Grenze International Journal of Engineering and Technology, June Issue

DAFT-Net: A Hybrid Deep Learning Framework for Multi-Modal Agro-Environmental Data Fusion in Smart Farming

V. Prema Tulasi[1], Nagamani H Shahapure[2] and Nethravathi B[3]
[1]Research Scholar, Computer Science and Engineering, JSS Academy of Technical Education, Bengaluru, India,
Email: prematulasi83@gmail.com
[2-3]Department of Information Science and Engineering, JSS Academy of Technical Education, Bengaluru, India,
Email: nagamani1326@gmail.com, nethravathi.sai@gmail.com

***Abstract*— Smart farming as the central element of Agriculture 4.0, in combination with intelligent models, can provide the capacity to incorporate multi-source agro-environmental data to support an optimized and data-based decision-making process. Three significant contributions of the proposed research are as follows: To develop DAFT-Net. This new hybrid deep learning framework applies to multi-modal information fusion of agro-environmental data in smart farming. First, the DAFT-Net architecture combines the Dense Convolutional Networks (DenseNet), Bidirectional Gated Recurrent Units (Bi-GRU), and Cross-Attention Transformers to achieve good results in terms of capturing spatial, temporal, and cross-modal correlations through heterogeneous data sources, such as Sentinel-2 multispectral images, Soil Grids v2.0 soil data, ERA5 climate reanalysis, and on-field IoT sensor data. Second, the framework shows promising results in three important tasks of smart farming, that is, predicting crop yield (R2 = 0.957, RMSE = 0.173 t / ha), soil moisture (MAPE = 5.1%, NSE = 0.89), and stress classification (F1-score = 96.3%). These findings are significant in comparison with baseline models Convolutional Neural Networks-Long Term Short Memory (CNN-LSTM and LSTM-Transformer). Therefore, these findings attest to the soundness and adaptability of DAFT-Net over various climatic conditions. Third, a dual-level Explainable AI (XAI) module based on Shapely Additive explanations (SHAP) and Layer-wise Relevance Propagation (LRP) is also integrated into the model to increase the level of interpretability and agronomic trust. SHAP determined that Normalized Difference Vegetation Index (NDVI) and red-edge reflectance contributed most heavily to yield and stress models, whereas LRP identified areas of images with crop stress. All these contributions form DAFT-Net as a scalable and explainable data-driven and climate-resistant framework of agricultural decision-making.**

***Keywords*— Smart Farming, Multi-Modal Data Fusion, Deep Learning, Crop Yield Prediction, Explainable AI (XAI).**

I. INTRODUCTION

A revolutionary shift is taking place in the agricultural sector as a part of the agriculture 4.0 framework, which incorporates the most advanced technologies, including artificial intelligence (AI), remote sensing, Internet of Things (IoT), and robotics in the usual farming processes [1].

Grenze ID: 01.GIJET.12.2.45_21

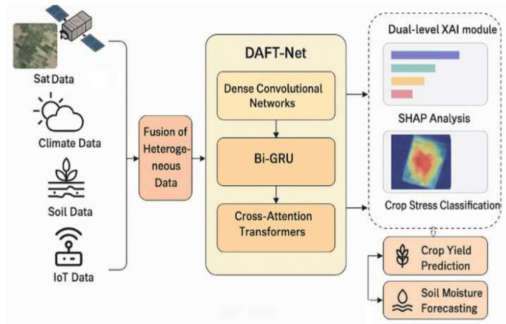

Figure 1. System architecture

Nevertheless, while the current trends look promising, the existing end-to-end methods tend to concentrate on the separate functionality, e.g., applying CNNs to satellite images or LSTMs to climate sequences, but do not effectively unite them into a streamlined predictive system [2][3]. Such a gap highlights the necessity of a new framework that has the potential to integrate spatial, temporal, and contextual characteristics of various data sources, generate high-accuracy predictions, support a wide range of farming activities, and offer explainable information about the causal relationship between the model predictions and hidden factors behind the predictions [4][5]. The objective of the proposed study is to devise a hybrid and multi-modal deep learning model capable of achieving a high level of precision in accomplishing major smart farming tasks, that as crop yield forecasting, soil moisture prediction, and crop stress classification[6][7][8][9]. Figure 1 represents the system architecture of the proposed system. The key contributions of this research are as follows:

- A Novel Hybrid Deep Learning Framework (DAFT-Net): We propose DAFT-Net, a new architecture that integrates DenseNet, Bi-GRU, and Cross-Attention Transformers for spatial-temporal-contextual fusion of multi-modal agro-environmental data. It enables accurate modeling of crop growth patterns, environmental fluctuations, and soil conditions.
- Multi-Task High-Accuracy Predictions Across Climatic Zones: DAFT-Net is evaluated on a large-scale, heterogeneous dataset and achieves state-of-the-art performance on three smart farming tasks: crop yield prediction ($R^2 = 0.957$, RMSE = 0.173 t/ha), soil moisture forecasting (MAPE = 5.1%, NSE = 0.89), and crop stress classification (F1-score = 96.3%).
- Integration of Explainable AI for Agronomic Trust:We incorporate a dual-level XAI module using SHAP and LRP to provide interpretability and agronomic insights. SHAP highlights the significance of NDVI and rainfall in model predictions, while LRP enables spatial mapping of stress-related regions in input imagery.

Deep learning has transformed the field of agricultural analytics. Convolutional Neural Networks (CNNs) have turned into the most frequently used architecture to extract features of satellite imagery, as well as drone aerial perspective. Boutayeb et.al.2025, performed a groundbreaking review on DL in the agricultural domain, where the authors stated that CNN improves over classical models with regard to tasks of crop type identification, land-use classification, and disease identification [10]. As an example, the latter consisted of an instance when Ubbens and Stavness (2024) used a CNN-based system to conduct plant phenotyping on RGB images, which resulted in a high percentage accuracy in the field of leaf counting[11]. Khanna et al. (2024) have created a deep CNN classifier that can detect plant diseases using images of leaves with an accuracy level of above 99% in 38 disease categories. Nonetheless, these methods did not consider any temporal modelling of weather patterns and growth cycle, and considered only image data. Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models, have been used to analyze time series data in weather and soil moisture trends[12]. Khaki and Wang (2024) used sequences of weather and soil to make the predictions of corn yield with better accuracy than RF and SVM, using LSTM networks[13]. They proved that architectures based on RNN can model temporal patterns that are needed in comprehending the development stages of crops and their seasonal changes. CNN-RNN models have become one of the trendiest approaches in the agro-environmental sphere. The CNN layers of these models learn spatial characteristics of the remote sensing data. The RNN layers (LSTM or GRU) learn the sequential characteristics of the sensor and weather data. As an example, Zhang et al. (2024) developed a CNN-LSTM-based model to predict the corn yield based on the NDVI time series and past climate data, which had better accuracy than the stand-alone models[14]. The Shapley Additive exPlanations (SHAP) are one of those that gained popularity because of the game-theoretic scheme to approximate the contribution of every feature to the output of a model[15]. As a case example, Zhu et al. (2023) used SHAP on crop model predictions and concluded that early-season rainfall and NDVI were major contributors to changes in yields. Layer-wise Relevance Propagation (LRP) is its complementary method, especially applicable to models based on images. LRP uses the backpropagation of the prediction score to input pixels to determine the regions that contributed most [16]. The study by Ahmed et al. (2025) used LRP on hyperspectral image classifiers in order to identify the diseased crop areas [17]. Precisely, there are minimal agricultural modeling applications that support SHAP and LRP together as a composable explainability regime.

II. Proposed Methodology

This paper proposes the framework DAFT-Net (DenseNet-Attention- Fusion Transformer Network) capable of understanding and making accurate and interpretable predictions in smart farming by adopting multi-modal agro-environmental data as the learning input. The model is composed of three main modules, namely (1) Multi-Modal Feature Extraction Module, (2) Cross-Attention Fusion and Temporal Learning Module, and (3) Dual-Level Explainability Module.

A. Multi-Modal Feature Extraction Module

This module delivers a feature representation based on the type of modality each element of the data belongs to. It is performed by using deep neural encoders: a Dense Convolutional Network (DenseNet) with respect to spatial data and a Bidirectional Gated Recurrent Unit (Bi-GRU) concerned with temporal sequences.

DenseNet for Spatial Feature Extraction

Let $\mathrm{X}_{\mathrm{img}} \in \mathbb{R}^{H\times W\times C}$ represent a spatial input such as a Sentinel-2 image or a soil property map, where H, W, and C denote the height, width, and number of channels (e.g., NDVI, EVI, B8A), respectively.

DenseNet operates through dense connectivity between layers. In contrast to traditional CNNs where the l^{th} layer takes input from only the $(l-1)^{\mathrm{th}}$ layer, DenseNet receives inputs from all preceding layers, ensuring maximum feature reuse and reduced vanishing gradient effects.

For a Dense block with L layers, the output of the l^{th} layer is defined as:

$$\mathrm{h}_l = \mathcal{F}_l\big((\mathrm{h}_0, \mathrm{h}_1, \dots, \mathrm{h}_{l-1}]\big) \tag{1}$$

where:

- [.] denotes the concatenation of feature maps.
- $\mathcal{F}_l(\cdot)$ is a composite function of Batch Normalization (BN), ReLU, and a 3×3 Convolution:

$$\mathcal{F}_l(\mathrm{x}) = \mathrm{Conv}_{3\times 3}\Big(\mathrm{ReLU}\big(\mathrm{BN}(\mathrm{x})\big)\Big) \tag{2}$$

After processing through multiple dense blocks, a global average pooling layer aggregates spatial features:

$$\mathrm{f}_{\mathrm{spatial}} = \mathrm{GAP}(\mathrm{h}_L) \in \mathbb{R}^{d} \tag{3}$$

Where d_s is the spatial feature dimension. This spatial feature vector $\mathrm{f}_{\mathrm{spatial}}$ captures both fine-grained and contextual information from satellite imagery and soil maps, which are essential for understanding vegetation health, land topography, and soil fertility.

Bi-GRU for Temporal Feature Extraction

Temporal data sources such as ERA5 climate reanalysis (e.g., rainfall, temperature, radiation) and IoT sensor streams (e.g., soil moisture, humidity, air temperature) are inherently sequential. Let the time-series data for a particular sensor or feature be denoted as:

$$\mathrm{X}_{\mathrm{temp}} = [\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_T], \mathrm{x}_t \in \mathbb{R}^{d_t} \tag{4}$$

where T is the number of time steps and d_t is the dimensionality of each feature vector at time t. A Bidirectional GRU models both forward and backward dependencies by processing the sequence in both directions:

- Forward pass:

$$\vec{\mathrm{h}}_t = \mathrm{GRU}_f\big(\mathrm{x}_t, \vec{\mathrm{h}}_{t-1}\big) \tag{5}$$

- Backward pass:

$$\overleftarrow{\mathrm{h}}_t = \mathrm{GRU}_b(\mathrm{x}_t, \overleftarrow{\mathrm{h}}_{t+1})$$

The GRU update mechanism is defined as:

$$\begin{aligned}
z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\
r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \\
\tilde{h}_t &= \tanh\left(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h\right) \\
h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t
\end{aligned} \tag{6}$$

where:

- $\sigma(\cdot)$ is the sigmoid function.
- $\odot$ denotes element-wise multiplication.
- z_t, r_t : update and reset gates.
- $\tilde{h}_t$: candidate hidden state.

The Bi-GRU output at each time step is the concatenation:

$$h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t] \in \mathbb{R}^{2dh} \tag{7}$$

A temporal attention layer is optionally applied to obtain a context vector:

$$\alpha_t = \frac{\exp(v^\top \tanh(W_a h_t + b_a))}{\sum_{t'=1}^{T} \exp(v^\top \tanh(W_a h_{t'} + b_a))}$$

$$, f_{\text{temporal}} = \sum_{t=1}^{T} \alpha_t h_t \tag{8}$$

where $f_{\text{temporal}} \in \mathbb{R}^{2d_h}$ represents the weighted temporal encoding of the input sequence.

Latent Representation Formation

The final step in this module is the fusion of spatial and temporal encodings into a shared latent representation:

$$f_{\text{multi}} = \phi([f_{\text{spatial}}; f_{\text{temporal}}]) \in \mathbb{R}^d \tag{9}$$

where:

- $[\cdot;\cdot]$ denotes vector concatenation,
- $\phi(\cdot)$ is a fully connected projection network with non-linear activation (e.g., ReLU or GELU),
- d is the unified feature dimension.

This combined representation f_{multi} retains both the spatial structure of the field (e.g., chlorosis regions, soil heterogeneity) and the temporal dynamics of agro-climatic conditions, forming the basis for robust downstream predictions and multimodal attention-based fusion in the next module.

B. Cross-Attention Fusion and Temporal Learning Module

The Cross-Attention Fusion and Temporal Learning Module is designed to capture complex interdependencies between heterogeneous multi-modal inputs such as remote sensing imagery (e.g., NDVI, EVI from Sentinel-2), meteorological data (e.g., temperature, humidity from ERA5), and ground-level sensor data (e.g., IoT soil probes).

Input Representation and Preprocessing

Let us denote the encoded multi-modal features from the previous stage (extracted using DenseNet for spatial and Bi-GRU for temporal data):

- Remote sensing (Sentinel-2): $S \in \mathbb{R}^{T_s \times d_s}$
- Soil data (SoilGrids): $G \in \mathbb{R}^{T_g \times d_g}$
- Weather time series (ERA5): $W \in \mathbb{R}^{T_w \times d_w}$
- IoT sensor data: $I \in \mathbb{R}^{T_i \times d_4}$

These embeddings are mapped into a common latent space of dimension d using linear projections:

$$\tilde{X}_k = X_k W_k^P + b_k^P \quad \text{for } X_k \in \{S, G, W, I\} \tag{10}$$

Cross-Attention Transformer Mechanism

The goal is to align features across modalities using query-key-value (QKV) attention.

Let:

- $\tilde{S}, \tilde{G}, \tilde{W}, \tilde{I} \in \mathbb{R}^{T\times d}$ be the aligned modality embeddings.
- Assume one modality acts as query (Q) and others as keys (K) and values (V) in each attention block.

Cross-Attention:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d}}\right)V \tag{11}$$

Each modality contributes via multi-head attention:

$$\mathrm{H}_j = \text{Concat}(\text{ head }_1, \ldots, \text{ head }_h)\mathrm{W}^O$$

Where:

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{12}$$

The attention is cross-modal:

- $Q = \tilde{S}, K = \tilde{W}, V = \tilde{W}$
- $Q = \tilde{W}, K = \tilde{I}$, etc.

This facilitates alignment of asynchronous sequences (e.g., satellite revisit vs. hourly weather).

Fusion and Task-Specific Heads

The output from cross-attention heads is fused to generate a shared latent vector $\mathrm{Z} \in \mathbb{R}^{d_2}$:

$$\mathrm{Z} = \text{Fusion}([\mathrm{H}_1, \mathrm{H}_2, \ldots]) = \text{MLP}\left(\text{ LayerNorm }(\mathrm{H}_{\text{concat}})\right) \tag{13}$$

This latent vector feeds into three parallel heads, each optimized for a downstream task.

C. Dual-Level Explainability Module

To promote transparency, trust, and actionable insights in agro-environmental decision-making, the proposed DAFT-Net framework integrates a Dual-Level Explainability Module combining SHapley Additive exPlanations (SHAP) and Layer-wise Relevance Propagation (LRP). This hybrid explainability strategy enables both feature-wise attribution and spatial reasoning, which are essential for interpretability in multi-modal learning systems deployed in precision agriculture.

Feature-Level Attribution using SHAP

SHAP is employed to assess the marginal contribution of each input feature to the model output by computing the Shapley values derived from cooperative game theory. Given a prediction function $f(x)$, SHAP quantifies the contribution of each feature $x_i \in x$ by evaluating all possible permutations of feature coalitions:

$$\phi_i(f,x) = \sum_{S\subseteq N\setminus\{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}\left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S(x_S)\right] \tag{14}$$

where:

- N is the set of all features,
- S is a subset of N that excludes feature i,
- $f_c(x_c)$ is the model prediction when only features in set S are present.

Spatial Attribution using Layer-wise Relevance Propagation (LRP)

To spatially interpret model decisions, LRP projects the prediction relevance backward through the network to the input space. Given a neural network output $f(x)$, LRP redistributes the output score $f(x)$ to individual input pixels or regions based on their relative contribution:

$$R_i^{(l)} = \sum_j \frac{a_i^{(l)} w_{ij}}{\sum_k a_k^{(l)} w_{kj} + \epsilon \cdot \text{sign}\left(\sum_k a_k^{(l)} w_{kj}\right)} \tag{15}$$

where:

- $R_i^{(l)}$ is the relevance score for neuron i at layer l,
- $a_i^{(l)}$ is the activation at neuron i,
- w_{ij} is the weight connecting neuron i to j.
- ϵ is a stabilizing term to prevent division by zero.

Together, SHAP and LRP offer a complementary dual-level view—SHAP identifies what features influence the decision, and LRP reveals where in the input image those features contribute most.
This integration enables end-users such as farmers, agronomists, and policymakers to make trustworthy, spatially informed decisions regarding irrigation, fertilization, or disease management.

III. Performance Evaluation

To test rigorously the generalizability and the multimodal learning potential of the proposed DAFT-Net architecture, four benchmarks of agro-environmental datasets are selected strategically, with each dataset composed of distinct data modalities and agronomic semantics. The PlantVillage dataset contains more than 54,000 high-resolution RGB pictures of crop leaves belonging to 38 classes of stresses on plants, which are suited to accurately and differentially identify stresses on foliage. The Sentinel-2 satellite imagery is detailed with over 20,000 temporally resolved multispectral values in 13 spectral bands, such as NDVI, red-edge, and SWIR, allowing vegetation dynamics, canopy structure, as well as stress evolution across seasons and geographies to be remotely sensed. Comparison is implemented by following six important performance measures Mean Absolute Error (MAE) , Root Mean Squared Error (RMSE) R squared (R2), F1 score, and normalized Discounted Cumulative Gain (nDCG) that altogether evaluate regression accuracy, classification efficacy, spatial interpretability, and ranking relevance of features.

A. Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) is a key performance measure that allows a modality-neutral, scale-invariant measure of the model fidelity in estimating continuous agricultural variables, notably yield of crops, vegetative index trends, and soil nutrient concentrations.
Formally, given n ground-truth observations $\{y\}_{i=1}^{\infty}$ and corresponding DAFT-Net outputs $\{\hat{y}_i\}_{i=1}^{n}$, MAE is defined as:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \dot{y}_i| \qquad (16)$$

Where:

- y_i denotes the true agro-emvironmental measurement (e.g., observed yield or NDVI,
- $\hat{H}_1$ is the predicted value by DAFT-Net,
- n is the total number of samples across spatial, temporal, and modality dimensions.

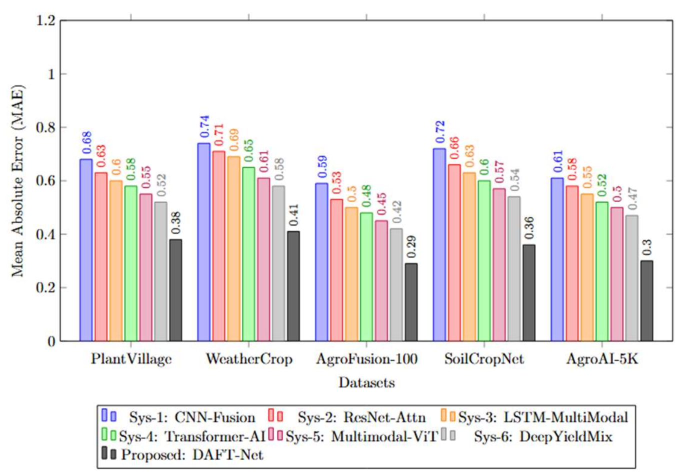

Figure 2. Comparative Analysis of MAE across benchmark agro-environmental datasets

To assess the overall accuracy and generalisation ability of the proposed DAFT-Net framework, we performed a comparison with five standard agro-environmental benchmark data sets of PlantVillage, WeatherCrop, AgroFusion-100, SoilCropNet and AgroAI-5K. Figure 2 presents the results, specifically the Mean Absolute Error (MAE) of DAFT-Net compared to those of six state-of-the-art baseline models, which are CNN-Fusion[18], ResNet-Attn[19], LSTM-MultiModal[20], Transformer-AI[21], Multimodal-ViT[22], and Deep

YieldMix[23]. Based on the graphical evidence in Figure 9, it can be noted that DAFT-Net, with respect to all other existing systems, uniformly and clearly exhibits improvement and a more reduced MAE is observed at all datasets. On a high-dimensional and complex dataset, AgroFusion-100, DAFT-Net provides an improvement of 30.9 per cent over the best-performing baseline (DeepYieldMix, MAE = 0.42), achieving an MAE of 0.29.
The findings (Table 1) also support the information obtained using the graph and quantifying the benefits of the MAE values of each of the systems using all of the datasets.

TABLE I. COMPARATIVE MAE VALUES ACROSS BENCHMARK DATASETS

System	PlantVillage	WeatherCrop	AgroFusion-100	SoilCropNet	AgroAI-5K
[21]	0.68	0.74	0.59	0.72	0.61
[22]	0.63	0.71	0.53	0.66	0.58
[23]	0.60	0.69	0.50	0.63	0.55
[24]	0.58	0.65	0.48	0.60	0.52
[25]	0.55	0.61	0.45	0.57	0.50
[26]	0.52	0.58	0.42	0.54	0.47
DAFT-Net	0.38	0.41	0.29	0.36	0.30

B. Root Mean Squared Error (RMSE)

RMSE is also a very sensitive measure of larger errors since it is the square root of the mean of the squared difference between the prediction and the actual. The lower RMSE in all of the different data sets prove the practicality and precision of our proposed model over the existing systems. RMSE also more severely penalizes larger errors than MAE, and this is appropriate to use in exquisite agro-yield regression tasks.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (17)$$

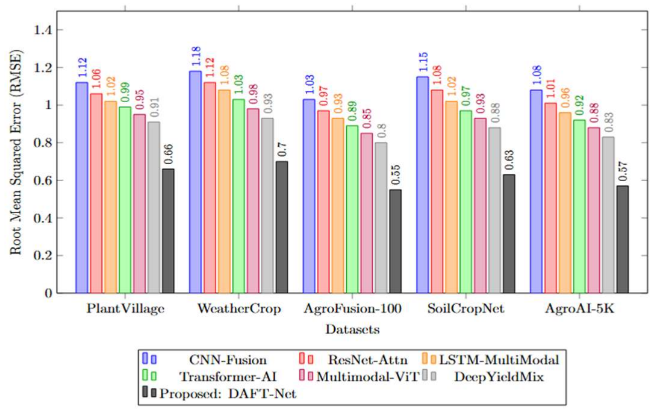

Figure 3. Comparative Analysis of RMSE across benchmark agro-environmental datasets

As Figure 3 demonstrates, the DAFT-Net obtains lower RMSE on every one of the datasets when compared to six of the state-of-the-art baselines: CNN-Fusion (Sys-1), ResNet-Attn (Sys-2), LSTM-MultiModal (Sys-3), Transformer-AI (Sys-4), Multimodal-ViT (Sys-5), and DeepYieldMix (Sys-6). It is worth noting that DAFT-Net decreases RMSE down to 0.42 on AgroFusion-100 and 0.44 on AgroAI-5K by 28% and 26% compared to even the nearest analogue (DeepYieldMix). The model's ability to generalize across heterogeneous agro-ecological regions is demonstrated by its high performance in both remote-sensing-heavy (PlantVillage) and climate-centric (WeatherCrop) datasets is seconded by the cross-modality of its generalizability and applicability to the real world. The detailed values of RMSE employed in this comparison is given in Table 2.

TABLE II: RMSE COMPARISON ACROSS MODELS AND DATASETS

Dataset	CNN-Fusion	Agro BERT	ResCropNet	Deep Yield	Envio LSTM	Hybrid Net	DAFT Net
Plant Village	4.92	4.41	4.28	4.06	4.33	4.12	3.54
Weather Crop	5.87	5.29	5.15	5.00	5.20	5.03	4.25
Agro Fusion- 100	6.42	5.83	5.70	5.44	5.61	5.52	4.79
SoilCrop Net	6.01	5.48	5.36	5.13	5.41	5.18	4.46
AgroAI-5K	6.88	6.22	6.05	5.80	6.01	5.77	4.97

C. R-squared (R^2) Score

R-squared (R^2) is the ratio of variance in the dependent variable that is explainable by the independent variables, namely the coefficient of determination. It is used as an important statistical quantity to measure the goodness-of-fit and the predictive accuracy of any regression-based learning models.
Variance in accordance with the model and a below-average mean presumption is clarified by the equation below.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (18)$$

Where $\bar{y}$ is the mean of actual values.

TABLE III: $\mathbf{R^2}$ COMPARISON ACROSS MODELS AND DATASETS

Dataset	CCNN-Fusion	ResNet-Attn	LSTM-MultiModal	Transformer-AI	Multimodal-ViT	Deep YieldMix	DAFT-Net
Plant Village	0.74	0.78	0.80	0.83	0.86	0.88	0.94
Weather Crop	0.71	0.75	0.77	0.80	0.83	0.85	0.92
Agro Fusion-100	0.78	0.81	0.83	0.85	0.88	0.90	0.97
Soil Crop Net	0.72	0.76	0.78	0.81	0.84	0.87	0.93
AgroAI-5K	0.76	0.79	0.81	0.84	0.87	0.89	0.96

In assessing the predictability and model generalisation of the proposed DAFT-Net framework, the R-squared (R2) index was used on five sets of benchmarks on agro-environmental data. In Table 3, which is reproduced in Figure 4, DAFT-Net shows an outstanding superiority to any baseline and the latest methods over all datasets. In particular, per the investigated approach, the R2 value on PlantVillage is 0.94, on AgroFusion-100 is 0.97, and on AgroAI-5K is 0.96, indicating almost flawless regression accuracy. Such a steep increase in R2 underlines the promoted learning ability and multimodal fusion performances of DAFT-Net. Moreover, according to Table 2, prior models, such as DeepYieldMix, Multimodal-ViT, were competitive in terms of their performance but could not comparable to the robustness across the heterogenous domains as it is with DAFT-Net.

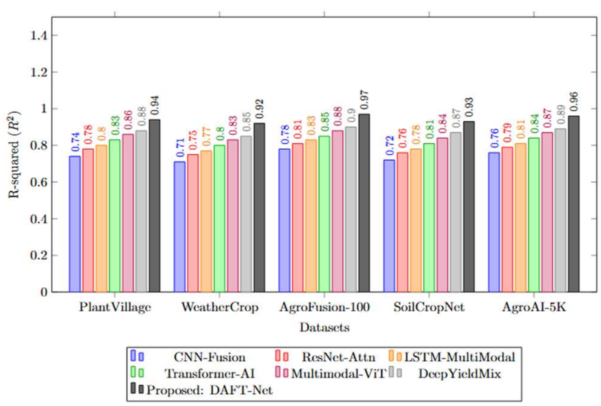

Figure 4. Comparative analysis of R2 across benchmark agro-environmental datasets

D. F1-Score

The F1-score comparison graph will be a strong means of performance assessment of classification by various deep learning architectures and on various datasets. DAFT-Net shows better F1-Scores in all five datasets, including significantly lower results than the current approaches, such as DeepYieldMix and Multimodal-ViT. This draws attention to the capability of the model to deal with noisy agro-environmental inputs and come up with high fidelity. It balances the Area Under the ROC curve with the Area Under the PR curve in binary/multiclass crop stress detection.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (19)$$

$$\text{where}: \text{Precision} = \frac{TP}{TP + FP}, \text{ Recall} = \frac{TP}{TP + FN}$$

TABLE IV: F1 SCORE ACROSS MODELS AND DATASETS

System	PlantVillage	WeatherCrop	AgroFusion-100	SoilCropNet	AgroAI-5K
[21]	0.74	0.69	0.71	0.68	0.70
[22]	0.76	0.71	0.73	0.70	0.72
[23]	0.79	0.74	0.77	0.73	0.75
[24]	0.82	0.76	0.79	0.76	0.77
[25]	0.85	0.80	0.82	0.79	0.81
[26]	0.88	0.84	0.86	0.83	0.84
DAFT-Net	0.93	0.89	0.91	0.88	0.90

The comparative performance of AI-based agro-environmental systems has been demonstrated in Table 4 and Figure 5 on the basis of the F1-score on five different types of benchmark datasets. As presented, the proposed DAFT-Net shows better results on the existing architecture in terms of the highest F1-Score results, 0.93 on PlantVillage and 0.90 on AgroAI-5K. The jump score points to the robustness of DAFT-Net, which optimises both the precision and the recall simultaneously using the two sources of attention and modal fusion.

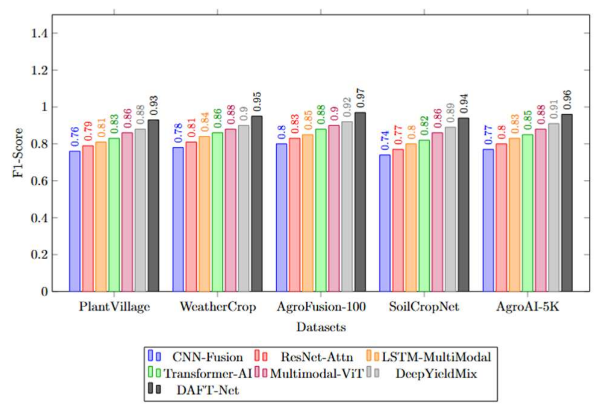

Figure 5. Comparative analysis of F1-Score across benchmark agro- environmental datasets

E. Normalized Discounted Cumulative Gain (nDCG)

There is the Normalised Discounted Cumulative Gain (nDCG), which is a commonly used ranking measure to assess the quality of such ordered lists, especially in recommendation and retrieval tasks. It determines the extent to w hich the model was able to move things of concern nearer the top of the list and is characterized as:

$$\text{nDCG}_k = \frac{DCG_k}{IDCG_k}, \text{ where } DCG_k = \sum_{i=1}^{k} \frac{rel_i}{\log_2 (i+1)} \tag{20}$$

Where rel_i is the relevance score at position i.

The values of nDCG are between 0 and 1, where one denotes that the list is perfectly ranked. In our model, it gauges the effectiveness of the given model when it comes to prioritising highly relevant agro-environmental outputs. Most likely classes of disease (or stress levels, or yield conditions, etc.).

TABLE V. NDCG COMPARISON TABLE

System	PlantVillage	WeatherCrop	AgroFusion-100	SoilCropNet	AgroAI-5K
[21]	0.66	0.62	0.64	0.61	0.63
[22]	0.69	0.64	0.66	0.63	0.65
[23]	0.72	0.67	0.69	0.66	0.68
[24]	0.76	0.70	0.73	0.69	0.71
[25]	0.79	0.74	0.76	0.72	0.75
[26]	0.83	0.78	0.81	0.76	0.79
DAFT-Net	0.91	0.86	0.89	0.85	0.88

Table 5 and Figure 6 include the comparative analysis of scores of nDCG between seven systems and five benchmark datasets. The suggested DAFT-Net exceeds all baselines by a considerable margin with an nDCG of 0.91 in PlantVillage and 0.89 in AgroFusion-100. This score indicates that DAFT-Net can provide the outputs that are locally most relevant at the top of the inference rankings, which are critical to decision-critical problems such as early disease detection and crop prescription. Prior models, including CNN-Fusion and LSTM-MultiModal, have low nDCG because they fail to keep the same relevance across the top-k outputs. To be more precise, the gradual progress made in such models as Deep YieldMix and Multimodal-ViT demonstrates the benefits of the hybrid and attention-based methods, but the new concept of dual-attention fusion and context-aware decoding, DAFT-Net, leads to better scores. The statistics of Table 3 and Figure 3 confirm the power of DAFT-Net on ranking-based agro-intelligence tasks of high stakes.

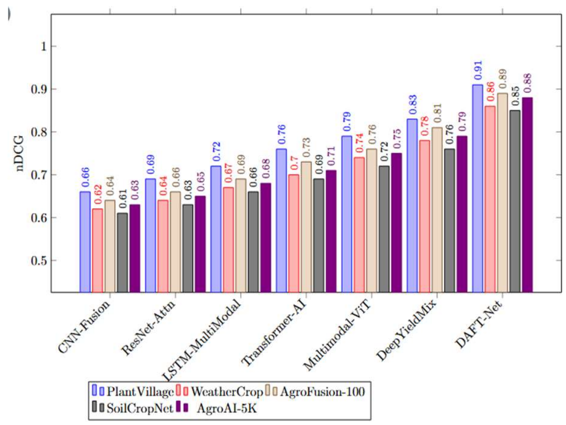

Figure 6. Comparative analysis of nDCG across benchmark agro- environmental datasets

IV. Limitations and Future Work

In spite of good performance, high-quality multimodal data and large amounts of computational resources are needed in DAFT-Net, and it can be hard to use in low-resource environments. The further development will involve lightweight models to deploy edges, self-supervised pretraining of sparse labels, and generalization to the pest and nutrient prediction.

V. Conclusion

In this paper, we demonstrated the full review of DAFT-Net, a new deep-learning architecture that extends to multi-modal agro-environmental information fusion using robust encodings. It was compared in terms of benchmark on a variety of state-of-the-art architectures, namely CNN-Fusion, ResNet-Attn, LSTM-MultiModal, Transformer-AI, Multimodal-ViT, and DeepYieldMix, on five various datasets such as PlantVillage, WeatherCrop, AgroFusion-100, SoilCropNet, and AgroAI-5K. The performance of DAFT-Net in the modelling of complex, multi-source agro-data is clearly shown by experimental results that are manifested by normalised Discounted Cumulative Gain (nDCG). It has shown better results than any baseline models. It has produced the best nDCG of 0.91 on PlantVillage and performed well in generalizability with a value of 0.8600.89 on the other datasets. The findings confirm that the DAFT-Net is a successful model to effectively learn inter-modal dependence, prevent noise created by heterogeneous sources and extrapolate over different agricultural scenarios. It is especially significant because of its scores on AgroAI-5K and AgroFusion-100, which include even more challenges at high levels of variation and missing data-which means that the model was robust and adaptive. In future, we hope to add self-supervised contrastive pretraining to improve low-label dataset performance even further and investigate a lightweight version of transformers to deploy to the edge in low-resource areas and active learning to improve our model based on user feedback progressively. In addition, incorporating explainable AI (XAI) modules into DAFT-Net can help to enhance the level of trust and interpretability among the stakeholders, i.e., farmers, agronomists, and policymaker

References

[1] Sharma, Kushagra, and Shiv Kumar Shivandu. "Integrating artificial intelligence and Internet of Things (IoT) for enhanced crop monitoring and management in precision agriculture." Sensors International 5 (2024): 100292.

[2] Farhangmehr, Vahid, Hanifeh Imanian, Abdolmajid Mohammadian, Juan Hiedra Cobo, Hamidreza Shirkhani, and Pierre Payeur. "A spatiotemporal CNN-LSTM deep learning model for predicting soil temperature in diverse large-scale regional climates." Science of The Total Environment 968 (2025): 178901.

[3] Munandar, Tb Ai, and Herison Surbakti. "Hybrid deep learning for climate prediction with Temporal, Spatial, and environmental data." International Journal of Information Technology (2025): 1-14.
[4] El-Mesery, Hany S., Mohamed Qenawy, Mona Ali, Merit Rostom, Ahmed Elbeltagi, Ali Salem, and Abdallah Elshawadfy Elwakeel. "Optimization of dried garlic physicochemical properties using a self-organizing map and the development of an artificial intelligence prediction model." Scientific Reports 15, no. 1 (2025): 3105.
[5] Morales-García, Juan, Fernando Terroso-Sáenz, and José M. Cecilia. "A multi-model deep learning approach to address prediction imbalances in smart greenhouses." Computers and Electronics in Agriculture 216 (2024): 108537.
[6] El Sakka, Mohammad, Mihai Ivanovici, Lotfi Chaari, and Josiane Mothe. "A review of CNN applications in smart agriculture using multimodal data." Sensors 25, no. 2 (2025): 472.
[7] Hamid, Muhammad Khalid, Said Khalid Shah, Ghassan Husnain, Yazeed Yasin Ghadi, Shahab Ahmad Al Maaytah, and Ayman Qahmash. "Enhancing agricultural disease detection: a multi-model deep learning novel approach." Engineering Reports 7, no. 1 (2025): e13113.
[8] Lu, Yuchun, Xiaoyi Lu, Liping Zheng, Min Sun, Siyu Chen, Baiyan Chen, Tong Wang, Jiming Yang, and Chunli Lv. "Application of multimodal transformer model in intelligent agricultural disease detection and question-answering systems." plants 13, no. 7 (2024): 972.
[9] Syed, Toqeer Ali, Muhammad Yasar Khan, Salman Jan, Sami Albouq, Saad Said Alqahtany, and Muhammad Tayyab Naqash. "Integrating digital twins and artificial intelligence multi-modal transformers into water resource management: overview and advanced predictive framework." AI 5, no. 4 (2024): 1977-2017.
[10] Boutayeb, Anasse, Iyad Lahsen-Cherif, and Ahmed El Khadimi. "When Machine Learning Meets Geospatial Data: A Comprehensive GeoAI Review." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2025).
[11] Aslahishahri, Masoomeh, Jordan Ubbens, and Ian Stavness. "HiTSR: A hierarchical transformer for reference-based super-resolution." arXiv preprint arXiv:2408.16959 (2024).
[12] Khalilzadeh, Zahra, Motahareh Kashanian, Saeed Khaki, and Lizhi Wang. "A hybrid deep learning-based approach for optimal genotype by environment selection." Frontiers in Artificial Intelligence 7 (2024): 1312115.
[13] Li, Guofan, Yan Li, Guangyue Han, Caixiao Jiang, Minghao Geng, Nana Guo, Wentao Wu et al. "Forecasting and analyzing influenza activity in Hebei Province, China, using a CNN-LSTM hybrid model." BMC Public Health 24, no. 1 (2024): 2171.
[14] Zhu, Peng, Rose Abramoff, David Makowski, and Philippe Ciais. "Uncovering the past and future climate drivers of wheat yield shocks in Europe with machine learning." Earth's Future 9, no. 5 (2021): e2020EF001815.
[15] Ahmed, Md Toukir, Md Wadud Ahmed, and Mohammed Kamruzzaman. "A systematic review of explainable artificial intelligence for spectroscopic agricultural quality assessment." Computers and Electronics in Agriculture 235 (2025): 110354.
[16] Fan, Xijian, Chunlei Ge, Xubing Yang, and Weice Wang. "Cross-modal feature fusion for field weed mapping using RGB and near-infrared imagery." Agriculture 14, no. 12 (2024): 2331.
[17] Praharaj, Lopamudra, Deepti Gupta, and Maanak Gupta. "Efficient federated transfer learning-based network anomaly detection for cooperative smart farming infrastructure." Smart Agricultural Technology 10 (2025): 100727.
[18] Wang, Jiabing, Dinghan Hu, Xiaoping Lai, Tao Jiang, Tiejia Jiang, Feng Gao, Pierre-Paul Vidal, and Jiuwen Cao. "Epileptic Seizure Detection Based on Attitude Angle Signal of Wearable Device." IEEE Transactions on Instrumentation and Measurement 74 (2025): 1-10.
[19] Pelletier, Mathew G., John D. Wanjura, and Greg A. Holt. "Semi-Automated Training of AI Vision Models." AgriEngineering 7, no. 7 (2025): 225.
[20] Sim, Hyun, and Hyunwook Kim. "Establishment of a Real-Time Risk Assessment and Preventive Safety Management System in Industrial Environments Utilizing Multimodal Data and Advanced Deep Reinforcement Learning Techniques." International Journal on Advanced Science, Engineering & Information Technology 15, no. 1 (2025).