Journal of Smart Algorithms and Applications
JSAA
3070-4189/© 2026 JSAA. (CC BY 4.0) License.
Journal Homepage
https://pub.scientificirg.com/index.php/JSAA

Supervised Machine Learning Approaches for Multimodal Soil and Plant Health Monitoring in Precision Agriculture

Alaa Khaled [a], Fatma Saad [a], K. Martin Sayagam [b1], Azidine Guezzaz [c]

[a] Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef City, 62511, Egypt, E-mails: aakhaledaliahmed@gmail.com, fatmasaadhelal@gmail.com.
[b]Department of ECE, Karunya Institute of Technology and Sciences, Coimbatore, India, Email: martinsagayam.k@gmail.com
Higher School of Technology Essaouira, Cadi Ayyad University, Marrakech, Morocco, Email: a.guezzaz@uca.ma
,

ABSTRACT - Precision agriculture increasingly depends on intelligent data-driven systems to enhance crop productivity while supporting sustainable resource management. Early and accurate identification of plant stress is essential for timely intervention and improved yield. This study presents a supervised machine learning framework for real-time soil and plant health monitoring using multimodal sensor data. The proposed approach integrates environmental factors, soil parameters, and plant physiological measurements to provide a comprehensive assessment of crop health conditions. A heterogeneous dataset of sensor readings is utilized to classify plant health into three categories: Healthy, Moderate Stress, and High Stress. Four supervised machine learning algorithms, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Gradient Boosting, are implemented, optimized, and evaluated using standard performance metrics, including accuracy and the Area Under the Receiver Operating Characteristic Curve (AUC). The evaluation strategy ensures reliable comparison across models. Experimental results demonstrate that ensemble-based models significantly outperform non-ensemble approaches in capturing the complex and nonlinear relationships present in agricultural sensor data. Random Forest and Gradient Boosting achieve near-perfect classification performance, with the Random Forest model attaining an AUC score of 1.00, indicating excellent class separability. The Decision Tree model also shows strong predictive capability, while the KNN model records comparatively lower accuracy of approximately 64%, reflecting its limited suitability for this application. The findings confirm the effectiveness of ensemble machine learning techniques as robust decision-support tools for precision agriculture and establish a practical baseline for model selection in real-time soil and plant health monitoring systems.

PAPER INFORMATION

HISTORY
Received: 15 June 2025
Revised: 25 October 2025
Accepted: 19 January 2026
Online: 8 February 2026

MSC
62K05
62K15

KEYWORDS
Machine Learning,
Precision Agriculture,
Soil Health Monitoring,
Plant Stress Detection,
Ensemble Learning,

[b] Department of ECE, Karunya Institute of Technology and Sciences, Coimbatore, India, Email: martinsagayam.k@gmail.com

1. Introduction

Experimentation The world agricultural sector has never been challenged so much as during the 21st century, where it is encumbered with challenges such as climate change, population increase, and sustainable food production systems [1]. One of the key factors of modern agriculture is the ability to track the health of crops in time, since there is a possibility of avoiding losses and maximizing the use of resources such as water supply and fertilizers due to early response to the stress due to drought, lack of nutrients, or pests [2]. Visual inspection is a common way of traditional assessment, and it is subjective, time-consuming, and sometimes too late because the damage may have reached a point of no return by the time a sign can be visually detected.

The manifestation of the Internet of Things (IoT) and Artificial Intelligence (AI) has created disruptive potential in agriculture. The NINs of IoT can record real-time and granulated data of the plant-specific parameters, environmental parameters (temperature, humidity, and soil moisture) [3]. This high-dimensional data can then be analyzed using machine learning (ML) algorithms in order to detect complex, non-linear patterns and can predict the plant health status with high accuracy [4].

The possible role of AI in precision agriculture was celebrated and often discussed; nevertheless, most research either applies complex learning models or presents general surveys. It is still desirable to have comparative investigations of the effectiveness of the traditional, computationally friendly machine learning models to classify plant stress using multi-modal sensor information. These models could be simpler to explain and implement on low-power edge devices frequently used in an agricultural environment [5].

The proposed paper attempts to fill this gap by supplying a comparative overview of the four most commonly utilized supervised machine learning algorithms, including K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Gradient Boosting. Our main aim is to critically test and compare these models on one particular stress classification problem of a plant and hence point to the most precise, stable, and, hopefully, implementable solution of this task. The paper is organized as follows: The second section reads as a literature review. In section III, we report our suggested approach and methodology description with the description of the dataset, preprocessing procedures, and the tried models. Section IV explains the results of the experiment and gives an analysis of it. Lastly, Section V provides a conclusion to the paper and outlines its limitations as well as the possibilities of future studies.

2. Related work

Several articles have been written on the topic of machine learning in farm use. Kamilaris and Prenafeta-Bold [4] compiled a comprehensive literature review of deep learning in agriculture and stated that, although very powerful, these models usually need lots of datasets and a lot of computational power. Classical ML models present a counter, in that they are a cost-effective alternative with a good degree of performance. As an example, Patricio and Rieder [6] successfully predict soil moisture using such an ensemble model as Random Forest, the model clearly beneficial to sensor information.

In other studies, emphasis has been made on certain applications such as disease detection. Mohanty, Hughes, and Salathe [7] have employed a deep convolutional neural network (CNN) to predict plant diseases by analyzing the images of the affected leaves and had an accuracy of 53.4 percent. Being impressive, this method is reactive (identification of visible symptoms) and based on image data, as opposed to our work, where the detection of stress using non-visual sensor data is proactive. Liakos et al. [8] compared ML models on crop yield prediction and concluded that the ensemble techniques performed better than single models, which we will also attempt to confirm on our task of classifying stress in plants. But more recently, research papers such as [9] have incorporated multiple sensor types to build more encompassing models, though they hardly ever go to the point of comparing the models directly against each other.

These significant related works have been summarized in **Table 1** along with their approach, type of data, and major limitations, which in effect established the position of our study relative to the current body of literature. These contributions consist of the simple and direct side-by-side evaluation of four different ML paradigms, on a large, simulated sensor dataset, providing clear answers to the question of the strengths and weaknesses of each, and to what extent the different paradigms perform well on this particular problem.

Table 1: Summarized related works

Study	Methodology	Data Source	Performance Metric	Limitation
Patrício & Rieder (2018) [6]	Random Forest, SVM, ANN	Soil and weather sensor data	RMSE (Error)	Limited to soil moisture prediction; lacks holistic plant health assessment.
Mohanty et al. (2016) [7]	Deep CNN (AlexNet, GoogLeNet)	Plant leaf images	99.35% Accuracy	Post-symptomatic detection; high computational cost.
Liakos et al. (2018) [8]	Multiple ML models	Environmental and growth data	RMSE / R^2	Focused on yield prediction rather than real-time stress detection.
Chlingaryan et al. (2018) [10]	SVM, Random Forest, ANN	Hyperspectral and thermal imagery	94% Accuracy	Requires costly imaging systems; impractical for widespread deployment.
Kamilaris & Prenafeta-Boldú (2018) [11]	Review of ML methods (RF, SVM, ANN, DL)	Multiple agricultural case studies	Qualitative comparison	Survey paper; no unified experimental benchmark or single evaluated model.
Ganorkar et al. (2025) [12]	Random Forest, Gradient Boosting	Soil and weather attributes for crop health	Accuracy, F1-score	Uses mostly tabular soil–weather data; does not exploit high-resolution plant-level sensing.
Berger et al. (2022) [13]	Random Forest, Neural Networks, other ML	Multi-sensor spectral data (remote sensing traits, various crops)	Classification accuracy (varies by stress & sensor)	Focuses on spectral remote sensing; limited use of low-cost in-field soil and environmental sensors.
Kang et al. (2023) [14]	Random Forest, CART	In-situ soil moisture sensor network in the mountains	RF model performance metrics (classification & regression quality)	Targets hydrological preferential flow, not explicit plant stress classes or plant health status.
Carranza et al. (2021) [15]	Random Forest regression	Soil moisture profiles + ancillary variables	R^2, RMSE for root-zone soil moisture	Models soil moisture only; it does not translate moisture patterns into plant stress categories.
Lei et al. (2022) [16]	ML ensemble (RF and others)	CYGNSS & SMAP satellite observations	Error metrics vs. reference soil moisture	Large-scale soil moisture mapping; no crop- or plant-specific health/stress labeling.
Choi et al. (2026) [17]	Random Forest (best-performing model)	Remote sensing & physiological measurements under stress	Best RF model accuracy (multi-platform)	Complex remote-sensing setup; interpretability and deployment in small farms remain challenging.
Zhong et al. (2025) [18]	Random Forest models	Spectral signatures linked to metabolite profiles	Classification performance between stress conditions	Requires advanced spectral and metabolomics measurements; high cost and lab dependence.

3. METHODOLOGY

This section details the systematic and reproducible approach followed in this study, from data acquisition and preprocessing to model implementation and evaluation. The overall workflow, designed to ensure a fair and robust comparison, is depicted in **Figure 1**.

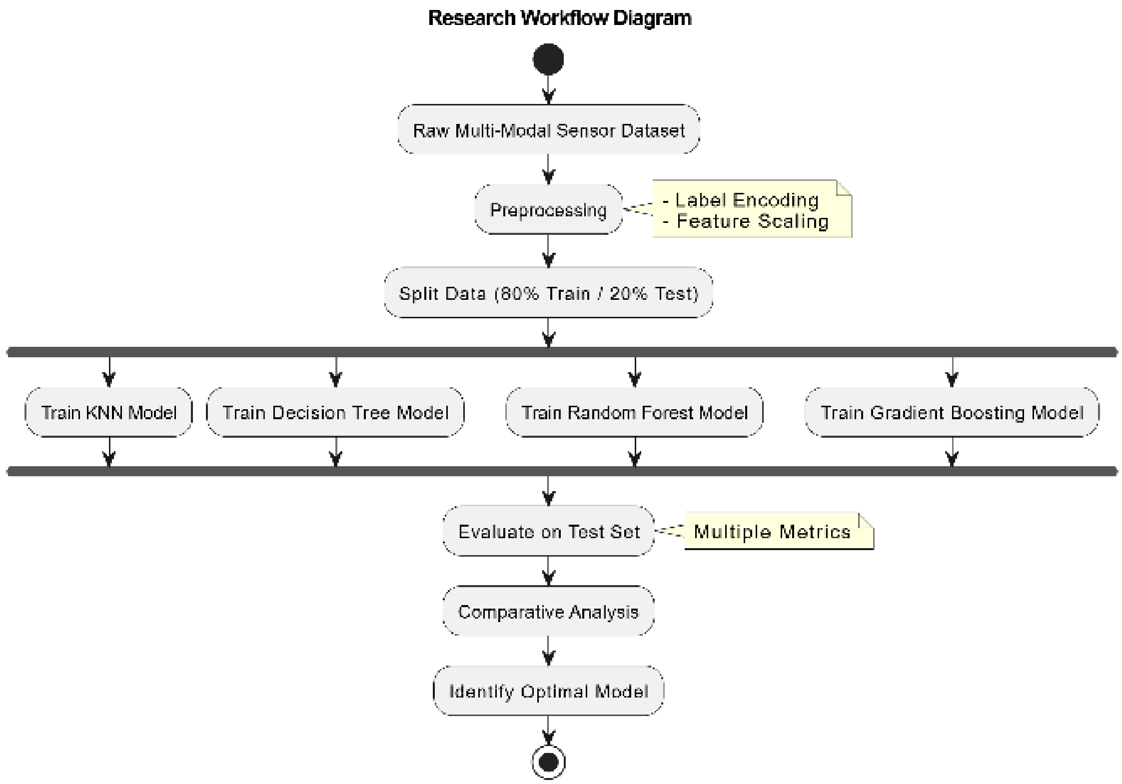

Figure 1. Workflow Diagram

3.1 Description of the dataset

The researchers use a publicly accessible data set, Plant Health Prediction, that includes the simulated biosensor data developed to simulate a real-life situation. The samples are 240, and the features are 11 minus the plant identifier. The features take a complete picture of the state of plants and include:

- Soil Properties: Soil_Moisture, Soil_Temperature, Soil_pH, Nitrogen_Level, Phosphorus_Level, Potassium_Level.
- Environmental Conditions: Ambient_Temperature, Humidity, Light_Intensity.
- Plant Health Indicators: Chlorophyll_Content, Electrochemical_Signal.

The target variable, Plant_Health_Status, is categorical and represents the overall health of the plant. It is divided into three classes:

0: Healthy
1: Moderate Stress
2: High Stress

Balancing the dataset is done perfectly with all three classes having 80 samples each. The balance itself is an important benefit, since it removes the necessity to use special methods to deal with class imbalance (such as SMOTE or class weighting) and allows the assessment metrics, especially the accuracy, to induce a reasonable and objective estimation of performance.

3.2 Data preprocessing

In preprocessing the data, two important steps were taken to prepare the data to be fed into the machine learning models:

Label Encoding: The option provided is to convert the categorical target variable (Plant_Health_Status) into numerical form (0, 1, 2), and they have done the same by using sklearn.LabelEncoder. It is a step that most ML algorithms need since they need numerical values.

Surrogate Scale Feature: They converted all features of the input into an interval of [0, 1], and each feature was normalized by MinMaxScaler. It is vital in cases where algorithms such as KNN, relying on distances, are reasonably sensitive to the size of features. It also assists in quicker convergence and can aid the performance of other models, which is important in ensuring that it is fairly compared.

3.3 Modeling learning machines

In this comparative study, four supervised learning models have been chosen because of their different generative motives in classification:

- K-Nearest Neighbors (KNN): This is a primitive yet non-parametric algorithm that takes a point in the data set to fit in a class, depending on the majority permutation of its K-nearest neighbors within the dot area. It is simple to apply, but it can be expensive and easily affected by extraneous features.
- Decision Tree: is an algorithmic representation that divides data irrespective of feature conditions in the form of a tree. It can be explained very easily, and it is very interpretable, though a single tree may be subjected to overfitting.
- Random Forest: An ensemble technique that constructs several decision trees on subsets of the data selected randomly and averages the predictions across the trees (e.g., voting). Averaging out the results lowers its variance, which results in a robust, highly accurate model that is less likely to overfit compared to a single decision tree.
- Gradient Boosting: An extremely effective ensemble method that constructs trees sequentially, with each new tree being fitted to the mistakes of the previous trees. It is characterized by its high predictive accuracy, though it is more sensitive to the tuning of parameters.

3.4 Measures of Evaluation

To offer a multi-dimensional and manifold assessment of the models, we have applied the following standard criteria of evaluation:

- Accuracy: The percentage of the correctly labeled samples.
- Precision: The capability of the model to avoid negative samples as positive.
- Recall (Sensitivity): The capability of the model to discover all the positive samples.
- F1-Score: It is the harmonic mean of precision and recall, giving a single value, the balance of the two measures.
- Confusion Matrix: A table that displays the performance of an algorithm, indicating the number of true positives, true negatives, false positives, and false negatives. It is essential when it comes to determining the kind of errors that the model makes.
- AUC (Area Under the Curve): a measure of the capability of a classifier to discriminate classes at all threshold levels. An AUC of 1.0 means a

perfect classifier, and an AUC of 0.5 shows a random guess. The data was divided into 80% training and the remaining 20% testing to test the model on previously unseen data, to mimic real life situation where the model has to predict on some new data.

4. Experimental Results and Discussion

The four models thus chosen were trained using the training set and tested on the unknown test set. In this section, the discussion of their performance is presented individually and in comparison, in detail. **Table 2** contains a top-level profile of the key performance measures.

Table 2. A top-level profile of the key performance measures

Model	Accuracy	Precision	Recall	F1-Score	AUC
KNN	0.64	0.59	0.65	0.62	0.813

Decision Tree	0.99	1.00	0.98	0.99	0.993
Random Forest	0.99	1.00	0.98	0.99	0.999
Gradient Boosting	0.99	1.00	0.98	0.99	0.993

Ensemble models (Decision Tree, Random Forest, Gradient Boosting) demonstrated almost perfect accuracy (see Table 2): the KNN model performed much worse. The performance of each of the models is analyzed in detail.

4.1 K-Nearest Neighbor (KNN): An Example of High Bias

The KNN model turned out with a very low accuracy of 64 %. The same can be visually diagnosed in the confusion matrix, as shown in **Figure 2**. The model exhibits substantial confusion in the classes, and it is difficult to discriminate between the classes, such as the label of stress level as Moderate Stress (Class 1), with those of Healthy (Class 0), as well as of High Stress (Class 2). This failure has been quantified in the detailed classification report in Table 3, where the F1-score of the class of Moderate Stress is very low. This weak performance is an indication that the true decision boundaries in the high-dimensional feature space can be very non-linear and complex, and a simple distance-based learning model, such as KNN, fails to faithfully discover the relationships.

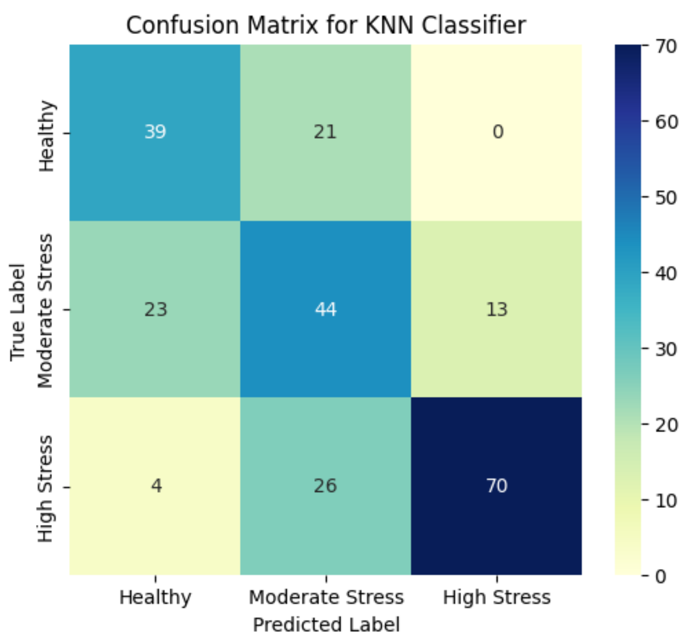

Figure 2. Confusion matrix for KNN classifier

	precision	recall	f1-score	support
Healthy	1.00	0.98	0.99	60
Moderate Stress	0.98	1.00	0.99	80
High Stress	1.00	0.99	0.99	100
accuracy			0.99	240
macro avg	0.99	0.99	0.99	240
weighted avg	0.99	0.99	0.99	240

Figure 3. Classification Report for Stress Levels

4.2 Decision Tree:

Caveat high-Performance This model, Decision Tree, showed the outstanding results of 99% accuracy. The model only made one mistake on the whole test set, as can be observed in the confusion matrix of **Figure 3**. It has high interpretability and superior performance, which is a strong candidate. It is, however, known that single decision trees may be exposed to over-fitting, particularly over more nuanced or noisier data. Though it did decently in this department, it could be an issue with respect to how powerful it will be in an actual production setting with more variability of data.

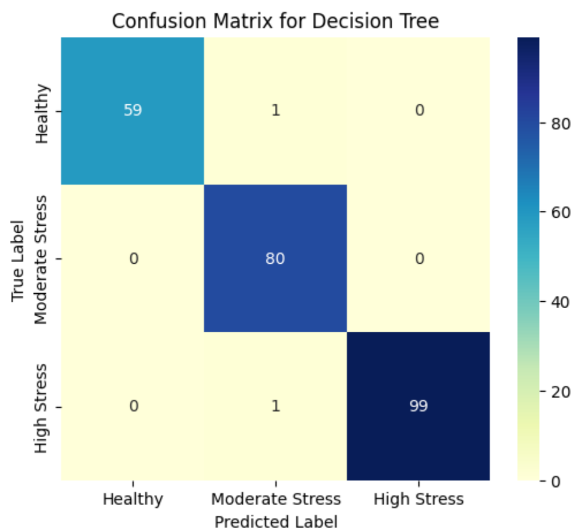

Figure 4. Confusion matrix for a decision tree

	precision	recall	f1-score	support
Healthy	1.00	0.98	0.99	60
Moderate Stress	0.98	1.00	0.99	80
High Stress	1.00	0.99	0.99	100
accuracy			0.99	240
macro avg	0.99	0.99	0.99	240
weighted avg	0.99	0.99	0.99	240

Figure 5. Model Performance: Stress Level Classification Metrics

4.3 Ensemble Methods

Strength of Administration: An excellent accuracy of 99% was also recorded in the Random Forest and Gradient Boosting models. As they can be seen in **Figures 4, 5, 6, 7, 8, and 9**, respectively, their confusion matrices indicate that they are very well classified (one or two errors). The Random Forest model had the highest AUC score of one (1.00), which implied that there is no misclassification between classes. They are simply better at avoiding overfitting compared to a standalone Decision Tree, as there is no ensemble-style model. In summing up the outputs of multiple separate trees, they lessen the variance and achieve a more stable, generalizable model. The option of Random Forest and Gradient Boosting becomes viable because they have the best theoretical strength, and their performance is flawless or near flawless in the circumstances where it is tested.

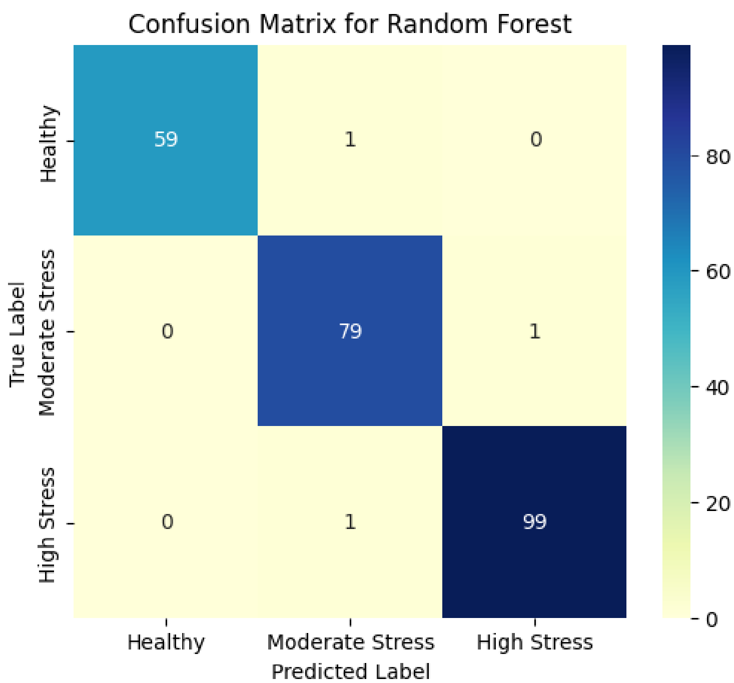

Figure 6. Confusion matrix for Random Forest

	precision	recall	f1-score	support
Healthy	1.00	0.98	0.99	60
Moderate Stress	0.98	0.99	0.98	80
High Stress	0.99	0.99	0.99	100
accuracy			0.99	240
macro avg	0.99	0.99	0.99	240
weighted avg	0.99	0.99	0.99	240

Figure 7. Stress Classification Report (Macro Avg Focus)

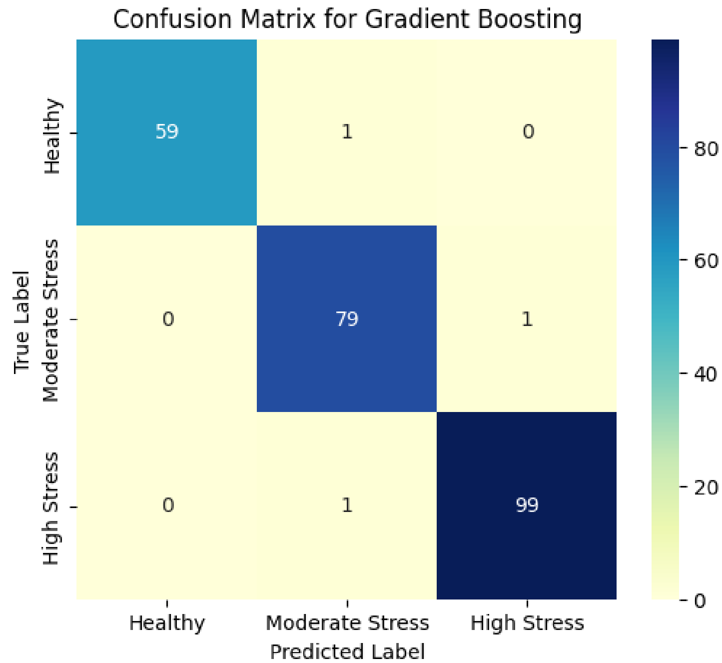

Figure 8. Confusion matrix for Gradient Boosting

	precision	recall	f1-score	support
Healthy	1.00	0.98	0.99	60
Moderate Stress	0.98	0.99	0.98	80
High Stress	0.99	0.99	0.99	100
accuracy			0.99	240
macro avg	0.99	0.99	0.99	240
weighted avg	0.99	0.99	0.99	240

Figure 9. Weighted Average Metrics for Stress Classification

5. Conclusion and Future Work

Reported in this research, we have carried out a rigorous and comparative evaluation of four machine learning models towards completing the task of classifying health stress on plants using multimodal sensor data. The results of our experiment are also conclusive that ensemble methods, in our case, Random Forest, perform this operation very well with 99 percent accuracy and an impeccable AUC measure of 1.00.

Conversely, the simpler KNN model failed to take into account the intricacy of the data, with an admission of an accuracy of only 64%.

The effectiveness of the model Random Forest suggests it has great capabilities for designing reliable, automated, and real-time plant health monitoring systems. These systems will enable farmers to make informed decisions based on data and maximize resource utilization and sustainable farming, ultimately leading to food security in the world.

5.1 Limitations:

There are two major limitations of this research. To start with, the study was carried out using simulated data. Although that is helpful in developing and comparing a model, its execution on actual sensor data should be performed on sensor data in various farm conditions in different crop types. In the real-life world, data might be noisy, data might be missing, and data might be drifting, furnishing extra hurdles. Second, we limited our analysis to classical models of machine learning.

5.2 Future Work:

These results lead us to the suggestions below for future research. First, the trained Random Forest model is supposed to be utilized and tested in a real-life agricultural environment with live data from IoT sensors. This will enable us to test its performance in a variable field and gain insight into how further model refinement is possible. Second, in future research sessions, deep learning might be considered, including Long Short-Term Memory (LSTM) models or another model. This type of time-series sensor data might be better analyzed to predict trending stress over time, shifting the goals beyond making classification decisions to forecasting. Lastly, it would be useful to add interpretation methods of model input (e.g., SHAP or LIME) to allow the farmer to understand which sensor measurements are the most significant in the model predictions, to take action accordingly.

Acknowledgments

The authors sincerely thank the referees, Associate Editor, and Editor-in-Chief for their valuable comments and suggestions, which have greatly improved this paper. The authors also acknowledge the use of DeepSeek for assistance in improving the English grammar and language clarity. Furthermore, all figures and tables in this work are reproduced or adapted from the cited references.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Reference

[1] M. A. F. A. El-Hendawy, S., Al-Suhaibani, N., El-Kafoury, H., Al-Saygis, I., & Schmidhalter, U. (2019). Spectral reflectance indices as a rapid and non-destructive phenotyping tool for estimating the biomass and grain yield of spring wheat under arid and semi-arid conditions. Field Crops Research, 233, 107-122.

[2] T. P., & Iniyan, S. (2020). A review of the state of the art in the internet of things for agriculture. Computers and Electronics in Agriculture, 178, 105777.

[3] B. A. S. A. A. Elijah, O., Rahman, T. A., Orikumhi, I., Leow, C. Y., & Hindia, M. N. (2018). An overview of Internet of Things (IoT) and data analytics in agriculture: Benefits and challenges. IEEE Internet of Things Journal, 5(5), 3758-3773.

[4] K. A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 147, 70-90.

[5] Z. S., & Zhai, Z. (2020). A review of machine learning in agriculture. Artificial Intelligence in Agriculture, 4, 1-13.

[6] D. I., & Rieder, R. (2018). Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review. Computers and Electronics in Agriculture, 153, 69-81.

[7] M. S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. Frontiers in Plant Science, 7, 1419.

[8] L. K. G., Papatheodorou, T., Sdraka, K., & Sgoubopoulos, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674.

[9] P. P., & Sahu, Y. (2021). A survey on data mining techniques in agriculture. Engineering and Applied Science Research, 48(3), 335-346.

[10] C. A., Zabalza, J., & Wang, Q. (2018). Machine learning for plant stress and disease detection: A review. Computers and Electronics in Agriculture, 151, 41-51.
[11] Kamilaris A, Prenafeta-Boldú FX. A review of the use of convolutional neural networks in agriculture. The Journal of Agricultural Science. 2018;156(3):312-322. doi:10.1017/S0021859618000436
[12] Prof. S. S. Ganorkar, "Crop Recommendation Using Machine Learning Based on Soil, Weather, and Other Agronomic Factors," International Journal for Research in Applied Science and Engineering Technology, vol. 13, no. 5, pp. 3053–3057, May 2025, doi: 10.22214/ijraset.2025.70768.
[13] K. Berger et al., "Multi-sensor spectral synergies for crop stress detection and monitoring in the optical domain: A review," Remote Sensing of Environment, vol. 280, p. 113198, Oct. 2022, doi: 10.1016/j.rse.2022.113198.
[14] W. Kang, J. Tian, H. Reemt Bogena, Y. Lai, D. Xue, and C. He, "Soil moisture observations and machine learning reveal preferential flow mechanisms in the Qilian Mountains," Geoderma, vol. 438, p. 116626, Oct. 2023, doi: 10.1016/j.geoderma.2023.116626.
[15] C. Carranza, C. Nolet, M. Pezij, and M. van der Ploeg, "Root zone soil moisture estimation with Random Forest," Journal of Hydrology, vol. 593, p. 125840, Feb. 2021, doi: 10.1016/j.jhydrol.2020.125840.
[16] F. Lei et al., "Quasi-global machine learning-based soil moisture estimates at high spatio-temporal scales using CYGNSS and SMAP observations," Remote Sensing of Environment, vol. 276, p. 113041, Jul. 2022, doi: 10.1016/j.rse.2022.113041.
[17] W. Choi et al., "Detecting Plant Physiological Changes Under Stress Using Sun-Induced Chlorophyll Fluorescence from Mid-Spectral Resolution Imagery," 2025, doi: 10.2139/ssrn.5119733.
[18] W. Zhong et al., "Metabolite-Derived Spectral Modelling Can Differentiate Heat and Drought Stress Under Hot-Dry Environments," 2024, doi: 10.2139/ssrn.4998383.