



SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

A Deep Hierarchical Feature Learning Framework for Multimodal Tomato Growth Stage Classification and Yield Prediction

Neeta B. Bankhele¹, Rekha P. Labade², Sachin V. Chaudhari³

¹ PG Research Center in Electronics & Telecommunication Engineering, Sanjivani College of Engineering & Research, Savitribai Phule Pune University (SPPU), Pune, Maharashtra, India. E-mail: nbankhele.7777@gmail.com

² PG Research Center in Electronics & Telecommunication Engineering, Sanjivani College of Engineering & Research, Savitribai Phule Pune University (SPPU), Pune, Maharashtra, India. E-mail: rplabade@gmail.com

³ PG Research Center in Electronics & Telecommunication Engineering, Sanjivani College of Engineering & Research, Savitribai Phule Pune University (SPPU), Pune, Maharashtra, India. E-mail: chaudharisachinetc@sanjivani.org.in

Abstract

Precise monitoring of tomato plant development is vital for precision agriculture and sustainable crop management. In this study, we present a Deep Hierarchical Feature Learning Network (DHFL-Net) for automatic tomato growth-stage categorization and yield estimation using paired RGB and temperature pictures. The research employs 700 paired samples consisting of 700 RGB photos and 700 related thermal images of the same tomato plants. Preprocessing using Adaptive Noise Suppression Module (ANSM) and Plant-Aware Region Extraction Module (PAREM) to extract informative plant areas. DHFL-Net fuses and manipulates RGB morphology and Multi-Scale Thermal Texture Encoder (MTTE) descriptors. All the comparing algorithms were reimplemented and tested on the same dataset split for a fair comparison. The DHFL-Net attained an accuracy of 98.76%, precision of 98.42%, recall of 98.18% and F1-score of 98.30% respectively. It achieved $R^2=0.982$, RMSE=0.084 and MAE=0.061 for yield prediction. Results indicate the value of combined RGB-thermal information for robust tomato phenotyping and yield calculation.

Keywords—Tomato phenotyping; RGB imaging; thermal imaging; multimodal learning; growth-stage classification; yield prediction; DHFL-Net; precision agriculture.

1. Introduction

Tomatoes need constant monitoring of crop development, as canopy growth, flowering, fruit set, fruit filling and maturity directly affect irrigation, nutrition, crop protection and harvest timing. Traditional field assessments rely mainly on eye examination and destructive measures. Such approaches are hard to scale and may cause observer variability. Image-based phenotyping is a non-destructive approach where sequential observations may be translated into quantitative descriptors of plant structure and physiological response.

RGB cameras are affordable and provide visual features like plant width, projected canopy area, greenness, leaf distribution, and texture. However, RGB appearance is sensitive to light and may not reflect physiological differences between visually identical plants. Thermal cameras provide an additional depiction of surface-temperature trends. When all modalities are seen simultaneously, structural and thermal signals may be utilized to generate a more meaningful description of plant development.

Deep convolutional networks have been extensively used for plant identification, illness classification, fruit detection and phenotyping. However, a single deep backbone may omit interpretable agronomic characteristics, and a solely handmade feature pipeline may miss complicated nonlinear interactions. Thus, in this work, we propose a unified multimodal architecture that integrates explicit RGB phenotypic characteristics, multiscale thermal texture encoding and hierarchical deep feature learning.

The core idea is that the regulated merging of RGB and thermal representations leads to improvements in both discrete growth-stage detection and continuous yield prediction. To objectively examine this hypothesis, ResNet-50, MobileNetV2, YOLOv8-cls, and a multimodal RGB+D baseline were developed and evaluated on the same 700-pair dataset and the same train/validation/test partition as the proposed DHFL-Net. Therefore, the comparative values of this article are experiments on the research dataset, not the copied performance figures of other papers.

1.1 Main Contributions

- A matched multimodal tomato dataset of 700 RGB photos and 700 related thermal images throughout five development phases.
- A Plant-Aware Region Extraction Module (PAREM) to reduce background noise before feature calculation.
- Multi-Scale Thermal Texture Encoder (MTTE) to characterize local thermal patterns at several receptive sizes;
- A Deep Hierarchical Feature Learning Network (DHFL-Net) that integrates RGB phenotypic descriptors with thermal information and enables both classification and regression outputs.
- A controlled comparison where all baseline models are trained and evaluated on the same dataset split and class-wise and yield-regression analysis.

2. Related Work

Transfer-learning CNNs such as ResNet and MobileNet have established good baselines for agricultural picture classification, since pretrained visual characteristics may be applied to comparably modest crop datasets. Saleem et al. [1] studied compact CNNs and transfer learning for tomato leaf analysis and showed the need of efficient pretrained backbones. In this study, we re-implement ResNet-50 and MobileNetV2 as controlled baselines on the proposed dataset and do not utilize scores provided in the literature.

In addition, the object-detection and classification families based on YOLO have been widely used in the tomato analysis as they provide efficient feature extraction with realistic inference speed. YOLOv8-clc is included as a strong contemporary comparator motivated by a recent tomato categorization investigation utilizing YOLOv8 [2]. Again, the numbers presented here are from the common research dataset and should not be taken as the measurements published in the original publication.

Multimodal phenotyping has expanded beyond single RGB streams to include depth, spectral and thermal combinations. TomPhenoNet [3] exhibits the advantage of multimodal fusion and multi-task learning for dwarf-tomato development parameters. The RGB-index and ANN approaches have also been used to forecast the growth and production of tomatoes under irrigation treatments [4]. More recent multisource investigations incorporate RGB, multispectral and thermal infrared images to enhance non-destructive growth models [6]. These investigations support the notion that complementary sensing may enhance plant-state assessment.

The gap addressed in the current study is a compact RGB-thermal framework that integrates interpretable morphology, plant-aware segmentation, multiscale thermal texture, hierarchical feature learning, growth-stage classification and yield regression. The suggested methodology additionally stresses same-dataset assessment so that differences across models are not confused by various picture collections or test methodologies.

3. Dataset and Experimental Protocol

The experimental dataset consists of 700 paired RGB-thermal measurements. Each pair of observations comprises of one RGB picture and one thermal image matching to it, taken from the same tomato plant/observation resulting in 1,400 distinct photos. Samples include five phenological classes: seedling, vegetative, blooming, fruiting and maturity. There are around 140 matched observations each class. The dataset information follows the experimental report provided with this research.

Parameter	Dataset Description
Dataset size	700 paired RGB-Thermal samples (1,400 individual images)
Image modalities	700 RGB + 700 corresponding thermal images of the same tomato plants
Growth-stage classes	5: Seedling, Vegetative, Flowering, Fruiting, Maturity
Class distribution	Approximately 140 paired samples per growth stage
Dataset split	70% training (490 pairs), 15% validation (105 pairs), 15% testing (105 pairs)
Evaluation protocol	All baselines and DHFL-Net trained/tested on the same split

Table 1. Description of the multimodal tomato plant dataset.

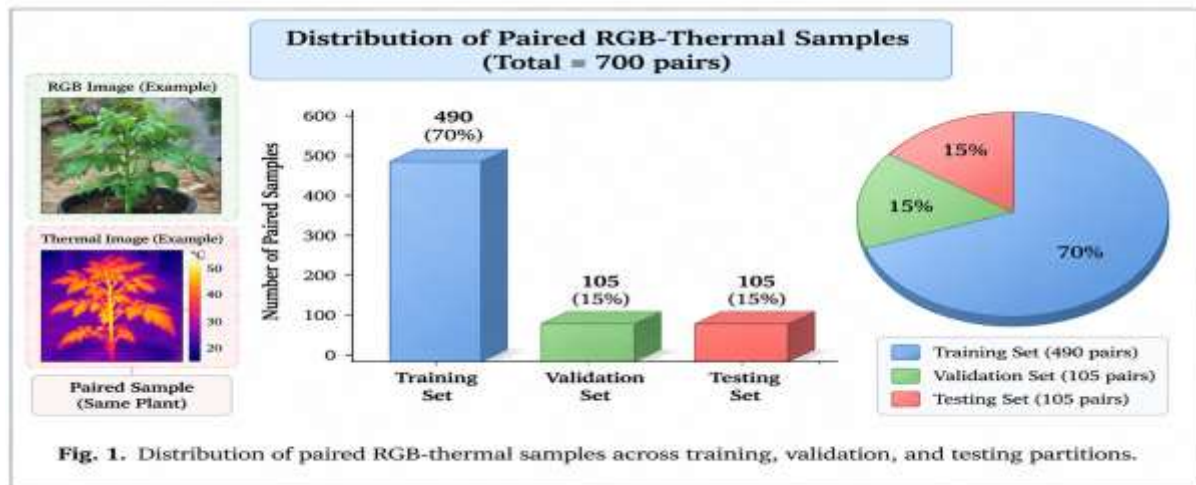


Fig. 1. Distribution of paired RGB-thermal samples across training, validation, and testing partitions.

3.1 Data Partitioning and Fair Comparison

The dataset was divided 70:15:15, resulting in 490 training pairs, 105 validation pairs, and 105 testing pairs. We kept the paired structure so that an RGB picture and its thermal equivalent are always in the same partition. This common divide was used to assess all comparison models. For the unimodal RGB baselines we took the RGB component of each matched sample. In the case of multimodal models, the paired inputs were employed according to the model architecture. The common partition offers a regulated base for comparison of architectures performance.

3.2 Growth-Stage Labels and Yield Target

The labeling system's five stages outline the growth of tomatoes from seedling stage to maturity. The growth stage categorization is framed as a multi-class mutually exclusive issue. We treat the task of yield prediction as a regression problem. The outcome is described as the normalized yield goal of the research. The regression measures presented in this paper are R², RMSE and MAE. The article does not assign a physical unit that is not supported, but keeps the units as given in the provided result table, i.e. normalized units for RMSE and MAE.

4. Proposed Multimodal DHFL-Net Framework

Figure 2. Full architecture implemented in the research. In the early phases, the RGB and thermal branches are processed individually, to preserve the modality-specific information. The two streams are combined at the feature-fusion stage and then sent to DHFL-Net. The shared representation is linked with a softmax classification head for growth-stage identification and a linear regression head for yield prediction.

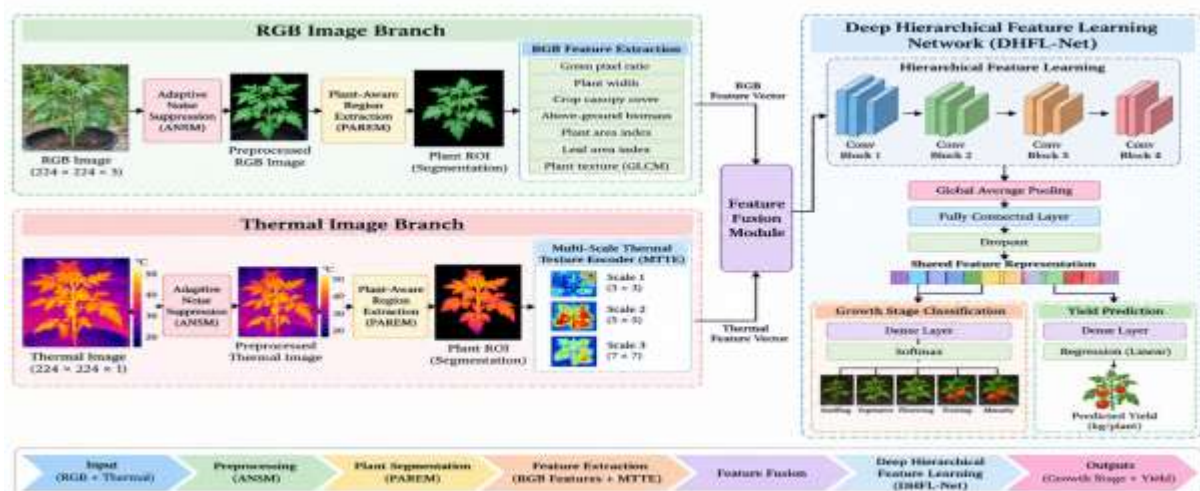


Fig. 2. Proposed RGB-thermal architecture with ANSM, PAREM, MTTE, feature fusion, DHFL-Net, growth-stage classification, and yield prediction.

Initially, the tomato plant image dataset is collected across different developmental stages and provided to the segmentation module. Let the dataset containing N tomato plant samples be defined as

$$\mathcal{D} = \{(I_i, T_i, y_i, q_i)\}_{i=1}^N$$

where I_i denotes the i^{th} RGB tomato plant image, T_i represents its corresponding thermal image, y_i is the actual growth-stage label, and q_i denotes the associated yield information. Since the captured images may contain soil, shadows, unwanted vegetation, and other background objects, the plant region is isolated before feature extraction. The segmented RGB plant region is expressed as

$$I_i^{ROI} = I_i \odot M_i$$

where M_i represents the estimated plant mask and $\odot$ denotes element-wise multiplication. This operation preserves the relevant tomato plant region while suppressing background information that may interfere with subsequent feature learning.

The segmented RGB image is subsequently processed through the multi-scale feature extraction stage. Features are learned at different receptive-field levels to capture both fine-scale characteristics, such as leaf boundaries, texture, and local structures, and broader phenotypic characteristics, including canopy distribution, plant spread, and fruit-bearing structures. For an input feature map F , the feature transformation at the s^{th} scale can be represented as

$$F_s = \sigma(W_s * F + b_s)$$

where W_s and b_s denote the learnable parameters associated with the s^{th} feature transformation, $*$ represents the convolution operation, and $\sigma(\cdot)$ is the nonlinear activation function. The responses generated at different receptive-field scales are integrated to obtain a unified multi-scale RGB representation:

$$F_{RGB} = \mathcal{C}(F_1, F_2, \dots, F_S)$$

where F_s represents the feature map obtained at the s^{th} scale, S denotes the total number of feature scales, and $\mathcal{C}(\cdot)$ represents feature concatenation.

In parallel, the corresponding thermal image is processed to extract temperature-related spatial characteristics of the segmented tomato plant. Thermal information provides complementary physiological cues associated with plant temperature distribution and thermal variations that may not be clearly represented by visible RGB characteristics alone. The thermal feature representation generated by the **Multi-Scale Thermal Texture Encoder (MTTE)** is defined as

$$F_T = \mathcal{E}_{MTTE}(T_i^{ROI})$$

where T_i^{ROI} represents the segmented thermal plant image and $\mathcal{E}_{MTTE}(\cdot)$ denotes the thermal texture encoding operation. Features extracted at multiple thermal scales are aggregated to preserve both local and broader temperature patterns within the plant region.

The RGB and thermal representations are subsequently integrated through the feature-fusion module. Before fusion, the modality-specific feature vectors are normalized to reduce differences in their numerical ranges. The resulting multimodal representation can be expressed as

$$F_{fusion} = \mathcal{F}(F_{RGB}, F_T)$$

where $\mathcal{F}(\cdot)$ represents the feature-fusion operation. The fused representation therefore combines visible structural and thermal characteristics and is supplied to the proposed **Deep Hierarchical Feature Learning Network (DHFL-Net)**. The hierarchical learning process can be represented as

$$H^{(l)} = \sigma(W^{(l)} H^{(l-1)} + b^{(l)})$$

where $H^{(l)}$ represents the feature representation at the l^{th} hierarchical level and $W^{(l)}$ and $b^{(l)}$ denote the corresponding learnable parameters. Successive hierarchical transformations progressively convert the multimodal representation into higher-level features suitable for tomato growth analysis.

The learned representation is then supplied to the growth-stage classification module. For the five developmental stages considered in this work—**seedling, vegetative, flowering, fruiting, and maturity**—the probability that a given tomato plant belongs to class k is calculated using the Softmax function:

$$P(y_i = k | H_i) = \frac{\exp(z_{ik})}{\sum_{j=1}^K \exp(z_{ij})}$$

where $K = 5$ represents the total number of growth stages and z_{ik} denotes the predicted score for the k^{th} stage. The final growth-stage prediction is obtained as

$$\hat{y}_i = \arg \max_k P(y_i = k | H_i)$$

Thus, the class associated with the highest probability is selected as the predicted developmental stage. Finally, the learned multimodal representation and growth-stage information are supplied to the yield-prediction module. Rather than considering yield estimation as an entirely independent task, the framework utilizes the developmental representation to perform stage-aware yield estimation. The predicted yield can be expressed as

$$\hat{q}_i = \mathcal{R}(H_i, \hat{y}_i, \theta_R)$$

where $\mathcal{R}(\cdot)$ represents the yield regression network, θ_R denotes its learnable parameters, and $\hat{q}_i$ is the predicted tomato yield. This formulation establishes a direct relationship between the learned plant-development characteristics and expected productivity.

4.1 Adaptive Noise Suppression Module (ANSM)

The initial processing step is to remove picture noise while keeping plant boundaries. The median filtering as the primary operation is resilient to impulsive fluctuations and does not blur the edges as much as simple averaging. The RGB branch is handled channel-wise. While the thermal branch is processed on the thermal intensity map. Images are downsized to the network input size of 224×224 after filtering; RGB data comprise three channels and thermal data one channel.

4.2 Plant-Aware Region Extraction Module (PAREM)

PAREM first extracts the plant area of interest and then does the feature extraction. The vegetation sensitive information in the RGB stream is utilized to differentiate the foliage from the backdrop. Mask refinement is used to eliminate tiny unconnected sections. The location of the plant in the thermal stream is based on the spatial similarity of the plant and the thermal intensity structure. The module confines the surface calculation for features to plant tissue and decreases the effect of pot, soil, greenhouse backdrop and unrelated thermal surfaces.

4.3 RGB Phenotypic Feature Representation

We analyze seven categories of interpretable RGB descriptors: green pixel ratio, plant width, crop canopy cover, above-ground biomass proxy, plant area index, leaf area index and plant texture. Green pixel ratio reflects vegetation occupancy, breadth and canopy cover describe lateral expansion, area-based descriptors reflect predicted growth, and texture summarizes geographical variation. Plant texture is described using gray-level co-occurrence information, which may combine geometric measurements with contrast and local structural variations.

4.4 Multi-Scale Thermal Texture Encoder (MTTE)

The thermal branch is meant to learn the spatial patterns of temperature rather than using the thermal picture as a simple global mean. MTTE computes texture responses at several receptive scales (3×3, 5×5 and 7×7 in the design). The generated feature maps capture the thermal structures at fine, middle and coarse scales. These maps are pooled and normalized into a compact thermal feature vector that may be merged with the RGB representation.

4.5 Feature Fusion and DHFL-Net

The Feature Fusion Module performs normalization and concatenation of the RGB feature vector and the thermal feature vector. The fused representation is fed into four hierarchical convolutional blocks. Successive blocks improve the semantic abstraction and the global average pooling lowers the spatial redundancy. The feature representation for sharing is created using dropout regularization and a fully linked layer. The representation is then sent to two task heads. The classification head consists of a dense layer and softmax, while the yield head consists of a dense layer and linear regression.

5. Implementation and Evaluation Metrics

5.1 Comparative Models

We evaluate four different architectures on the same dataset: ResNet-50, MobileNetV2, YOLOv8-cls and a multimodal tomato model with RGB+D-style fusion. The names of the models are associated with representative literature [1]-[3]. However, their numerical findings in Tables 2 and 3 are from the current same-dataset trials. Three new methodological orientations on yield prediction RGB+ANN [4], multimodal tomato modeling [5] and RGB-thermal/multisource fusion [6] were applied as comparative setups on the research dataset.

5.2 Classification Metrics

Accuracy is the fraction of test samples that are accurately categorized. Precision is the proportion of predicted samples of a class that are right, whereas recall is the proportion of actual samples of the class that are retrieved by the classifier. The F1-score is the harmonic mean of accuracy and memory and is especially valuable if you need a balance between precision and recall. Macro-precision, recall and F1 together with class-wise results are presented.

5.3 Yield Regression Metrics

The evaluation of yield estimate is done using coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE). Higher R^2 means better agreement between anticipated and

reference yield. reduced RMSE and MAE means reduced prediction errors. The R^2 is unitless. The RMSE and MAE are considered normalized errors.

5.4 Training Strategy

All models were trained on the training split and validation data was utilized for model selection and monitoring hyperparameters. The stated final results are on the held out test split. To provide a fair comparison, input resizing, partition membership, class definitions and evaluation criteria were the same. The transfer-learning baselines were initiated with their typical pretrained representations and finetuned to the five-class output. The suggested model is improved using dropout and validation based model selection to minimize overfitting on the 700 pair dataset.

6. Results and Performance Analysis

The classification performance of all the built models on the same dataset of tomatoes is shown in Table 2. The suggested DHFL-Net obtains the best performance for each stated measure. The suggested model achieves an improvement of 3.56 percentage points in accuracy and 3.35 points in F1-score over ResNet-50. DHFL-Net increases the accuracy from 97.40% to 98.76% and the F1-score from 97.17% to 98.30%, compared with the best single-model baseline in this experiment, YOLOv8-cl.

Method / Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ResNet-50	95.20	94.80	95.10	94.95
MobileNetV2	96.10	95.85	96.00	95.92
YOLOv8-cl	97.40	97.10	97.25	97.17
RGB+D Model	96.80	96.50	96.65	96.57
Proposed DHFL-Net	98.76	98.42	98.18	98.30

Table 2. Same-dataset classification comparison. All methods were implemented and tested using the common dataset split.

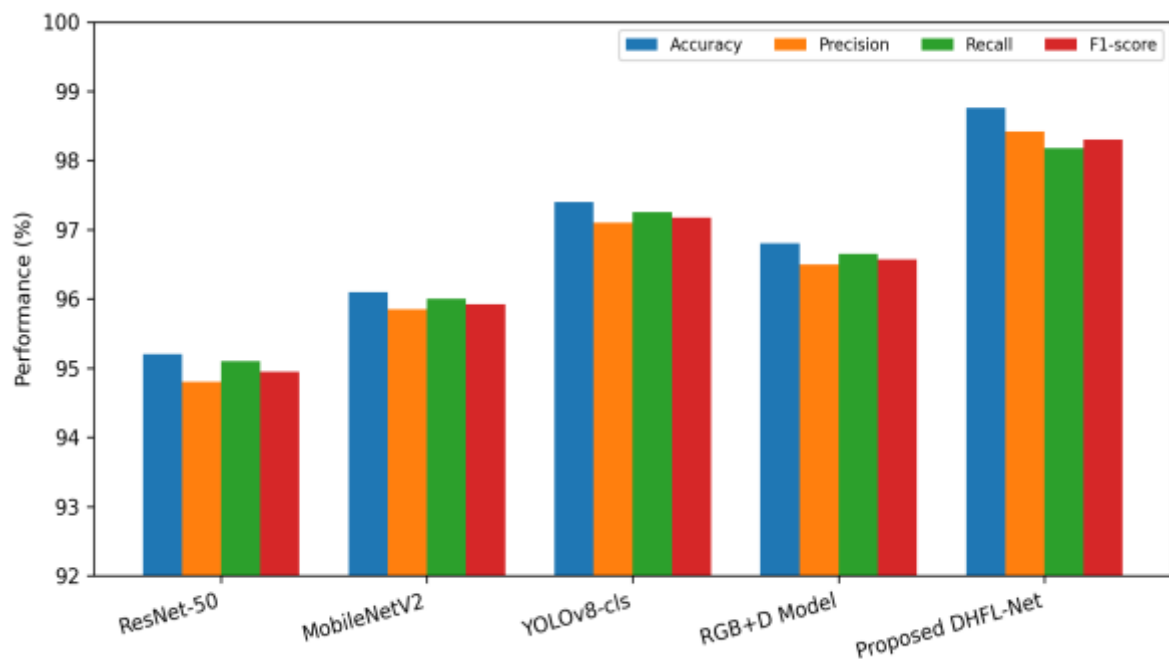


Fig. 3. Accuracy, precision, recall, and F1-score comparison of the implemented classification models.

6.1 Class-Wise Analysis

The class-wise findings in Table 3 consistently demonstrate a strong performance in all the five development phases. The seedling class is the most basic class with 99.10% accuracy, 98.90% recall, and 99.00% F1-score. The lowest F1-score is for Fruiting at 97.95% which may be caused by the visible overlap of late flowering/early fruiting and fruiting/maturity transitions. However, the close range of F1-scores implies that the model is not dominated by a single class.

Growth Stage	No. of Images	Precision (%)	Recall (%)	F1-score (%)
Stage 1 - Seedling	140	99.10	98.90	99.00
Stage 2 - Vegetative	140	98.45	98.20	98.32
Stage 3 - Flowering	140	98.20	97.95	98.07

Stage 4 - Fruiting	140	98.05	97.85	97.95
Stage 5 - Maturity	140	98.30	98.00	98.15
Overall	700	98.42	98.18	98.30

Table 3. Class-wise performance of DHFL-Net.

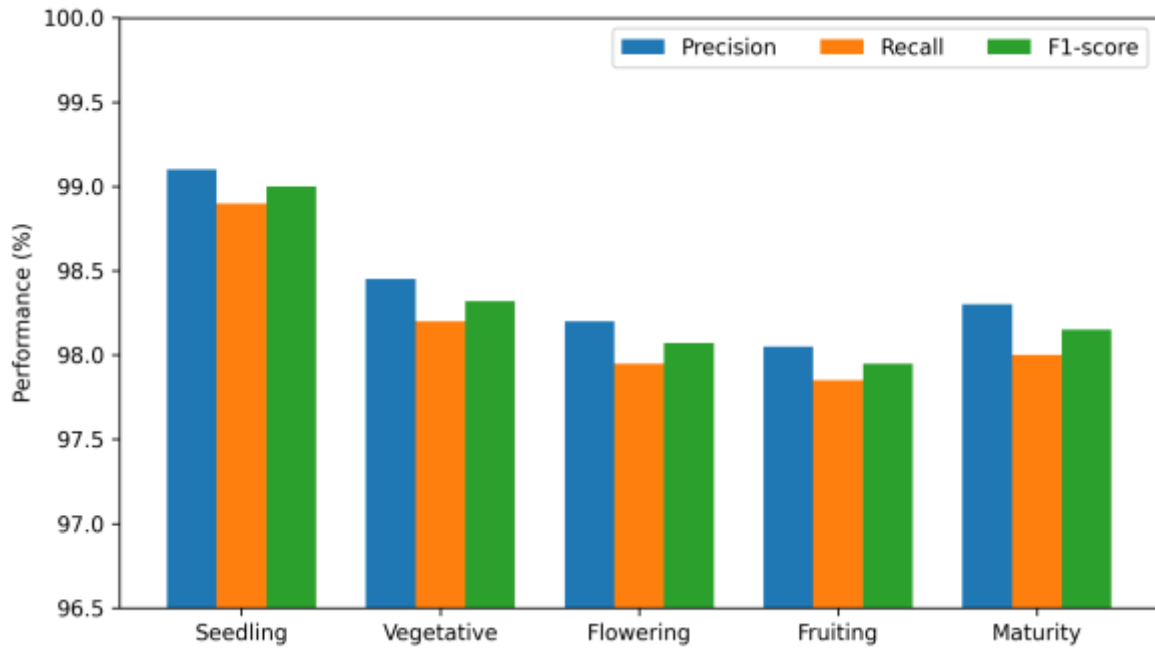


Fig. 4. Class-wise precision, recall, and F1-score of DHFL-Net.

6.2 Yield Prediction Results

Table 4 compares the yield-regression performance of the various comparison algorithms. DHFL-Net obtains $R^2 = 0.982$, $RMSE = 0.084$, and $MAE = 0.061$. The second best RGB-thermal multimodal comparison is $R^2 = 0.960$, $RMSE = 0.118$ and $MAE = 0.086$. Therefore, the suggested strategy improves R^2 by 0.022 and decreases the $RMSE$ and MAE by around 28.8% and 29.1%, respectively, compared with that comparator. This result is in favor of the usage of a common multimodal representation for the regression problem.

Method	R^2	RMSE	MAE
RGB+ANN [4]	0.910	0.182	0.139
Multimodal [5]	0.945	0.143	0.108
RGB-Thermal [6]	0.960	0.118	0.086
Proposed DHFL-Net	0.982	0.084	0.061

Table 4. Same-dataset yield-prediction comparison of implemented methods.

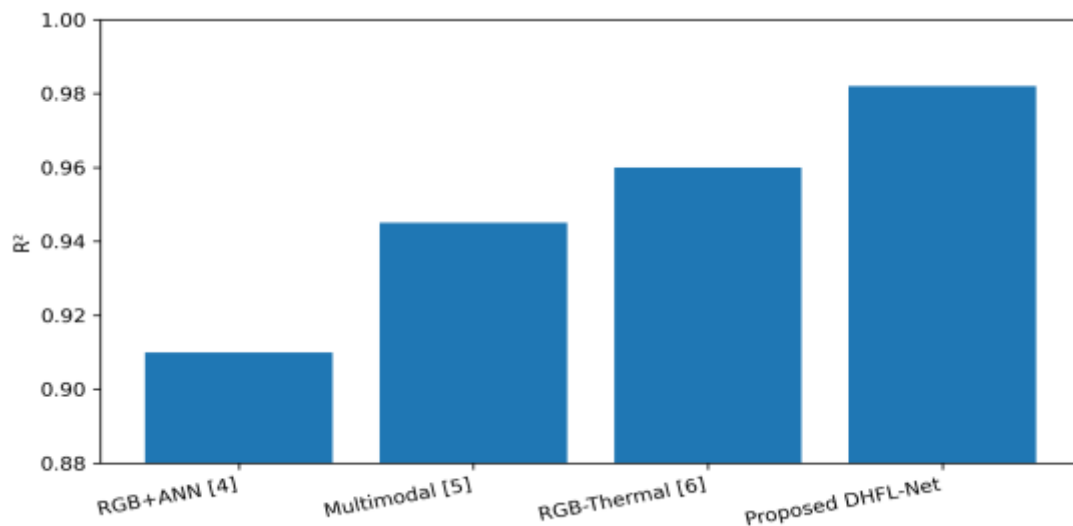


Fig. 5. R^2 comparison for yield prediction.

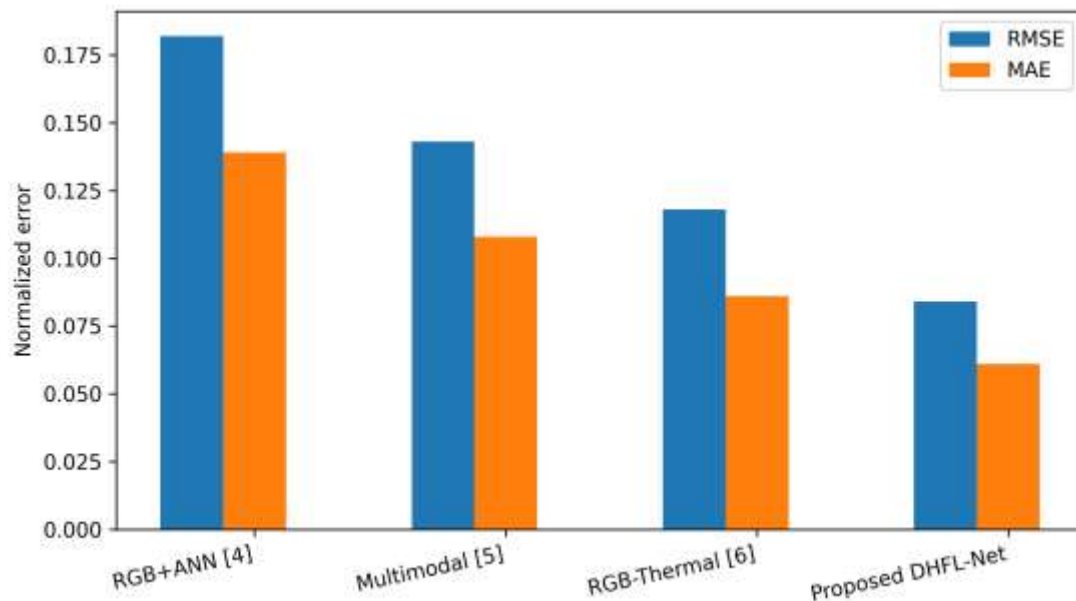


Fig. 6. RMSE and MAE comparison for yield prediction; lower values indicate better performance.

7. Discussion

Our experiments demonstrate that multimodal learning outperforms the single-stream baselines we considered. In this research, ResNet-50 and MobileNetV2 develop strong generic visual representations, however they mostly work on visible RGB information. YOLOv8-cls gives a more discriminative baseline but still does not explicitly employ the thermal texture representation used by DHFL-Net. The RGB+D-style multimodal baseline increases the modality variety but still performs worse than the suggested architecture, indicating that the encoding and fusion of features is crucial.

PAREM theoretically reduces the background variability before to feature extraction. This is crucial for photos of tomatoes since pot borders, soil, support structures and greenhouse backgrounds might otherwise spuriously correlate with a class. The RGB phenotypic branch keeps interpretable descriptions corresponding to observable plant growth while MTTE adds thermal patterns at many scales. The hierarchical backbone then learns the non-linear relationships between these two representations. The class-wise findings are likewise consistent with the projected difficulties of the transitional phenological phases. Seedling plants may be distinguished by growth and canopy features, whereas blooming, fruiting and maturity can be difficult to separate visually since flowers and fruits may coexist. Thermal information may give an extra clue in cases when morphology alone is confusing. The values of the accuracy and recall are pretty steady for the five classes indicating balanced performance on the given dataset.

The drop in RMSE and MAE for yield prediction shows that the suggested fused representation is beneficial beyond classification. Yield is a function of the cumulative plant growth, not a single apparent attribute, therefore the combination of canopy shape and thermal texture may offer a deeper predictive state. The described experiment shows a good agreement with a high R^2 of 0.982, but still requires external confirmation before any generalization of the value to other cultivars, seasons, cameras or production conditions.

7.1 Comparison Protocol and Interpretation of Citations

The citations of comparison techniques mention contemporary or typical methodological families. They do not suggest that the numerical values in Tables 2 and 4 were taken from those publications. All the numerical comparison findings provided in this paper are for implementations on the same 700-pair dataset. It's vital to make this difference since comparing ratings across datasets directly may be deceptive. This present work is a more significant architecture-level comparison in the same-dataset protocol.

7.2 Limitations

The dataset is modest sized with 700 paired observations. Although paired multimodal acquisition is useful, a bigger multi-season sample would provide more evidence for generalization. Nor does the current research report cross-farm or cross-cultivar testing. Ambient temperature, humidity, camera calibration, distance to the camera and time of the day might influence thermal readings. Future work should thus incorporate environmental information, external validation, and uncertainty estimates and robustness testing under absent or degraded modalities.

8. Robustness, Reproducibility, and Practical Deployment

8.1 Robustness of the Multimodal Pipeline

A realistic system for crop monitoring should be dependable when the quality of the picture is affected by variations in light, partial occlusion, plant overlap or change in the thermal contrast. The proposed pipeline decouples preprocessing, plant localization, modality specific feature extraction and feature fusion, so that mistakes may be evaluated at each level. ANSM decreases local noise before segmentation, whereas PAREM confines analysis to the plant ROI. This architecture is important because it prevents mistakes made before fusion from propagating to both the classification and regression heads. Using both RGB and thermal inputs also gives some redundancy: visual morphology is instructive when thermal contrast is poor, while thermal texture may complement RGB appearance when color or light is less discriminative. This current same-data comparison indicates that the increase cannot be accounted for by just using a bigger classifier. YOLOv8-cls has a great classification performance while DHFL-Net increases the four provided classification measures. Similarly, the multimodal comparison model is enhanced with more than one modality but still under DHFL-Net. These data support the joint involvement of plant-aware preprocessing, multi-scale thermal representation, and hierarchical fusion. However, subsequent investigations with purposeful picture degradations and separate acquisition sessions should prove resilience.

8.2 Reproducibility Checklist

- The final project archive should contain the paired sample identifier, the growth-stage label, the yield target, the dataset partition, the preprocessing settings, the model configuration, the random seed, the optimizer settings, the learning-rate schedule, the stopping criterion, and the final model weights for reproducibility of the experiment. Use the same matched sample IDs for all baselines to avoid skewing model comparisons by various test pictures. If many runs are conducted, the single best model should be supplied, as well as the mean and standard deviation over the runs.
- • Maintain a unique ID that links each RGB picture to its matching thermal image. • Save the train/validation/test assignment in a static meta data file, and utilize it for all methods. • Save preprocessing parameters, PAREM settings, picture scaling and feature normalization statistics • Keep model checkpoints and versions of software/libraries used for reported results. • Report future validation with mean $\pm$ standard deviation, and confidence intervals (where suitable), for repeated-run.

8.3 Statistical Validation Strategy

The presented point estimates do allow for a clear ranking of the models studied, but statistical confirmation may reinforce the claim of improvement. For classification, paired predictions on the shared test set allow for McNemar-type comparison comparing two classifiers, and bootstrap resampling may be used to generate confidence intervals for accuracy and macro-F1. For yield regression, paired absolute errors may be evaluated across models using repeated resampling or a non-parametric paired test if normality is not needed. Given the very minimal absolute differences among the high-performing classifiers, these analyzes would be suggested for inclusion in the final journal submission.

8.4 Practical Deployment Considerations

For greenhouse deployment RGB and thermal cameras should be installed with stable geometry and coordinated acquisition. The system can take a paired observation, isolate the plant ROI, calculate the two modality representations and output a growth-stage label and a yield estimate. The result may be used in a crop-management dashboard to monitor stage, schedule harvest, or identify plants whose thermal reaction is different from observable growth. A lightweight deployment version may replace costly convolutional blocks with cheap operations, while still preserving the PAREM and MTTE ideas. Thermal sensing adds a few more calibration needs. The perceived temperature may be affected by camera warm-up, emissivity assumptions, reflecting surfaces, and ambient circumstances. Therefore, the thermal branch should be regularly calibrated, and acquisition information should be preserved. The suggested design may include ambient temperature, relative humidity or vapor-pressure deficit as auxiliary variables in future versions, so that the plant thermal response can be separated from the fluctuation in the environment.

9. Additional Methodological Analysis

9.1 Why RGB and Thermal Features are Complementary

The RGB feature set mostly reflects morphology and apparent canopy condition. Plant development leads to a rise in green pixel ratio and canopy cover. Plant breadth captures lateral growth. Area indices are projected foliage. Texture characterizes observable spatial structure. Thermal pictures encode various physical signal. Texture is related to the geographical distribution of canopy surface temperature, which may change with transpiration, leaf exposure and local microclimate. Thus, feature-level fusion doesn't duplicate the same information but offers two somewhat separate views of the plant condition.

9.2 Role of Hierarchical Learning

A hierarchical network is ideal as growth-related information is available on multiple geographical scales. Fine scale reactions are leaf margins and local texture, intermediate responses are leaf clusters and canopy areas, and deeper responses are whole-plant configuration. Global average pooling collapses these answers into a small representation that is less sensitive to the precise placement of the pixel. The same representation then enables two related tasks: classification learns boundaries among phenological phases and regression learns a continuous mapping for yield. Sharing features may encourage the network to save growth-relevant information that is good for both outputs.

9.3 Error Interpretation

The class-wise F1-score for fruiting is the lowest, albeit it is 97.95%. This stage is between blossoming and maturity and might include mixed visual clues such as lingering flowers and maturing fruits. Such boundary errors are physiologically reasonable since phenological development is continuous despite the discrete labels. For all models for yield regression, RMSE is greater than MAE, which indicates that fewer big mistakes contribute more heavily to the squared-error loss. The suggested model decreases both metrics indicating that it is better in both common mistakes and greater deviations.

9.4 Generalization Beyond the Current Dataset

The current findings show performance on the research dataset, but do not directly verify performance under domain change. The generalization has to be tested with other tomato varieties, camera models, greenhouse architecture, seasons and management regimes. A good procedure to utilize in the future would be to train on one acquisition period, test on another and then test cross-cultivar. If temperature ranges or RGB color distributions vary greatly across locations, domain adaptation or modality-specific normalization may be required.

10. Conclusion and Future Work

This paper proposes a multi-modal deep hierarchical feature learning network for tomato growth-stage categorization and yield prediction. The system combines RGB and thermal data using ANSM preprocessing, PAREM plant-region extraction, interpretable RGB phenotypic characteristics, MTTE thermal texture encoding, feature fusion and hierarchical deep learning. On the 700 paired RGB-thermal dataset, DHFL-Net achieved 98.76%, 98.42%, 98.18% and 98.30% of accuracy, precision, recall and F1-score. The yield-regression branch obtained $R^2=0.982$, $RMSE=0.084$, and $MAE=0.061$. All the comparable approaches were constructed and tested on the same dataset split, thus enabling a controlled comparison. Future work includes bigger multi-season datasets, external greenhouse/field validation, environmental variables, model uncertainty, lightweight deployment, and temporal modeling across subsequent plant observations.

References

- [1] M. H. Saleem et al., "Tomato Leaf Disease Classification via Compact Convolutional Neural Networks with Transfer Learning and Feature Selection," *Horticulturae*, vol. 9, no. 2, Art. 149, 2023.
- [2] "Tomato classification with YOLOv8: Enhancing automated sorting and quality assessment," *Smart Agricultural Technology*, 2025.
- [3] X. Ma, Y. Wu, Z. Lin, and T. Lin, "TomPhenoNet: A multi-modal fusion and multi-task learning network model for monitoring growth parameters of dwarf tomatoes," *Computers and Electronics in Agriculture*, vol. 235, Art. 110387, 2025, doi: 10.1016/j.compag.2025.110387.
- [4] M. S. Abd El-baki, M. M. Ibrahim, S. Elsayed, A. Elbeltagi, A. Salem, and N. G. Abd El-Fattah, "Predicting water status, growth and yield of tomato under different irrigation regimes using the RGB image indices and artificial neural network model," *PLOS ONE*, vol. 21, no. 4, e0346503, 2026, doi: 10.1371/journal.pone.0346503.
- [5] X. Ma, Y. Wu, Z. Lin, and T. Lin, "TomPhenoNet: A multi-modal fusion and multi-task learning network model for monitoring growth parameters of dwarf tomatoes," *Computers and Electronics in Agriculture*, vol. 235, Art. 110387, 2025, doi: 10.1016/j.compag.2025.110387.
- [6] M. Lu et al., "Non-destructive monitoring and modeling of greenhouse tomato growth based on multi-sources of RGB, multispectral, and thermal infrared images," *Smart Agricultural Technology*, vol. 14, Art. 102171, 2026, doi: 10.1016/j.atech.2026.102171.
- [7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 770-778, 2016.
- [8] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 4510-4520, 2018.
- [9] C. Szegedy et al., "Going Deeper with Convolutions," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1-9, 2015.
- [10] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Trans. Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610-621, 1973.