

# A Deep Feature–Enhanced Multi-Sensor Model for Predictive Crop Development in Smart Greenhouses

Pulime Satyanarayana<sup>1\*</sup>, Sankati Aakanksha Reddy<sup>2</sup>, M. Navyasri<sup>2</sup>, T. Ganesh<sup>2</sup>

<sup>1</sup>Associate Professor, <sup>2</sup>UG Student, <sup>1,2</sup>Department of Computer Science and Engineering (AI&ML),

<sup>1,2</sup>Kommuri Pratap Reddy Institute of Technology, Ghanpur, Ghatkesar, 501301, Telangana, India.

\*Correspondence: Pulime Satyanarayana (snpulime@gmail.com)

## ABSTRACT

The analysis of plant growth and development in controlled agricultural environments has gained importance due to the increasing demand for sustainable and efficient food production. Traditional plant monitoring methods relied on manual observation and basic statistical techniques, which offered limited understanding of complex environmental interactions. With advancements in sensor technologies and smart farming systems, large volumes of environmental and plant-related data are continuously generated. However, conventional analytical approaches struggle to process high-dimensional datasets and often fail to accurately predict plant development patterns. To address these challenges, this study proposes a machine learning (ML)-based analytical framework that integrates data preprocessing, exploratory data analysis (EDA), classification models, regression models, and hybrid deep learning techniques for plant development analysis. Classification algorithms such as Support Vector Classifier (SVC), Bernoulli Naive Bayes Classifier (BNC), and Multinomial Naive Bayes Classifier (MNC) are employed to identify plant development stages. Regression models including Decision Tree Regressor (DTR), Support Vector Regressor (SVR), and Ridge Regressor (RR) are utilized to predict plant growth parameters. Furthermore, two hybrid models are introduced to enhance predictive performance: Feed Forward Neural Network (FFNN) combined with Gaussian Naive Bayes (GNB), termed Deep Feature Probabilistic Classifier (DFPC), and FFNN integrated with RR, referred to as Hybrid Deep Ridge Predictor (HDRP). Experimental results demonstrate that DFPC achieves a classification accuracy of 94.42%, while HDRP attains a highly accurate prediction performance with an  $R^2$  score of 0.999. These findings highlight the effectiveness of combining deep feature extraction with machine learning techniques for precise plant development analysis and improved decision-making in controlled agricultural systems.

**Keywords:** Controlled environment agriculture, plant growth analysis, plant development monitoring, sensor data, data preprocessing, exploratory data analysis, classification, regression

## 1. INTRODUCTION

Crop yield hinges on the growth and differentiation of crops, as well as the distribution and accumulation of photosynthetic products, constituting a pivotal area of study in crop science. The precise and early prediction of crop yield consistently stands as one of the most important areas of investigation in precision agriculture [1]. Oilseed rape (*Brassica napus* L.) is a significant global oil crop that can be used to produce high-quality feed, cooking oil, dishes, biodiesel, and so on [2,3]. The timely and accurate yield prediction of oilseed rape is of guiding significance for agricultural producers to carry out cultivation, breeding, production, and sales activities reasonably and effectively. As shown in fig 1 conventionally the prevalent approaches for estimating crop yield involved establishing correlations between agronomic or climatic factors and crop yields through manual surveys or employing statistical analysis methods [4,5], which were not uniform, time-consuming and laborious.

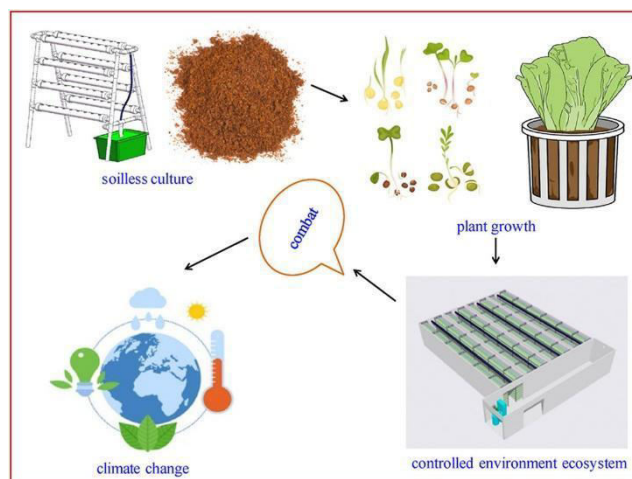


Fig. 1: Controlled environment ecosystem.

At present, the large-scale commercial breeding of oilseed rape has led to a sharp increase in the combination of breeding materials. It was difficult to ensure the uniformity and timeliness of data for estimating yield. Therefore, there is an urgent need to devise a swift, dependable, and non-invasive method capable of accurately forecasting oilseed rape yields on a field scale. Remote sensing is a real-time means of obtaining phenotypic information for crops, with a large monitoring range and low damage to crops, so it was widely used in agricultural information acquisition [4]. The demand for the timely and non-destructive evaluation of crop yield across diverse scales has resulted in the growing acceptance of remote sensing technologies [5]. Utilizing unmanned aerial vehicles (UAVs) for remote sensing has overcome the limitations associated with satellites, aircraft, and ground-based platforms. UAVs enable frequent flight tests at specific locations and times, facilitating the monitoring of crops through multi-scale and multi-temporal images. This capability allows for the observation of fine-scale spatial patterns in a more efficient manner.

## 2. LITERATURE SURVEY

Han S et al. [6] developed in the application of deep learning methods including CNNs, RNNs, LSTMs, Transformers, GANs, and VAEs to tasks such as growth monitoring, yield prediction, pest and disease identification, and phenotypic analysis. It further examines prominent research themes, including multimodal data fusion, transfer learning, and model interpretability. Additionally, it discusses key challenges related to data scarcity, model generalization, and real-world deployment. Finally, the review outlines prospective directions for future research, aiming to inform the integration of deep learning with phenomics and intelligent IoT systems and to advance plant monitoring toward greater intelligence and high-throughput capabilities.

Li B et al. [7] investigated the combined influence of these factors on the photosynthetic rate of chili pepper plants; a predictive model was developed for their net photosynthetic rate (Pn) during the peak fruiting period. A multifactorial nested experimental design with irregular sampling intervals was used to systematically examine the interacting effects of light intensity, CO<sub>2</sub> concentration, and temperature on photosynthesis. Based on the collected data, a support vector regression (SVR) algorithm was trained and its performance was compared with that of a backpropagation (BP) neural network, a radial basis function (RBF) neural network, and a random forest (RF) algorithm. To optimize performance, a grid search with five-fold cross-validation was conducted to identify optimal hyperparameters.

Roy AM et al. [8] proposed a model that has been improved to optimize for both detection speed and accuracy and applied to multi-class apple plant disease detection in the real environment. The mean average precision (mAP) and F1-score of the detection model reached up to 91.2% and 95.9%,

respectively, at a detection rate of 56.9 FPS. The overall detection result demonstrates that the current algorithm significantly outperforms the state-of-the-art detection model with a 9.05% increase in precision and 7.6% increase in F1-score. The proposed model can be employed as an effective and efficient method to detect different apple plant diseases under complex orchard scenarios. Manuelito C et al. [9] evaluated the effectiveness of two sensor-based approaches—proximal sensing with a FLAME spectrometer and remote sensing via UAV-mounted multispectral imaging compared with foliar chemical analyses as the reference standard, for diagnosing the nutritional status of olive trees. They were conducted in Elvas, Portugal, between 2022 and 2023, across three olive cultivars ('Azeiteira', 'Arbequina', and 'Koroneiki') subjected to different fertilisation regimes. Machine learning (ML) models showed strong correlations between sensor data and nutrient levels: the multispectral sensor

Xu R et al. [10] designed the multi-sensor unmanned aerial system was open sourced. A data processing pipeline was proposed to preprocess the raw data and to extract phenotypic traits at the plot level, including morphological traits (canopy height, canopy cover, and canopy volume), canopy vegetation index, and canopy temperature. Protocols for both field and laboratory calibrations were developed for the RGB, multispectral, and thermal cameras. The system was validated using ground data collected in a cotton field. Temperatures derived from thermal images had a mean absolute error of 1.02 °C, and canopy NDVI had a mean relative error of 6.6% compared to ground measurements. Lee D-H et al. [11] introduced an artificial intelligence (AI)-powered model that utilizes unmanned aerial systems (UAS)-based multi-sensor data to predict Napa cabbage fresh weight. The model was developed using high-resolution RGB, multispectral (MSP), and thermal infrared (TIR) imagery collected throughout the 2020 growing season. The imagery was used to extract various vegetation indices, crop features (vegetation fraction, crop height model), and a water stress indicator (CWSI). The deep neural network (DNN) model consistently outperformed support vector machine (SVM) and random forest (RF) models.

Zhu H et al. [12] explored the feasibility and potential for predicting oilseed rape yield through the utilization of a platform equipped with RGB and multispectral cameras. Genetic algorithm–partial least square was employed and evaluated for effective wavelength (EW) or vegetation index (VI) selection. Additionally, different machine learning algorithms, i.e., multiple linear regression (MLR), partial least squares regression (PLSR), least squares support vector machine (LS-SVM), back propagation neural network (BPNN), extreme learning machine (ELM), and radial basis function neural network (RBFNN), were developed and compared. With multi-source data fusion by combining vegetation indices (color and narrow-band VIs), robust prediction models of yield in oilseed rape were built. Balan T et al. [13] described a multi-sensor system for agriculture consisting of a soil probe, an air probe and a smart data logger. The implementation details will focus on the integration element and the innovative Artificial Intelligence based gas identification sensor. Furthermore, they focused on the analytics and decision support system implementation that provides farming recommendations and is enhanced with a feedback loop from farmers and a social trust index that will increase the reliability of the system.

Fei S et al. [14] studied on multi-sensor data fusion and ensemble learning for grain yield prediction in wheat. A set of thirty wheat cultivars and breeding lines were grown under three irrigation treatments i.e., light, moderate and high irrigation treatments to evaluate the yield prediction capabilities of a low-cost multi-sensor (RGB, multi-spectral and thermal infrared). Multi-sensor data fusion-based yield prediction showed higher accuracy compared to individual-sensor data in each ML model. Grunwald S et al. [15] collected DSC sensor data and spatially co-located soil cores that were sliced into narrow layers for laboratory-based soil measurements. Partial Least Squares Regression (PLSR) cross-validation was used to compare the results of testing four data integration methods. Method A reduced

the high-resolution sensor data to discrete values paired with layer-based soil laboratory measurements. Method B used stochastic distributions of sensor data paired with layer-based soil laboratory measurements. Method C allocated the same soil analytical data to each one of the high-resolution multi-sensor data within a soil layer. Method D linked the high-density multi-sensor soil data directly to crop responses (crop performance and behavior metrics), bypassing costly laboratory soil analysis. The soil models derived from Method C outperformed Methods A and B. Soil predictions derived using Method D were the most cost-effective for directly assessing soil–crop relationships, making this method well suited for industrial-scale precision agriculture applications.

### 3. PROPOSED SYSTEM

The proposed methodology provides a structured framework for analysing multi-sensor agricultural data to understand plant development and environmental influences in controlled environments. As shown in fig 2 the workflow begins with dataset acquisition, preprocessing, and exploratory analysis, followed by the application of analytical models for classification and regression tasks. These models analyse environmental and plant growth features to identify development stages and estimate growth parameters. The framework integrates traditional machine learning techniques with hybrid deep learning models to improve prediction accuracy and reliability. Neural network layers are used for feature extraction to capture complex relationships among variables, while probabilistic and regression models perform final predictions. A graphical interface enables users to upload datasets, preprocess data, train models, visualize results, and generate predictions. Efficient data handling and visualization tools support better understanding of variable relationships, while continuous model evaluation and retraining allow the system to adapt to new sensor data.

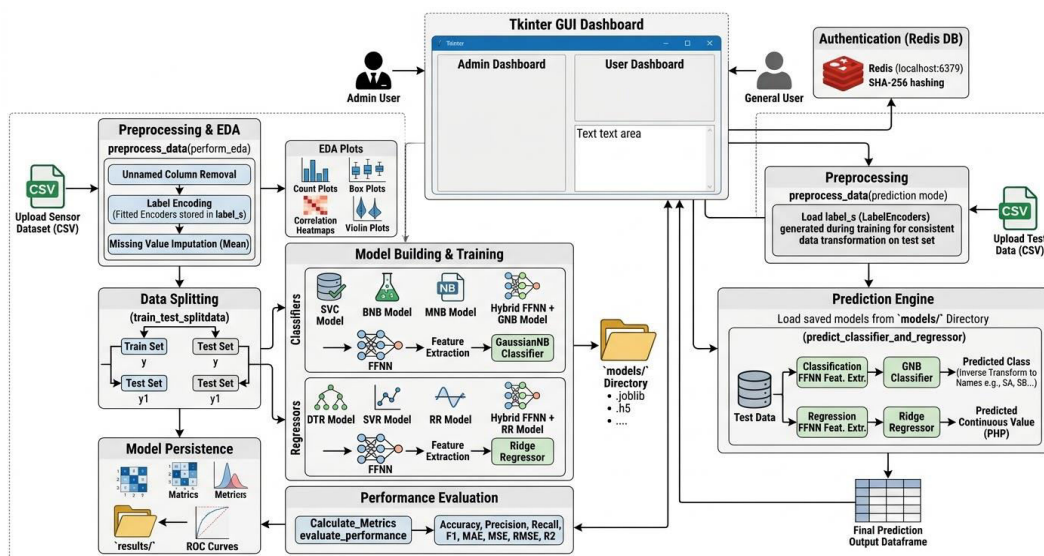


Fig. 2: Proposed system architecture.

#### User Interface (Tkinter Application)

- The user interacts with the system through a desktop graphical interface developed using Tkinter.
- The interface facilitates critical operations including user login, dataset uploading, preprocessing, model training, EDA visualization, and result generation.
- All actions are triggered via GUI buttons and passed to internal processing modules.

- The UI serves as the primary output display for model results, predictions, and analytical summaries.

#### **Authentication System (Redis Storage)**

- Includes a secure module for user registration and role-based login (Admin/User).
- User credentials are encrypted using hashed storage within a Redis database.
- Ensures that sensitive data analysis and prediction features are accessible only to authorized personnel.

#### **Dataset Input (CSV Sensor Data)**

- The primary input consists of CSV files containing multi-sensor agricultural data.
- The data includes environmental and physiological attributes such as root weight and vegetative growth indicators.
- Once uploaded via the UI, the data is forwarded directly to the preprocessing pipeline.

#### **Data Preprocessing and Feature Engineering**

- The raw dataset is cleaned by handling missing values and removing unnecessary columns.
- Categorical features are converted into numerical formats using Label Encoding.
- The final output is a structured feature vector optimized for training analytical models.

#### **Exploratory Data Analysis (EDA)**

- Provides a visual understanding of the dataset's characteristics and variable relationships.
- Generates automated visualizations including count plots, box plots, violin plots, and correlation heatmaps.
- Assists in identifying feature importance, data patterns, and potential anomalies before modeling.

#### **Existing Baseline Classification Models**

- The preprocessed features are passed to classical models to establish performance baselines:
  - **SVC**: Uses hyperplane-based separation for development classes.
  - **BNB**: A probabilistic approach based on feature distributions.
  - **MNB**: Analyzes feature frequency distributions for classification.
- These models predict development categories such as SA, SB, SC, TA, TB, and TC.

#### **Proposed DFPC**

- This is the core hybrid classification mechanism of the framework, consisting of two stages:
  1. **FFNN**: A Feed Forward Neural Network learns high-level, nonlinear representations of the sensor data through multiple hidden layers.
  2. **GNB**: The extracted deep features are fed into a Gaussian Naive Bayes classifier to improve classification reliability and accuracy.

#### **Existing Baseline Regression Models**

- Used to estimate continuous plant growth parameters alongside classification:

- **DTR:** Uses hierarchical rules to predict growth values.
- **SVR:** Utilizes margin-based optimization for value estimation.
- **RR:** Implements L2 regularization to prevent overfitting in linear estimations.

### Proposed HDRP

- The primary regression model for enhanced stability:
  1. Deep Feature Extraction (FFNN): A neural network processes sensor vectors to extract meaningful nonlinear patterns.
  2. Ridge Regression Prediction: The deep features are forwarded to a Ridge model which applies regularized regression to predict indicators like PHR.

### Prediction Output and Results

- The system generates dual outputs:
  - **Classification:** Identifies the specific plant development stage.
  - **Regression:** Provides numerical estimates of plant growth values.
- These results are visualized in the Tkinter UI for interpretation by the researcher.

### Model Evaluation and Performance Metrics

- **Classification Metrics:** Evaluated via Accuracy, Precision, Recall, and F1-score.
- **Regression Metrics:** Evaluated via MAE, MSE, RMSE, and  $R^2$  score.
- These metrics determine the effectiveness and reliability of both baseline and proposed models.

### Model Retraining and Adaptive Learning

- Supports an iterative lifecycle where models can be updated with new sensor data.
- Admins can upload fresh datasets through the interface to trigger retraining.
- Ensures the system adapts to evolving environmental conditions and maintains high accuracy over time.

## 4. RESULTS ANALYSIS

The results and discussion section presents the outcomes obtained from applying classification and regression models on the multi-sensor plant dataset. This stage evaluates how effectively the implemented models analyse environmental features and predict plant development stages and growth values. The experimental results are analysed using standard evaluation metrics such as accuracy, precision, recall, F1-score, mean absolute error, mean squared error, and  $R^2$  score. Graphical visualizations and statistical comparisons are used to observe the performance differences among the implemented models. The results provide insights into how different machine learning algorithms interpret sensor data and identify plant growth patterns. This analysis helps determine the most reliable models for accurate classification and prediction within the analytical framework.

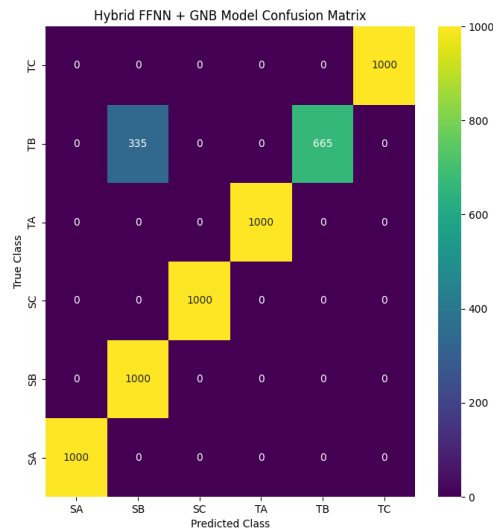


Fig. 3: Confusion matrices obtained using DFPC Model.

Fig. 3 shows confusion matrices for different models, with "Hybrid FFNN + GNB Model Confusion Matrix" with 3000 for SA vs. SA, 1000 for TA vs. TA, and 635 for SC vs. SC.

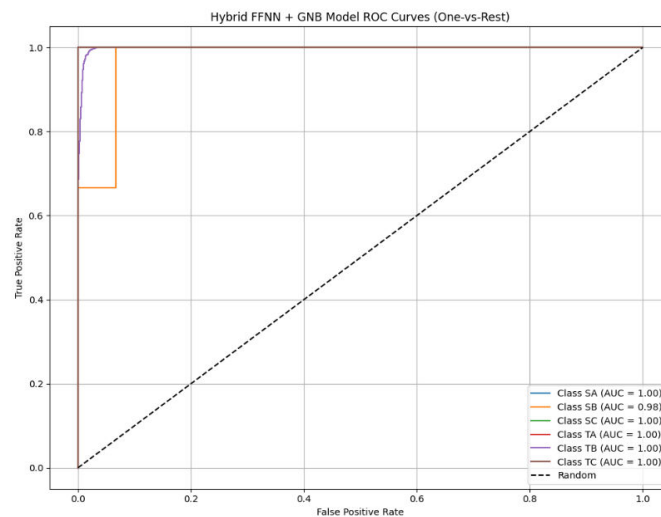


Fig. 4: AUC-ROC Curves obtained using DFPC Model.

Fig. 4 shows AUC-ROC curves obtained using different models, with "Hybrid FFNN + GNB Model ROC Curves (One-vs-Rest)" with Class SA (AUC = 1.00), Class SB (AUC = 1.00), Class SC (AUC = 1.00), Class TA (AUC = 1.00), Class TB (AUC = 1.00), and Class TC (AUC = 1.00) against a random AUC of 0.50.

Fig. 5 shows the best fit line (Predicted values vs True values) obtained using different models, with "HDRP Model vs Predicted" presenting a red diagonal line with blue data points also ranging from 0 to 60, indicating the performance of each model in aligning predicted and true values.

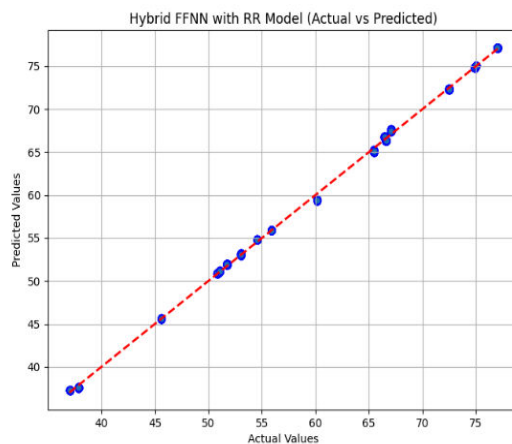


Fig. 5: Best fit line (Predicted values vs True values) obtained using Proposed HDRP Model.



Fig. 6: Sample predictions on new test data.

Fig. 6 shows sample predictions on new test data, displaying results such as "Class as => SA, Predicted Value => 54.84059143066406," "Class as => SB, Predicted Value => 75.0870819091769," "Class as => TA, Predicted Value => 53.201995849609375," and "Class as => TB, Predicted Value => 65.1302490234375," presented in a text format within the GUI under the "Prediction on Test Data" section.

#### 4.1 Comparative Analysis

The comparative analysis section evaluates the performance of different machine learning models implemented for plant development classification and growth prediction. This analysis compares how effectively each model interprets multi-sensor environmental data and generates accurate predictions. Performance metrics such as accuracy, precision, recall, and F1-score are used for classification models, while regression models are evaluated using mean absolute error, mean squared error, root mean squared error, and  $R^2$  score. By comparing these metrics, the study identifies the strengths and limitations of each analytical approach. The comparison helps determine which models provide the most reliable results when analyzing environmental factors that influence plant growth. This evaluation also highlights the effectiveness of hybrid models in improving predictive performance compared to traditional machine learning techniques.

The comparative analysis of classification models shows in table 1 significant differences in their ability to identify plant development stages from multi-sensor data. The SVC achieved an accuracy of 77.92%, with precision of 80.70%, recall of 77.92%, and an F1-score of 77.60%, indicating moderate classification performance. The BNB model produced very low performance with an accuracy of 16.53%, showing that it is not suitable for this dataset. The MNB model performed better with an accuracy of 83.70%, achieving precision of 85.05%, recall of 83.70%, and an F1-score of 83.55%. Among all models, the DFPC demonstrated the best performance with an accuracy of 94.42%, precision of 95.82%, recall of 94.42%, and an F1-score of 94.26%. These results indicate that the hybrid DFPC model provides superior classification capability by effectively extracting deep features and improving prediction accuracy compared to traditional machine learning models.

Table. 1: Performance comparison of classification models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVC	77.92	80.70	77.92	77.60
BNB	16.53	5.54	16.53	8.09
MNB	83.70	85.05	83.70	83.55
DFPC	94.42	95.82	94.42	94.26

Table. 2: Performance comparison of regression models.

Model	MAE	MSE	RMSE	R <sup>2</sup> Score
DTR	6.77	76.75	8.76	0.459
SVR	6.31	74.18	8.61	0.477
RR	4.27	38.39	6.20	0.729
HDRP	0.22	0.08	0.28	0.999

The comparative analysis of regression models shows in table 2 noticeable differences in their ability to predict plant growth values from multi-sensor data. The DTR produced a mean absolute error of 6.77, mean squared error of 76.75, root mean squared error of 8.76, and an R<sup>2</sup> score of 0.459, indicating relatively low prediction accuracy. The SVR showed slightly improved performance with a mean absolute error of 6.31, mean squared error of 74.18, root mean squared error of 8.61, and an R<sup>2</sup> score of 0.477. The RR performed significantly better with a mean absolute error of 4.27, mean squared error of 38.39, root mean squared error of 6.20, and an R<sup>2</sup> score of 0.729. Among all models, the HDRP demonstrated the best performance with a very low mean absolute error of 0.22, mean squared error of 0.08, root mean squared error of 0.28, and an R<sup>2</sup> score of 0.999. These results indicate that the hybrid HDRP model provides highly accurate growth prediction by combining deep feature extraction with regularized regression analysis.

## 5. CONCLUSION

The study demonstrates an effective analytical framework for interpreting multi-sensor agricultural data to classify plant development stages and predict plant growth indicators in controlled environments. The implementation integrates data preprocessing, exploratory data analysis, and multiple machine learning models to analyse environmental sensor data. Experimental results show that the DFPC

achieves the highest classification performance with an accuracy of 94.42%, outperforming traditional models such as SVC, BNB, and MNB. Similarly, for regression analysis, the HDRP produces highly accurate growth predictions with an  $R^2$  score of 0.999, significantly improving prediction reliability compared to conventional regression models. The hybrid models enhance analytical performance by combining deep feature extraction with probabilistic and regularized learning techniques. These results indicate that integrating deep learning with classical machine learning algorithms can significantly improve predictive performance when analysing complex environmental datasets.

## REFERENCES

- [1] Mia, M.S.; Tanabe, R.; Habibi, L.N.; Hashimoto, N.; Homma, K.; Maki, M.; Matsui, T.; Tanaka, T.S. Multimodal deep learning for rice yield prediction using UAV-based multispectral imagery and weather data. *Remote Sens.* 2023, 15, 2511.
- [2] Su, C.; Hong, J.; Wang, J.; Yang, Y. Quick and Accurate Counting of Rapeseed Seedling with Improved YOLOv5s and Deep-Sort Method. *Phyton* 2023, 92, 2611.
- [3] Natarajan, S.; Basnayake, J.; Wei, X.; Lakshmanan, P. High-throughput phenotyping of indirect traits for early-stage selection in sugarcane breeding. *Remote Sens.* 2019, 11, 2952.
- [4] Guo, Y.; Zhang, X.; Chen, S.; Wang, H.; Jayavelu, S.; Cammarano, D.; Fu, Y. Integrated UAV-based multi-source data for predicting maize grain yield using machine learning approaches. *Remote Sens.* 2022, 14, 6290.
- [5] Ahmed, S.; Nicholson, C.E.; Rutter, S.R.; Marshall, J.R.; Perry, J.J.; Dean, J.R. Use of an unmanned aerial vehicle for monitoring and prediction of oilseed rape crop performance. *PLoS ONE* 2023, 18, e0294184.
- [6] Han S, Wang H. Application of Deep Learning Technology in Monitoring Plant Attribute Changes. *Sustainability.* 2025; 17(17):7602. <https://doi.org/10.3390/su17177602>
- [7] Li B, Qiao B, Zhao Q, Yang D, Zhu R, Wang Z, Yang Y. A Predictive Model of the Photosynthetic Rate of Chili Peppers Using Support Vector Regression and Environmental Multi-Factor Analysis. *Horticulturae.* 2025; 11(5):502. <https://doi.org/10.3390/horticulturae11050502>
- [8] Roy AM, Bhaduri J. A Deep Learning Enabled Multi-Class Plant Disease Detection Model Based on Computer Vision. *AI.* 2021; 2(3):413-428. <https://doi.org/10.3390/ai2030026>
- [9] Manuelito C, Deus Jd, Damásio M, Leitão A, Conceição LA, Arias-Calderón R, Inês C, Cordeiro AM, Fernandes E, Albino L, et al. Multi-Sensor Comparison for Nutritional Diagnosis in Olive Plants: A Machine Learning Approach. *Applied Biosciences.* 2025; 4(3):32. <https://doi.org/10.3390/applbiosci4030032>
- [10] Xu R, Li C, Bernardes S. Development and Testing of a UAV-Based Multi-Sensor System for Plant Phenotyping and Precision Agriculture. *Remote Sensing.* 2021; 13(17):3517. <https://doi.org/10.3390/rs13173517>
- [11] Lee D-H, Park J-H. Development of a UAS-Based Multi-Sensor Deep Learning Model for Predicting Napa Cabbage Fresh Weight and Determining Optimal Harvest Time. *Remote Sensing.* 2024; 16(18):3455. <https://doi.org/10.3390/rs16183455>
- [12] Zhu H, Liang S, Lin C, He Y, Xu J-L. Using Multi-Sensor Data Fusion Techniques and Machine Learning Algorithms for Improving UAV-Based Yield Prediction of Oilseed Rape. *Drones.* 2024; 8(11):642. <https://doi.org/10.3390/drones8110642>

- [13] Balan T, Dumitru C, Dudnik G, Alessi E, Lesecq S, Correvon M, Passaniti F, Licciardello A. Smart Multi-Sensor Platform for Analytics and Social Decision Support in Agriculture. *Sensors*. 2020; 20(15):4127. <https://doi.org/10.3390/s20154127>
- [14] Fei S., Hassan, M.A., Xiao, Y. et al. UAV-based multi-sensor data fusion and machine learning algorithm for yield prediction in wheat. *Precision Agric* **24**, 187–212 (2023). <https://doi.org/10.1007/s11119-022-09938-8>
- [15] Grunwald S, Murad MOF, Farrington S, Wallace W, Rooney D. Multi-Sensor Soil Probe and Machine Learning Modeling for Predicting Soil Properties. *Sensors*. 2024; 24(21):6855. <https://doi.org/10.3390/s24216855>