DOI: https://doi.org/10.54393/fbt.v6i2.243

FUTURISTIC BIOTECHNOLOGY

https://fbtjournal.com/index.php/fbt
ISSN (E): 2959-0981, (P): 2959-0973
Vol 06 Issue 02, (April-June, 2026)

Original Article

OPEN ACCESS

Statistical Characterization of Plant Biosensor Data: Distributional Properties and Inter-Variable Relationships

Maham Riaz[1*]

[1]School of Biochemistry and Biotechnology, University of the Punjab, Lahore, Pakistan

ARTICLE INFO

Keywords:
Precision Agriculture, Plant Health Monitoring, Biosensors, Soil Moisture, Environmental Sensing, Statistical Analysis

How to Cite:
Riaz, M. (2026). Statistical Characterization of Plant Biosensor Data: Distributional Properties and Inter-Variable Relationships: Plant Biosensor Data: Distributional Properties and Inter-Variable Relationships. Futuristic Biotechnology, 6(2), 33-38. https://doi.org/10.54393/fbt.v6i2.243

***Corresponding Author:**
Maham Riaz
School of Biochemistry and Biotechnology, University of the Punjab, Lahore, Pakistan
mahamriaz533@gmail.com

Received Date: 7th May, 2026
1st Revision Received: 16th May, 2026
2nd Revision Received: 2nd June, 2026
Acceptance Date: 22nd June, 2026
Published Date: 30th June, 2026

ABSTRACT

Precision agriculture is now increasingly being monitored with continuous biosensors. But many datasets are analyzed for predictive modeling without any basic statistical properties of the data being characterized, which is required to be able to determine whether it is appropriate for downstream use. **Objectives:** To characterize the distributional properties of the key environmental and soil biosensor variables, to assess the linear relations among variables, and to determine if the biosensor readings vary systematically across individual plants of a biosensor-monitored dataset. **Methods:** A de-identified dataset of plant biosensor readings was used, consisting of 120 readings per plant from 10 plants spanning an approximate 6-hour time interval. The soil moisture, ambient temperature, soil temperature, humidity, light intensity, soil pH, nitrogen level, and phosphorus level were analyzed. Descriptive statistics were conducted, normality was tested, and ANOVA was used. **Results:** All variables showed very peaked and leptokurtic distribution shapes (skewness and excess kurtosis values close to zero). The patterns here are not a formal goodness-of-fit test against a uniform distribution, so they are not definitive statements of uniformity, but rather hypotheses that should be considered. **Conclusions:** The distributional patterns of the biosensor variables were very symmetrical and flattened (indicating uniformity) and had very low inter-variable correlation. The statistical characteristics in these readings indicate that single-timepoint readings alone might not be suitable for predicting an outcome in a meaningful way without extra labeled outcome data and engineered temporal properties. The results provide a baseline of statistics to be used in subsequent research with this data.

INTRODUCTION

One of the most important trends in the emerging field of precision agriculture is the application of biosensors to measure environmental and soil parameters in real time, providing advantageous opportunities for early detection of plant stress and for timely management decisions on irrigation, fertilization, and crop management [1]. Biosensor technologies offer the potential for continuous or near real-time measurement of relevant biological and environmental parameters in agricultural contexts, such as in the field, which is not possible with traditional monitoring methods using periodic sampling and laboratory analysis. In-field and wearable biosensors can monitor physiological and environmental changes in plants in real time, such as changes that can take place before the observable signs of stress emerge [2]. This early detection capacity can enable farmers to adjust to varying crop conditions before significant crop damage or loss of yield. The sensitivity, selectivity, and response time of the biosensor platforms have been further enhanced by recent developments in the field of nanotechnology. Several important soil parameters, such as soil moisture and N and P levels, which can be difficult, time-consuming, or expensive to measure again and again with traditional laboratory methods, have been monitored in a rapid and highly sensitive way by means of new nanobiosensor systems [3, 4]. These technologies can deliver greater detail of the spatial and

DOI: https://doi.org/10.54393/fbt.v6i2.243

temporal variation of soil conditions, which will help to develop site-specific management strategies and more efficient use of agricultural resources. Internet of Things (IoT) technologies and capacitive-based soil-moisture sensors are especially well suited for long-term field monitoring, as they can continuously monitor and share information related to soil water availability [5]. Moreover, developments in the integration of biosensors, internet of things (IoT) systems, and data analysis systems have opened new avenues for automated agricultural monitoring. Data collected from the sensors can be used to detect changes in the environmental conditions, assess plant reactions, and inform management decisions that are based on data. As for plant-health monitoring, however, the predictive modeling approach relies heavily on the statistical features and quality of data generated by the sensors [6–8]. The distribution of sensor measurements, relationships between different variables, the temporal fluctuations, the variation in measurements, and differences between the individual plants or field locations are all important. Therefore, it is crucial that these characteristics are analyzed properly and that reliable predictive models are developed. The integration of real-time biosensing with sophisticated statistical and predictive techniques could potentially lead to more precise plant-stress detection, more effective resource use, and more sustainable and efficient precision agriculture.

Although machine learning-based methods are effective for plant stress detection, relying on plant data, there is limited research that systematically characterizes the descriptive statistical properties of raw biosensor data and tests the difference between measurements from individual plants under similar environmental and soil conditions. These basic statistical characteristics of data are crucial before it can be used effectively for predictive modelling, but are rarely done and often left out of the equation when developing the model. In addition, the data set employed in this study has not been previously described regarding its distribution properties, correlation structure, or between-plant variability, which creates a gap in the knowledge for researchers who wish to use this data set for modeling applications. The objective of this study was to fill this gap with a detailed statistical characterization of a dataset of plant biosensors. This study aimed to characterize the central tendency, dispersion, and shape of the distribution of eight environmental and soil parameters, to assess the strength and direction of linear relationships among all parameter pairs, to test for systematic differences in biosensor measurements within individual plants, and to summarize plant-level mean profiles based on which potential systematic trends may be recognized across parameters.

METHODS

This was a cross-sectional retrospective study based on a de-identified plant biosensor dataset obtained from Kaggle [9]. The dataset ("Plant-Health-Data," Version 1.0), curated by user "ziya07," was accessed on January 2026, and is available at https://www.kaggle.com/datasets/ziya07/plant-health-data. It comprises simulated environmental and soil sensor readings generated between October and November 2024, with no accompanying metadata on plant species, growth stage, sensor calibration, or health status [9]. The analytic sample consisted of 10 individually identified plants (Plant_ID 1-10), with 120 readings per plant taken at approximately 6-hour intervals over 30 days, totaling 1,200 readings. Eight biosensor variables were analyzed: soil moisture (%), ambient temperature (°C), soil temperature (°C), humidity (%), light intensity (lux), soil pH, nitrogen level (mg/kg), and phosphorus level (mg/kg) [10, 11]. Records were included if they had: (1) a valid timestamp, (2) a valid Plant_ID (1-10), and (3) complete numeric data for all eight variables. Exclusion criteria were: (1) missing data, (2) duplicate Plant_ID-Timestamp combinations, and (3) physically implausible values (e.g., soil moisture or humidity below 0, or pH outside the range 0-14). Distributional properties were assessed using multiple approaches. The Shapiro-Wilk test was used to evaluate deviation from normality [10]. Skewness and excess kurtosis were calculated to characterize the shape of observed distributions [11]. Formal goodness-of-fit testing (e.g., Kolmogorov-Smirnov test against a uniform distribution) was not conducted; therefore, interpretations of distributional shape are descriptive.

Pearson correlation coefficients were calculated to quantify the magnitude of linear relationships between variables [12]. Given the evidence of non-normality (Shapiro-Wilk $p<0.050$), interpretations were based primarily on the correlation coefficient (r) rather than p-values, as the coefficient estimate itself is robust to distributional shape. One-way ANOVA was used for between-plant comparisons because (i) the F-statistic is robust to moderate non-normality given the balanced design and large sample size (n=120 per group), and (ii) the effect size (η^2) provides a practically meaningful measure of variance explained by plant identity [13]. To confirm these findings under the observed non-normal distributions, the non-parametric Kruskal-Wallis test was conducted as a sensitivity analysis [14]. To address multiple comparisons, Bonferroni correction was applied (α adjusted = 0.05/8 = 0.00625). All statistical analyses were performed using R version 4.2.0 (R Core Team, 2022).

DOI: https://doi.org/10.54393/fbt.v6i2.243

RESULTS

The central tendency and dispersion of the eight biosensor variables were computed to characterize their typical values and variability. Skewness values were near zero for all variables (-0.02 to 0.03), indicating high symmetry. Excess kurtosis values consistently clustered around -1.2 (range: -1.18 to -1.24), which approximates the theoretical excess kurtosis of a perfectly uniform distribution (-1.2). The Shapiro-Wilk test confirmed significant deviation from normality for all variables ($p<0.001$), consistent with the descriptive evidence of non-normal, flat distributions. While these descriptive statistics suggest near-uniform distributions, formal goodness-of-fit testing (e.g., Kolmogorov-Smirnov test against a uniform distribution) would be necessary to confirm uniformity definitively. This characterization is important because the observed distributions differ markedly from the normal distributions often assumed in many biological studies (Table 1).

Table 1: Descriptive Statistics for All Biosensor Variables (n=1,200)

Variables	Mean ± SD
Soil Moisture (%)	25.11 ± 8.68
Ambient Temperature (°C)	24.00 ± 3.44
Soil Temperature (°C)	19.96 ± 2.93
Humidity (%)	54.85 ± 8.78
Light Intensity (lux)	612.64 ± 228.32
Soil pH	6.52 ± 0.58
Nitrogen Level (mg/kg)	30.11 ± 11.51
Phosphorus Level (mg/kg)	30.26 ± 11.47

Pairwise Pearson correlation coefficients were calculated to assess linear relationships between the eight variables (Table 2).

Table 2: Pearson Correlation Matrix Across All Eight Biosensor Variables

Variables	Soil Moisture	Ambient Temperature	Soil Temperature	Humidity	Light Intensity	Soil pH	Nitrogen Level	Phosphorus Level
Soil Moisture	1	-0.009	-0.027	0.054	-0.009	-0.044	-0.025	0.019
Ambient Temperature	-0.009	1	0.037	0.004	-0.025	-0.021	-0.012	-0.001
Soil Temperature	-0.027	0.037	1	0.023	0.004	-0.025	-0.033	-0.04
Humidity	0.054	0.004	0.023	1	-0.024	0.014	-0.023	-0.003
Light Intensity	-0.009	-0.025	0.004	-0.024	1	0.008	-0.013	-0.041
Soil pH	-0.044	-0.021	-0.025	0.014	0.008	1	-0.006	0.023
Nitrogen Level	-0.025	-0.012	-0.033	-0.023	-0.013	-0.006	1	0.013
Phosphorus Level	0.019	-0.001	-0.04	-0.003	-0.041	0.023	0.013	1

All coefficients fall between -0.044 and 0.054 in magnitude, indicating negligible linear association between any pair of variables. Corresponding p-values for all correlation coefficients were >0.050, confirming that none reached statistical significance (Figure 1).

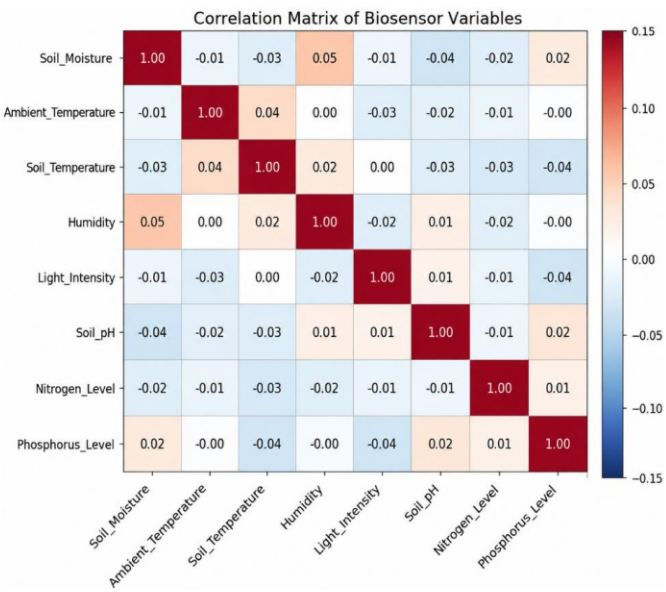

Figure 1: Correlation Matrix of Biosensor Variables

All the skewness values are close to zero, which means that they are very symmetrical. Excess kurtosis values are always around the theoretical value of a perfectly uniform distribution of –1.2, with the larger, but still consistent, values being around –1.2. This descriptive evidence provides support for the visual evidence of flat and symmetric distributions, but there was no formal goodness-of-fit testing to verify a strict uniform distribution (Table 3).

Table 3: Skewness and Kurtosis Values for All Biosensor Variables (n=1,200)

Variables	Skewness	Kurtosis
Soil Moisture	-0.02	-1.23
Ambient Temperature	0.01	-1.19
Soil Temperature	-0.01	-1.22
Humidity	0.03	-1.18
Light Intensity	0.00	-1.21
Soil pH	-0.02	-1.20
Nitrogen Level	0.01	-1.24
Phosphorus Level	-0.01	-1.22

One-way ANOVA was used to test whether each variable's mean differs significantly across the 10 plants. To account for multiple comparisons, Bonferroni correction was applied (adjusted $\alpha = 0.05/8 = 0.00625$). Seven of eight

DOI: https://doi.org/10.54393/fbt.v6i2.243

variables showed no statistically significant between-plant differences. Nitrogen Level showed a nominally significant difference (F=2.175, p=0.021) (Table 4).

Table 4: One-way ANOVA Results by Variable, Grouped by Plant ID

Variables	F-statistic	p-value
Soil Moisture	0.758	0.655
Ambient Temperature	1.061	0.388
Soil Temperature	1.317	0.222
Humidity	0.419	0.925
Light Intensity	0.86	0.561
Soil pH	0.718	0.692
Nitrogen Level	2.175	0.021
Phosphorus Level	0.65	0.754

To assess the individual plant profiles directly, the mean values of the eight parameters per plant were calculated. The levels of the plants are generally similar across all variables, with Plant 3 (33.03%) and Plant 9 (28.05%) representing the extremes at the observed high and low ends of Nitrogen Level. There was no evidence of any plant identity across the whole set of variables being either systematically raised or lowered, and the effect size was small (η^2 = 0.02), meaning that only 2% of the variance in Nitrogen Level is attributable to plant identity (Table 5).

Table 5: Mean Value of Each Biosensor Variable by Plant ID (n=120 Readings Per Plant)

Plant ID	Soil Moisture (%)	Ambient Temp. (°C)	Soil Temp. (°C)	Humidity (%)	Light Intensity (lux)	Soil pH	Nitrogen Level	Phosphorus Level
1	24.76	23.83	19.96	55.33	581.18	6.49	29.76	29.60
2	25.85	24.58	20.16	54.06	601.67	6.51	28.46	30.29
3	26.08	24.34	19.80	55.20	623.03	6.45	33.03	31.31
4	25.12	23.92	19.69	55.00	622.96	6.52	31.40	29.72
5	24.54	24.16	20.23	55.32	641.17	6.56	29.72	31.59
6	25.40	23.83	19.26	55.08	608.93	6.51	30.43	30.67
7	24.91	23.44	20.01	53.81	641.72	6.52	30.28	28.96
8	25.82	23.78	20.02	54.75	600.93	6.50	28.58	30.90
9	24.72	23.94	20.25	54.89	599.26	6.61	28.05	30.09
10	23.86	24.18	20.20	55.10	605.52	6.57	31.36	29.52

DISCUSSION

The data set exhibited highly flattened and symmetric distributional patterns descriptively. However, without formal goodness-of-fit testing, these findings should be interpreted cautiously as indicative of non-normality rather than conclusive proof of a true uniform distribution, with negligible inter-variable correlations and limited between-plant variation, and were the subject of statistical analysis. This information has implications for the use of this data for predictive modelling of plant health. The actual distribution shape was close to uniformity, and there was no correlation between physically related variables, such as ambient temperature and soil temperature (r=0.037), which differs from what is expected of physically or biologically coupled systems, where covariation is usually correlated with a common environmental factor [15]. It has been found in previous studies that there is a moderate to strong correlation between temperature and humidity under field conditions [16] and a strong correlation between soil moisture and humidity and temperature [17]. The low correlation values found here indicate that these data might have been collected in a simulated or highly controlled environment instead of in the natural field, or that the monitoring configuration sampled different microenvironments [18]. This result is consistent with the work done by Zhang *et al.* who observed that the correlation structures in environmental monitoring data are highly sensitive to both the placement and calibration of the sensors [5]. While no other differences were found, the small but nominally significant difference in Nitrogen Level between plants suggests that there may be differences in the distribution of nutrients in the soils. This is in line with precision agriculture practice literature that has identified soil nutrient variability, especially for nitrogen and phosphorus, as a critical management variable that should be managed in separate areas and using separate plans and practices [10, 19]. The near-zero correlation between nitrogen and other variables, such as soil pH (r = -0.006) and phosphorus (r=0.013), is surprising because the availability of nutrients is expected to be influenced by soil pH [11] and light intensity, to affect photosynthesis aside from nutrient status [20]. This independence could be due to independent sensor sampling, or not necessarily a scenario that is representative of coupled field conditions [21]. Machine learning techniques for plant stress detection have been shown to benefit from the use of labeled stress/health data, multi-modal sensing, or engineered temporal features to achieve good predictive performance [7, 18]. Near-uniform distributions and weak inter-variable correlations observed here indicate that simple linear models based on raw single-timepoint readings are unlikely to provide good predictive signals. This is consistent with Elvanidi and Katsoulas and Garg *et*

DOI: https://doi.org/10.54393/fbt.v6i2.243

al. who demonstrated that temporal dynamics (e.g., rate of change, deviation from plant-specific baselines) and multimodal sensor fusion were a better way to express plant stress and health status than just static readings of the environment [7, 18]. There are three ways to extend this dataset towards plant-health predictive applications. First, in supervised modelling [7], an outcome variable that is labelled is required. Second, the ability to see structure that isn't evident in a single-timepoint reading can be gained with time-based feature engineering, including rolling averages, rate-of-change calculations, or deviations from plant-specific baselines [7, 18]. Thirdly, due to the non-normal distributions, future hypothesis testing should be conducted using non-parametric statistical methods [21].
There are several restrictions that should be noted. Firstly, the study team was not involved in the pre-collection of the data, so they had no control over the calibration, placement, and data collection process of the sensors, and were unable to validate the accuracy of the measurements. Secondly, there is no metadata available about plant species, growth stage or cultivation environment to limit conclusions for plant-wise comparisons. Third, because no health or stress outcome variable was used, this analysis only describes characteristics of input variables, not their predictive ability. Fourth, the cross-sectional structure does not take full advantage of the repeated measures, time-series structure of the data; the use of within-plant autocorrelation was not addressed in this study. Fifth, the uniform distribution and near-zero correlation structure seem to indicate that the variables measured may not fully correspond to physical processes, but could have been sampled independently, depending on the sensor documentation, and thus may need to be verified before applied modeling. The current data set should be expanded with additional data with an outcome variable that is labeled for plant health or stress to allow supervised predictive modelling. Researchers are advised to develop temporal features (e.g., rate of change, deviation from plant-specific baseline), and apply non-parametric or longitudinal mixed-effects models to explore the dynamic stress responses due to the low cross-sectional correlations and non-normal distributions. Furthermore, future deployments should incorporate detailed calibration, placement, and contextual information to address the near-zero correlation issue between the physically coupled variables, and evaluate if configuration sampling requirements need to be adjusted to reflect field conditions.

CONCLUSION

This cross-sectional statistical analysis of a plant biosensor monitoring dataset revealed that eight environmental and soil variables were complete and exhibited near-uniform distributions with negligible inter-variable correlations. Only Nitrogen Level showed a nominally significant between-plant difference (p=0.021), but with a small effect size (η^2=0.02), and no individual plant displayed systematic elevation or depression across multiple variables simultaneously. The results provide the statistical baseline that is necessary for this data set, and suggest that in their current state, raw biosensor readings from a single timepoint do not have the linear structure required for robust predictive plant health modeling. Improvements in the future should include the acquisition of labeled outcome variables, use of non-parametric statistical methods for the relevant distributions, and the creation of engineered variables for time to allow for the capture of temporal dynamics that are not evident in the cross-sectional readings alone. These measures are important in order to turn this statistical basis into a good basis for precision agriculture decisions.

Author's Contribution

Conceptualization: MR
Methodology: MR
Formal analysis: MR
Writing and Drafting: MR
Review and Editing: MR

The author approved the final manuscript and take responsibility for the integrity of the work

Conflicts of Interest

The author declare no conflict of interest.

Source of Funding

The author received no financial support for the research, authorship and/or publication of this article.

REFERENCES

[1] Dhanaraju M, Chenniappan P, Ramalingam K, Pazhanivelan S, Kaliaperumal R. Smart Farming: Internet of Things (IoT)-Based Sustainable Agriculture. Agriculture. 2022 Oct; 12(10): 1745. doi: 10.3390/agriculture12101745.
[2] Li XH, Li MZ, Li JY, Gao YY, Liu CR, Hao GF. Wearable Sensor Supports in-Situ and Continuous Monitoring of Plant Health in the Precision Agriculture Era. Plant Biotechnology Journal. 2024 Jun; 22(6): 1516-35. doi: 10.1111/pbi.14283.
[3] Jayabharathi SP and Theresa K. Nanobiosensors: Enabling Next-Generation Soil and Plant Health Monitoring. Frontiers in Nanotechnology. 2026 Aug;

DOI: https://doi.org/10.54393/fbt.v6i2.243

8: 1832447. doi: 10.3389/fnano.2026.1832447.
[4] Chaturvedi A, Tripathi D, Ranjan R. Nano-Enabled Biosensors in Early Detection of Plant Diseases. Frontiers in Nanotechnology. 2025 Apr; 7: 1545792. doi: 10.3389/fnano.2025.1545792.
[5] Zhang X, Feng G, Sun X. Advanced Technologies of Soil Moisture Monitoring in Precision Agriculture: A Review. Journal of Agriculture and Food Research. 2024 Dec; 18: 101473. doi: 10.1016/j.jafr.2024.101473.
[6] Narware J, Chakma J, Singh SP, Prasad DR, Meher J, Singh P *et al.* Nanomaterial-Based Biosensors: A New Frontier in Plant Pathogen Detection and Plant Disease Management. Frontiers in Bioengineering and Biotechnology. 2025 Apr; 13: 1570318. doi: 10.3389/fbioe.2025.1570318.
[7] Elvanidi A and Katsoulas N. Machine Learning-Based Crop Stress Detection in Greenhouses. Plants. 2022 Dec; 12(1): 52. doi: 10.3390/plants12010052.
[8] Walsh JJ, Mangina E, Negrão S. Advancements in Imaging Sensors and AI for Plant Stress Detection: A Systematic Literature Review. Plant Phenomics. 2024 Mar; 6: 0153. doi: 10.34133/plantphenomics.0153.
[9] Dziadkowiec O, Durbin J, Muralidharan VJ, Novak M, Cornett B. Improving the Quality and Design of Retrospective Clinical Outcome Studies That Utilize Electronic Health Records. Hospital Corporation of America: Healthcare Journal of Medicine. 2020 Jun; 1(3): 131. doi: 10.36518/2689-0216.1094.
[10] Yuan Y, Shi B, Yost R, Liu X, Tian Y, Zhu Y *et al.* Optimization of Management Zone Delineation for Precision Crop Management in an Intensive Farming System. Plants. 2022 Oct; 11(19): 2611. doi: 10.3390/plants11192611.
[11] Barrow NJ. The Effects of pH on Phosphate Uptake from the Soil. Plant and Soil. 2017 Jan; 410(1): 401-10. doi: 10.1007/s11104-016-3008-9.
[12] Shapiro SS and Wilk MB. An Analysis of Variance Test for Normality (Complete Samples). Biometrika. 1965 Dec; 52(3-4): 591-611. doi: 10.1093/biomet/52.3-4.591. Pearson K. Note on Regression and Inheritance in the Case of Two Parents. Proceedings of the Royal Society of London. 1895 Jan; 58: 240-2. doi: 10.1098/rspl.1895.0041.
[13] Fisher RA. Statistical Methods for Research Workers. 1934.
[14] Joanes DN and Gill CA. Comparing Measures of Sample Skewness and Kurtosis. Journal of the Royal Statistical Society: Series D (The Statistician). 1998 Apr; 47(1): 183-9. doi: 10.1111/1467-9884.00122.
[15] Dong Y, Werling B, Cao Z, Li G. Implementation of an in-field IoT system for Precision Irrigation Management. Frontiers in Water. 2024 Feb; 6: 1353597. doi: 10.3389/frwa.2024.1353597.
[16] Borah S, Kumar R, Mukherjee S. Design and Characterization of a Low-Cost Capacitive Soil Moisture Sensor System for IoT-based Agriculture Applications. 2023.
[17] Garg D, Singh H, Shacham-Diamand Y. AdapTree: Data-Driven Approach to Assessing Plant Stress Through The AI-Sensor Synergy. Sensors. 2025 May; 25(10): 3149. doi: 10.3390/s25103149.
[18] Althubaiti A. Sample Size Determination: A Practical Guide for Health Researchers. Journal of General and Family Medicine. 2023 Mar; 24(2): 72-8. doi: 10.1002/jgf2.600.
[19] Tang W, Guo H, Baskin CC, Xiong W, Yang C, Li Z *et al.* Effect of Light Intensity on Morphology, Photosynthesis and Carbon Metabolism of Alfalfa (Medicago sativa) Seedlings. Plants. 2022 Jun; 11(13): 1688. doi: 10.3390/plants11131688.
[20] Kruskal WH and Wallis WA. Use of Ranks in One-Criterion Variance Analysis. Journal of the American Statistical Association. 1952 Dec; 47(260): 583-621. doi: 10.1080/01621459.1952.10483441.