applying kernel functions.

For this model, the training parameters used are a regularization parameter $C = 1$ and radial basis function as kernel.

b) Decision Tree: The decision tree classifier is an often-used supervised machine learning algorithm for classification tasks. This model takes input data conditions or features and makes decisions in a flowchart-like manner. By traversing a binary tree structure, the decision tree classifier predicts the class label of a given sample. Each node in the tree can be a decision node, where a specific feature is used to make a decision or a leaf node that represents a class prediction. The decision tree is designed to efficiently categorize input data by utilizing a criterion that measures the effectiveness of decision splits in dividing the data into classes. The two most common splitting criteria are Gini impurity and information gain [18]. The trained decision tree models employ Gini impurity as splitting criterion in this work.

c) Random Forest Classifier (RF): In machine learning, ensembles enhance performance by combining multiple models. One such ensemble technique is the random forest, which merges multiple decision trees and can be applied to both regression and classification tasks. The concept involves generating numerous decision trees, each trained on a random subset of the training data, and constructing the trees using only a subset of features for their nodes. This approach mitigates the risk of overfitting, as each tree is trained on distinct samples and feature subsets, thereby reducing the model's sensitivity to specific patterns or noise within the dataset. Each random forest model trained during this work employs 100 decision trees.

d) Artificial Neural Network: The artificial neural network is a learning model inspired by the functioning of the human brain. It can be employed in a wide range of problems and comprises different layers of connected artificial neurons. The input layer receives the initial inputs, propagating through one or more hidden layers, ultimately reaching the output layer that generates the desired output. The network training is performed by minimizing the difference between the actual and the predicted output using the backpropagation technique. The following hyperparameters are used for training: 2 hidden layers with 10 neurons, batch size equal to 20, 300 training epochs, and ReLU as activation function.

IV. Results

Each model is accompanied by a table that presents the mean and standard deviation of the evaluation metrics calculated over the 5 cross-validation folds.

A. SVM results

TABLE I
PERFORMANCE OF SUPPORT VECTOR MACHINE CLASSIFIER

	Acc(%)	F1(%)	AUC(%)
All features	86.53 ± 1.10	85.63 ± 1.27	93.30 ± 0.57
No imp. phase	83.22 ± 0.48	81.99 ± 0.54	92.10 ± 0.46
No imp. modulus	81.26 ± 1.13	76.37 ± 1.46	86.52 ± 0.82

Table I illustrates the outcomes achieved with the support vector machine (SVM). The overall results are promising, as a classification accuracy of 86.53 % is attained when utilizing impedance and environmental measurements. However, upon removing the impedance phase, there is a decline in performance by 3.31%. Similarly, excluding the impedance modulus leads to a decrease in accuracy by 5.27% compared to the initial scenario. These findings suggest that impedance is a relevant feature for assessing water stress, as evidenced by the model's inferior performance when these features were eliminated from the dataset.

B. Decision Tree results

TABLE II
PERFORMANCE OF DECISION TREE CLASSIFIER

	Acc(%)	F1(%)	AUC(%)
All features	96.69 ± 0.51	96.50 ± 0.54	96.67 ± 0.52
No imp. phase	95.65 ± 0.68	95.40 ± 0.72	95.64 ± 0.68
No imp. modulus	91.85 ± 0.87	91.41 ± 0.95	91.84 ± 0.89

Table II presents the outcomes of the decision tree model. The classification performance is excellent, surpassing the SVM when all features are utilized by a margin of 10.16%, resulting in an accuracy of 96.69%. Furthermore, upon removing the impedance phase, a slight decrease in performance of 1.04% is observed. Finally, excluding the impedance modulus from the training leads to a drop in accuracy by 4.84% compared to the initial case.

C. Random Forest Results

.

TABLE III
PERFORMANCE OF RANDOM FOREST CLASSIFIER

	Acc(%)	F1(%)	AUC(%)
All features	98.48 ± 0.49	97.42 ± 0.52	99.87 ± 0.06
No imp. phase	97.57 ± 0.49	97.42 ± 0.52	99.75 ± 0.08
No imp. modulus	94.78 ± 0.66	94.33 ± 0.69	98.82 ± 0.25

Table III presents the results obtained from the random forest classifier. In this scenario, the performance is enhanced compared to the decision tree case, achieving an accuracy of 98.48%

TABLE IV
PERFORMANCE OF NEURAL NETWORK CLASSIFIER

	Acc(%)	F1(%)	AUC(%)
All features	90.00 ± 1.72	89.46 ± 1.83	96.01 ± 1.26
No imp. phase	87.36 ± 2.42	86.53 ± 2.64	94.99 ± 1.54
No imp. modulus	82.13 ± 1.22	79.98 ± 1.34	89.35 ± 1.37

D. Neural network results

The neural network was the final model analyzed, and the results are presented in Table IV. The achieved classification accuracy is 90.00% when utilizing all the features. Notably, removing the impedance phase led to a slight reduction in accuracy by 2.64% compared to the initial case, whereas excluding the impedance modulus resulted in a more significant drop of 7.87%.

V. Conclusion

In this work, machine learning models are applied to detect water stress in tobacco plants using environmental and impedance measurements. In particular, four different algorithms were considered: support vector machine, decision tree, random forest, and neural network. All the above models performed well, and the best results were obtained using a random forest classifier, achieving 98.48 % of correct predictions. The decision tree achieved a little worse performance but was still excellent with 96.69 %. The neural network reached a classification accuracy of 90 %, while the model that overall performed worse was the support vector machine with 86.53 % accuracy.

Models' performance was also analyzed by alternatively removing the impedance phase and modulus to assess their impact on water stress detection. The results revealed the significance of both quantities in achieving accurate predictions, as all models experienced a decrease in performance when either was removed. Notably, removing the impedance modulus resulted in a more pronounced decline in performance across all considered models, emphasizing its greater relevance as a feature for water stress detection compared to the impedance phase.

Future work aims to broaden the dataset by incorporating new samples from more tobacco plants and extracting additional features intrinsic to the plant. In addition, the approach of stem impedance monitoring together with machine learning models could also be extended to early identification of harmful organisms in the plants, such as fungi and bacteria. Furthermore, emphasizing neural network architectures is desirable, given the accessibility of tools to implement such models on low-cost microcontrollers. This approach opens the potential to develop a low-power, cost-effective, and reliable edge system.

Acknowledgment

This study was carried out within the Agritech National Research Center and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI

RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1032 17/06/2022, CN00000022). This article was funded under the National Recovery and Resilience Plan (NRRP), Mission 4 - Component 2 - Investment 3.1 - Call for tender No. n. 3264 of 28/12/2021 of Italian Ministry of Research funded by the European Union – NextGenerationEU - Project code: IR0000027, Concession Decree No. 128 of 21/06/2022 adopted by the Italian Ministry of Research, CUP: B33C22000710006, Project title: iENTRANCE. This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

REFERENCES

[1] J. Huang, G. Zhang, Y. Zhang, X. Guan, Y. Wei, and R. Guo, "Global desertification vulnerability to climate change and human activities," *Land Degradation & Development*, vol. 31, no. 11, pp. 1380–1391, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ldr.3556

[2] "World population prospects 2022." [Online]. Available: https://population.un.org/wpp/publications/

[3] A. Dinar, A. Tieu, and H. Huynh, "Water scarcity impacts on global food production," *Global Food Security*, vol. 23, pp. 212–226, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2211912417301220

[4] S. I. Hassan, M. M. Alam, U. Illahi, M. A. Al Ghamdi, S. H. Almotiri, and M. M. Su'ud, "A systematic review on monitoring and advanced control strategies in smart agriculture," *IEEE Access*, vol. 9, pp. 32 517–32 548, 2021.

[5] U. Garlando, L. Bar-On, P. M. Ros, A. Sanginario, S. Peradotto, Y. Shacham-Diamand, A. Avni, M. Martina, and D. Demarchi, "Towards optimal green plant irrigation: Watering and body electrical impedance," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2020, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/9181290

[6] A. Sharma, A. Jain, P. Gupta, and V. Chowdary, "Machine learning applications for precision agriculture: A comprehensive review," *IEEE Access*, vol. 9, pp. 4843–4873, 2021.

[7] H. F. Pardede, E. Suryawati, D. Krisnandi, R. S. Yuwana, and V. Zilvan, "Machine learning based plant diseases detection: A review," in *2020 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, 2020, pp. 212–217.

[8] D. Oppenheim and G. Shani, "Potato disease classification using convolution neural networks," *Advances in Animal Biosciences*, vol. 8, no. 2, p. 244–249, 2017.

[9] P. K. Sethy, N. K. Barpanda, A. K. Rath, and S. K. Behera, "Deep feature based rice leaf disease identification using support vector machine," *Computers and Electronics in Agriculture*, vol. 175, p. 105527, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169919326997

[10] M. Kerkech, A. Hafiane, and R. Canals, "Vine disease detection in uav multispectral images using optimized image registration and deep learning segmentation approach," *Computers and Electronics in Agriculture*, vol. 174, p. 105446, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016816991932558X

[11] Y. Peng, Y. Xiao, Z. Fu, Y. Dong, Y. Zheng, H. Yan, and X. Li, "Precision irrigation perspectives on the sustainable water-saving of field crop production in china: Water demand prediction and irrigation scheme optimization," *Journal of Cleaner Production*, vol. 230, pp. 365–377, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959652619314441

[12] M. Bettelli, F. Vurro, R. Pecori, M. Janni, N. Coppedè, A. Zappettini, and D. Tessera, "Classification and forecasting of water stress in tomato plants using bioristor data," *IEEE Access*, vol. 11, pp. 34 795–34 807, 2023. [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3265597

[13] M. Barezzi, F. Cum, U. Garlando, M. Martina, and D. Demarchi, "On the impact of the stem electrical impedance in neural network algorithms for plant monitoring applications," in *2022 IEEE Workshop on Metrology for Agriculture and Forestry (MetroAgriFor)*, 2022, pp. 131–135.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[15] G. Canbek, S. Sagiroglu, T. Taskaya Temizel, and N. Baykal, "Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights," 10 2017, pp. 821–826.

[16] U. Garlando, L. Bar-On, P. M. Ros, A. Sanginario, S. Calvo, M. Martina, A. Avni, Y. Shacham-Diamand, and D. Demarchi, "Analysis of in vivo plant stem impedance variations in relation with external conditions daily cycle," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/9401242

[17] J. Huang and C. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.

[18] L. Raileanu and K. Stoffel, "Theoretical comparison between the gini index and information gain criteria," *Annals of Mathematics and Artificial Intelligence*, vol. 41, pp. 77–93, 05 2004.