OPEN ACCESS Plant Protection, 10 (01) 2026. 125-133 DOI:10.33804/pp.010.01.6011

Check for updates

Available Online at EScience Press

Plant Protection

ISSN: 2617-1287 (Online), 2617-1279 (Print)

http://esciencepress.net/journals/PP

Research Article

Chlorophyll Attention Network Using MobileViTv2 for On-Device Plant Health Monitoring

[a,b]Rabia Zafar, [a]Muhammad Shahid Farid, [a]Muhammad Hassan Khan, [c]Rashid Mahmood

[a] *Department of Computer Science, University of the Punjab, Lahore, Pakistan.*
[b] *University of Engineering and Technology, Lahore, Narowal Campus, Pakistan.*
[c] *Department of Soil Science, University of the Punjab, Lahore, Pakistan.*

ARTICLE INFO

Article history
Received: 27th November, 2025
Revised: 24th February, 2026
Accepted: 27th February, 2026

Keywords
Vision Transformer
Precision agriculture
Mobile deployment
Deep learning
Non-destructive sensing

ABSTRACT

Accurate assessment of leaf chlorophyll content is critical for precision nitrogen management in maize (*Zea mays* L.), yet conventional SPAD meters are labor-intensive and limited in scalability. To address this, a regression framework integrating MobileViT, Chlorophyll Attention (CA), and a multi-modal fusion mechanism for rapid, image-based SPAD value prediction was proposed. An ablation study quantified the contribution of each architectural component. The full model achieved a baseline coefficient of determination (R^2) of 0.6832, confirming robust predictive capability. Removal of the MobileViT block caused the largest performance decline (R^2 = 0.6564), emphasizing its critical role in feature representation. Excluding the CA module (R^2 = 0.6721) or the multi-modal fusion mechanism (R^2 = 0.6368) further highlighted the necessity of synergistic component integration for optimal performance. Using 5-fold cross-validation, the base model attained a mean R^2 of 0.7099, which was further enhanced to 0.7345 through ensemble learning and four-way test-time augmentation. The improved model achieved a Root Mean Square Error (RMSE) of 4.264 SPAD units and Mean Absolute Error (MAE) of 3.649 SPAD units. Comparative evaluation demonstrated that the proposed approach outperformed state-of-the-art architectures, including EfficientNet-B3 (R^2 = 0.7078) and deeper ResNet variants, highlighting its robustness and generalization capability. The model's rapid inference ensures suitability for real-time, edge-device deployment in field conditions. The study confirms that each component enhances predictive accuracy, and that ensemble learning with TTA further improves reliability, providing a scalable, high-throughput solution for efficient, data-driven chlorophyll assessment.

Corresponding Author: Rabia Zafar
Email: rabiazafar600@gmail.com *ORCID ID: https://orcid.org/0000-0002-4231-410X*

Introduction

Nitrogen concentration is a fundamental physiological indicator strongly correlated with plant vigor, chlorophyll content, and photosynthetic efficiency. Accurate assessment of plant nitrogen status is therefore essential for informed decision-making in precision agriculture and sustainable crop management (Berger et al., 2020; Zhao et al., 2022; Saudy and Mohamed El-Metwally, 2023; Li et al., 2024a). Conventional nitrogen estimation methods, including destructive chemical analysis and non-destructive handheld meters (e.g., SPAD meters), provide reliable measurements. However, these approaches are

labor-intensive, time-consuming, limited to discrete point sampling, and often fail to capture spatial heterogeneity across the leaf surface or crop canopy.

Early non-contact estimation techniques employed machine learning (ML) algorithms such as support vector machines (SVM) and random forests. Although these models demonstrated promising predictive capability, their performance was constrained by reliance on manually engineered features. Consequently, chlorophyll and SPAD value prediction has progressively shifted from traditional ML approaches toward deep learning and hyperspectral imaging techniques (Li et al., 2024b). Deep learning models automate feature extraction and frequently achieve superior predictive accuracy (Li et al., 2024a; Wang et al., 2025). Nevertheless, hyperspectral systems remain costly, computationally demanding, and require specialized expertise, limiting their large-scale field deployment.

Recent advances in computer vision and deep learning provide a compelling alternative for plant phenotyping. Convolutional neural networks (CNNs) and, more recently, vision transformers (ViTs) have significantly enhanced automated plant disease diagnosis and trait analysis (Wang et al., 2024). Vision transformers have emerged as a powerful complement to CNNs in agricultural computing, excelling in modeling long-range dependencies and global contextual relationships through multi-head attention mechanisms (Borhani et al., 2022; Khan et al., 2022). Furthermore, lightweight and mobile-optimized variants such as MobileViT and EdgeViTs improve computational efficiency while maintaining competitive performance (Mehta and Rastegari, 2021; Li et al., 2023).

The integration of attention mechanisms enables models to dynamically prioritize diagnostically relevant regions while suppressing background noise, thereby improving interpretability, robustness, and generalization. Hybrid CNN-Transformer architectures that incorporate domain-specific attention modules represent the current frontier in intelligent plant phenotyping, aiming to deliver accurate, computationally efficient, and explainable solutions. Image-based approaches offer several advantages: they are non-contact, support rapid high-throughput analysis, allow spatial mapping of chlorophyll distribution across entire leaves, and are suitable for deployment on mobile and edge devices for real-time field applications.

To address key challenges in image-based SPAD prediction, this study proposes a novel hybrid architecture termed the Mobile Chlorophyll Attention Network (MoCA-Net), designed to achieve an optimal balance between computational efficiency and regression accuracy.

The major contributions of this work are as follows:

1. Chlorophyll Attention (CA) Module: A specialized domain-specific attention mechanism that extends beyond generic attention strategies to enhance chlorophyll-sensitive feature learning.
2. Hierarchical Multi-Scale Feature Fusion: A structured framework for integrating features across multiple spatial scales to improve robustness and predictive precision of SPAD values.
3. Domain-Tailored Training Framework: A comprehensive training paradigm incorporating advanced data augmentation and regularization techniques to mitigate overfitting and address the limitations of small agricultural datasets.

Collectively, these innovations bridge the gap between controlled laboratory research and practical field deployment, supporting both high-precision and resource-efficient agricultural systems.

Materials and Methods

Dataset acquisition and preprocessing

A dataset comprising 795 maize leaf images, captured using smartphone cameras and paired with corresponding SPAD meter readings, was utilized in this study. The images were collected under natural field conditions at the Maize & Millet Research Institute, Sahiwal, Pakistan. Each image was associated with a ground-truth chlorophyll measurement obtained using a SPAD meter, enabling supervised regression modeling.

All images were resized to a standardized pixel resolution to ensure uniformity and computational efficiency during model training. To eliminate background interference and enhance feature extraction, a U-Net-based semantic segmentation model was employed to isolate maize leaves from complex field backgrounds. This process effectively removed noise arising from soil, hands, shadows, and other extraneous objects, thereby improving the signal-to-noise ratio in the input data as shown in Figure 1.

To enhance model robustness and generalization across diverse environmental conditions, an extensive data augmentation pipeline was implemented. The augmentation techniques included random cropping, horizontal and vertical flipping, rotation, and color jittering. These transformations artificially increased dataset variability, enabling the model to better adapt to variations in illumination, orientation, and field heterogeneity.

Cross-validation strategy

The experimental design incorporated stratified five-fold cross-validation to ensure robust performance estimation and to minimize bias associated with specific training-validation splits. This approach partitions the dataset into five approximately equal-sized folds while preserving the distribution of SPAD values across all folds. Maintaining this distribution ensures that each fold represents the full range of chlorophyll variability present in the dataset.

In each iteration, four folds (80% of the data) were used for model training, while the remaining fold (20%) served as the validation set. This process was repeated five times, with each fold serving exactly once as the validation set. Consequently, five independent performance estimates were obtained and subsequently averaged to generate the final evaluation metrics.

This validation strategy is particularly appropriate for agricultural datasets, where sample sizes are often limited and biological variability is inherently high due to genetic diversity, environmental influences, and differences in phenological stages. By reducing variance in performance estimation and improving generalizability, stratified cross-validation enhances the reliability of the reported model outcomes.

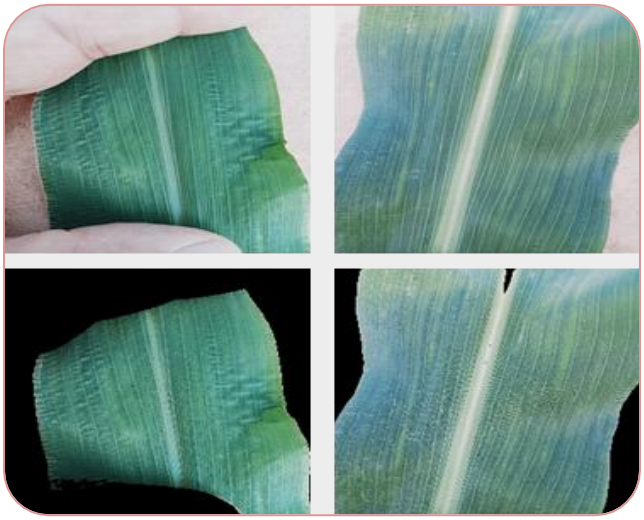

Figure 1. Representative leaf images and their corresponding segmentation outputs generated using a U-Net model.

Model architecture

The proposed model utilizes a Mobile Sparse Vision Transformer architecture as the feature extraction backbone (Figure 2), representing a shift from conventional CNNs toward transformer-based frameworks optimized for computational efficiency and mobile deployment. This hybrid design effectively integrates local feature extraction with global contextual modeling, thereby enhancing representational capacity while maintaining a lightweight structure.

The backbone network comprises an initial convolutional stem followed by five sequential MobileViT blocks with progressively increasing feature dimensions, enabling the learning of hierarchical feature representations. The convolutional stem is responsible for capturing low-level visual features, including edges, color gradients, and fundamental texture patterns, which provide the structural foundation for subsequent high-level semantic learning. The stacked MobileViT blocks further refine these representations by combining convolutional operations for local spatial encoding with transformer-based self-attention mechanisms for global dependency modeling.

This architectural configuration allows the model to efficiently capture both fine-grained local details and long-range contextual relationships, leading to improved feature discrimination and robustness. The detailed architectural specifications are summarized in Table 1.

CA module

The architecture incorporates a dedicated CA mechanism designed to enhance discriminative feature learning by adaptively reweighting informative channels and spatial regions while suppressing irrelevant background noise. This mechanism applies both channel-wise and spatial attention to progressively refine feature representations. The refined feature map is computed as:

$$\boldsymbol{X_{CAA}} = (\boldsymbol{X} \odot \boldsymbol{C_A}) \odot \boldsymbol{S_A}$$

Where C_A is set of learned weights for each channel and S_A 2D map that highlights important spatial regions.

Following each CA module, three sequential Dense Residual Blocks (DRBs) are employed to further refine features and capture patterns at multiple abstraction levels. Each DRB integrates efficient depth wise separable convolutions with residual connections to promote computational efficiency and stable gradient propagation in deep networks. The processing pathway of a single DRB is formulated as:

$$H = BN(PWCONV_{1X1}\left(BN\left(DWCONV_{3X3}(X_{in})\right)\right)$$

$$X_{OUT} = X_{IN} + Dropout\left(GELU(H)\right)$$

Where DWCONV and PWCONV denote depthwise and pointwise convolutions, respectively; BN represents batch normalization; and GELU is the Gaussian Error Linear Unit activation function.

Multi-Scale Feature Fusion

To exploit hierarchical representations, features extracted at different network depths are aggregated through a structured multi-scale fusion strategy. First, global average pooling (GAP) is applied to each feature map:

$$F_i = GAP(X_i) = \left(\frac{1}{H_i W_i}\right) \sum_h \sum_w (X_i(h, w))$$

Where X_i represents the feature map from level i with spatial dimensions $H_i W_i$. Each pooled feature vector then passes through a linear projection layer to map all features to a common dimensionality

$$F_i' = W_i F_i + b_i$$

Where W_i and b_i represent learned parameters specific to each hierarchical level. The projected features are concatenated to form a comprehensive multi-scale representation:

$$F_{concat} = [F_1' \,||F_2'||F_3'||F_4']$$

Where || denotes concatenation along the channel dimension. This concatenated vector preserves information from all hierarchical levels without privileging any particular scale.

A fusion network consisting of two fully connected layers with GELU activation integrates these multi-scale representations, learning optimal combinations of features from different scales:

$$H_{fusion} = GELU(W_1 F_{concat} + b_1)$$

$$F_{fused} = W_2 H_{fusion} + b_2$$

Where W_1 implements dimensionality reduction, W_2 compresses to the final dimensional representation, and b_1, b_2 are bias vectors.

Table 1. Configuration of the MobileViT backbone architecture illustrating the hierarchical feature extraction pipeline. The table presents input and output dimensions, spatial resolution progression across stages, and transformer block parameters (number of attention heads and layers) at each level, highlighting the network's design for efficient multi-scale feature learning in agricultural image analysis.

Block	Input Dim	Output Dim	Spatial Resolution	Attention Heads	Transformer Layers
Conv Stem	3	128	H/2 × W/2	-	-
MobileViT-1	128	128	H/4 × W/4	4	2
MobileViT-2	128	256	H/8 × W/8	4	3
MobileViT-3	256	384	H/16 × W/16	8	3
MobileViT-4	384	512	H/32 × W/32	8	4
MobileViT-5	512	576	H/32 × W/32	8	4

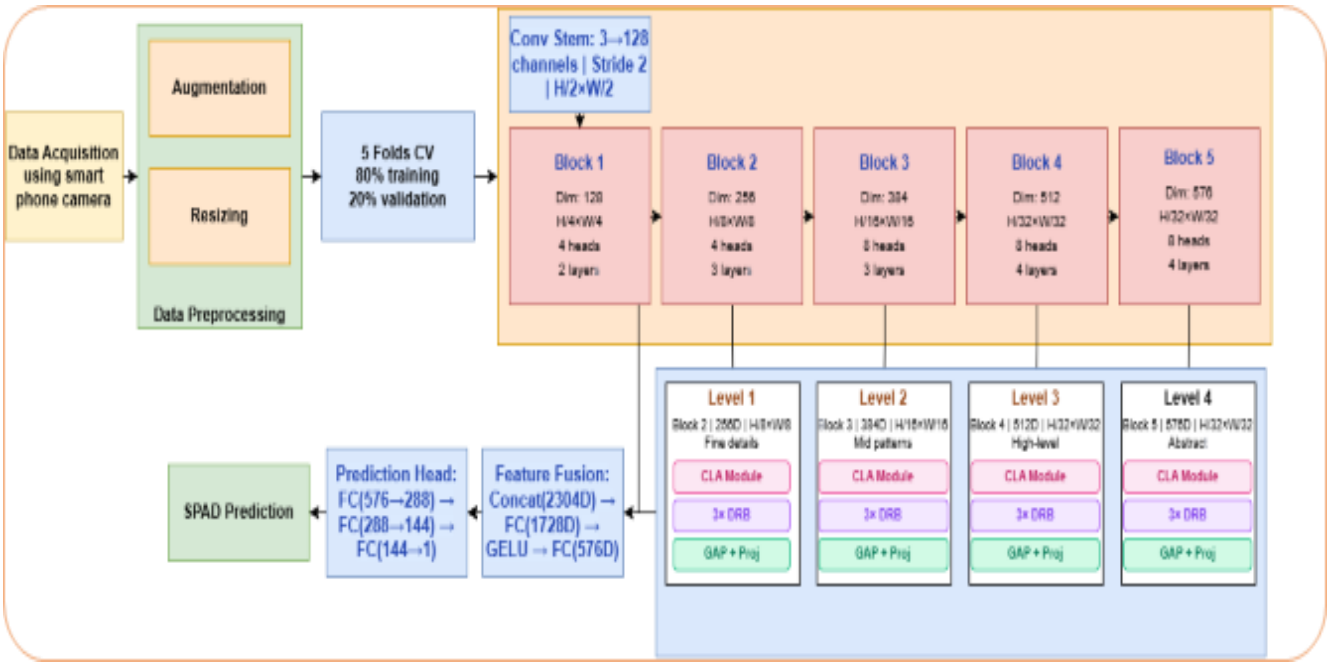

Figure 2. Schematic overview of the proposed framework for SPAD prediction. The workflow begins with image acquisition, followed by preprocessing and augmentation to standardize and enhance the input data. The processed images are then fed into the core model architecture, which extracts hierarchical features and generates the final SPAD value prediction.

Regression output generation

The fused feature vector is passed to a regression head consisting of two fully connected layers with batch normalization, dropout, and GELU activation. The final output layer produces a continuous scalar corresponding to the predicted SPAD value. During training, predictions were optimized using the Mean Squared Error (MSE) loss:

$$L_{MSE} = (\frac{1}{N}) \sum_{i} (\hat{y}_i - y_i)^2$$

Where N is the batch size, $\hat{y}_i$ are predicted SPAD values, and y_i are measured SPAD values.

Training strategy and optimization

A systematic grid search was conducted to determine the optimal hyperparameter configuration (Table 2). This exhaustive approach minimizes selection bias associated with a single validation split and provides a more reliable estimate of generalization performance. Although computationally intensive, this strategy enhances experimental robustness and strengthens the scientific validity of the reported results.

Table 2. Final model configuration, showing the optimal hyperparameters identified through grid search and the corresponding training dynamics.

Hyperparameter	Value
Learning Rate (η)	2×10^{-4}
Dropout Rate (p)	0.3
Model Dimension (d)	576
Weight Decay (λ)	0.01
Optimizer	AdamW
Training Dynamics	
Batch size (Training)	8
Batch size (Validation)	16
Epoch	50

Ensemble strategy and inference

To reduce prediction variance and enhance model robustness, an ensemble strategy was employed by combining three independently trained models with test-time augmentation (TTA). During inference, each model generates predictions from four augmented versions of the same input image, thereby improving generalization and mitigating sensitivity to input variability. The ensemble prediction was computed by averaging the outputs of the three models:

$$\hat{y}_{ensemble} = \left(\frac{1}{M} \sum_{i=1}^{M} T_i(X)\right)$$

Where M = 3 denotes the number of models, T_i represents the i^{th} trained model, and X is the input image.

The complete inference pipeline combines both strategies, implementing two-level averaging across models and augmentations thus predictions are averaged to produce a more stable final output.

$$y^{final} = \frac{1}{M} \sum_{i=1}^{M} [\frac{1}{N} \sum_{j=1}^{N} T_i(A_j)]$$

The complete architecture comprises 45.29 million trainable parameters, corresponding to a storage requirement of 172.76 MB under FP32 precision. In terms of computational complexity, the model requires 10.41 giga multiply-accumulate operations (GMac), which is approximately equivalent to 20.81 GFLOPs, considering that one MAC operation corresponds to two floating-point operations.

The majority of the parameters are concentrated within the MobileViT backbone and the multimodal fusion modules. As demonstrated in the ablation study, these components are critical for achieving optimal performance. Overall, the proposed framework achieves a favorable trade-off between representational capacity and computational efficiency, making it suitable for practical real-world deployment scenarios.

Results

The performance of the proposed regression model was evaluated using the coefficient of determination (R^2). An R^2 value approaching 1 indicates that the model explains a substantial proportion of the variance in the observed data, whereas a value close to 0 reflects limited explanatory capability.

The quantitative results of the ablation study are summarized in Table 3. The observed performance variations associated with the inclusion or exclusion of individual components provide clear insights into their respective contributions to the overall model architecture.

Table 3. Results of the ablation study illustrating the impact of each key component on model performance.

Configuration	Parameters	R^2
Proposed Model	45.29	0.6832
without CA	44.79	0.6721
Without MobileViT	31.99	0.6564
without Fusion	39.31	0.6368

Ablation study

The ablation study systematically quantifies the contribution of each proposed module to the overall performance of the architecture. The complete model (without ensemble learning and TTA, incorporating all architectural components, achieved the highest coefficient of determination (R^2 = 0.6832), thereby establishing a robust baseline.

Importantly, the removal of any individual component resulted in a measurable decline in performance, confirming that each module contributes synergistically to the predictive capability of the network. The most pronounced degradation was observed when the MobileViT module was excluded, with R^2 decreasing sharply to 0.6564. This substantial decline, despite the reduction in model parameters, highlights that the representational capacity provided by the MobileViT block is fundamental to effective feature extraction and contextual modeling.

Similarly, the removal of the CA module led to a notable reduction in performance (R^2 = 0.6721), highlighting its importance in enhancing spatial-channel feature refinement. Furthermore, eliminating the multi-modal fusion mechanism resulted in the lowest performance (R^2 = 0.6368), demonstrating that naive feature concatenation is insufficient and that the proposed dedicated fusion strategy is essential for effectively integrating complementary information across modalities.

Overall, the ablation results validate that each architectural component plays a necessary and complementary role in achieving optimal predictive performance.

Model performance and comparative analysis

The experimental evaluation demonstrates the robustness and reliability of the proposed architecture for SPAD value prediction. Using rigorous 5-fold cross-validation, the base model achieved a mean R^2 of 0.7099, establishing a strong foundation for further optimization.

The integration of ensemble learning combined with TTA yielded substantial performance gains, improving the R^2 to 0.7345. The enhanced model also achieved a RMSE of 4.264 SPAD units and a Mean Absolute Error (MAE) of 3.649 SPAD units.

Figure 3 illustrates the relationship between actual and predicted SPAD values, further confirming the high predictive accuracy and generalization capability of the proposed framework.

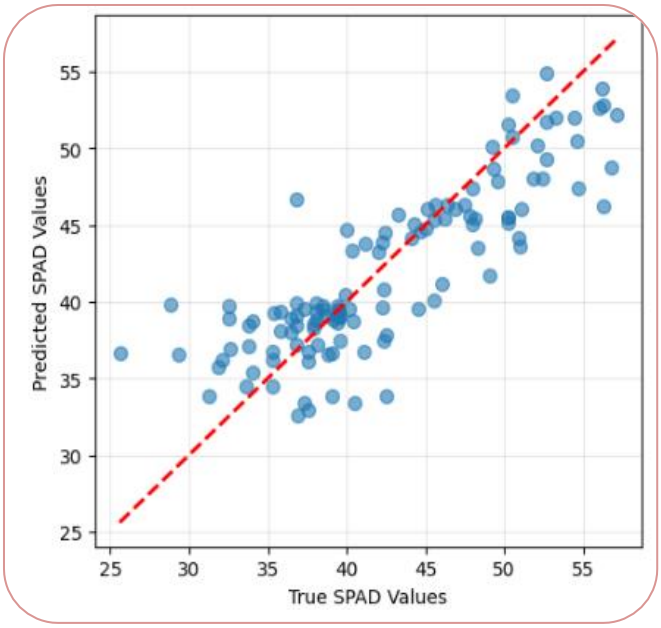

Figure 3. Comparison of true and predicted values for the validation set. The red dashed line represents the ideal fit (y = x). The tight clustering of points around this line, along with an R^2 value of 0.73, indicates low prediction bias and high model accuracy.

Comparative evaluation against established deep learning architectures demonstrates the competitive superiority of the proposed framework. As presented in Table 4, the enhanced model (Ensemble + TTA) achieved the highest predictive performance, surpassing all baseline models, including the strongest competitor, EfficientNet-B3 (R^2 = 0.7078).

Table 4. Comparative performance of SPAD prediction models, evaluated using the coefficient of determination (R^2) and reported as mean ± standard deviation across validation folds.

Model	R^2 (Mean ± Std)
CNN	0.6231 ± 0.0399
ResNet-18	0.7065 ± 0.0286
ResNet-34	0.7015 ± 0.0278
ResNet-50	0.6862 ± 0.0324
ResNet-101	0.6857 ± 0.0324
EfficientNet-B0	0.7096 ± 0.0259
EfficientNet-B1	0.6631 ± 0.0346
EfficientNet-B2	0.7097 ± 0.0267
EfficientNet-B3	0.7078 ± 0.0380
Proposed Model (5-Fold Avg)	0.6832 ± 0.0325
Proposed Model (Ensemble + TTA)	0.7345 ± 0.0321

Interestingly, deeper ResNet variants exhibited a slight decline in performance (ResNet-50: R^2 = 0.6862; ResNet-101: R^2 = 0.6857), indicating potential limitations of deeper transfer learning architectures for this specific regression task. This degradation may be attributed to overfitting, increased model complexity relative to dataset size, or insufficient extraction of features directly relevant to chlorophyll-associated traits.

Beyond predictive accuracy, the proposed system exhibits strong practical attributes that support real-world deployment. The model achieves an inference time of approximately 40 milliseconds per sample on standard edge-computing hardware, enabling real-time processing required for field-based applications.

The progressive performance gains observed from the baseline to the enhanced implementation emphasize the effectiveness of the proposed ensemble and TTA strategies. The relative improvement in the R^2 further highlights the importance of robust inference frameworks in agricultural computer vision systems, where environmental variability and image heterogeneity can substantially affect predictive reliability.

Discussion

The experimental results demonstrate that the proposed hybrid architecture achieves a competitive position in vision-based SPAD prediction, effectively balancing predictive accuracy with computational efficiency for practical field deployment. The obtained R^2 = 0.7345 compares favorably with previously reported RGB-based approaches. Although ensemble strategies applied to spring maize have reported higher predictive performance (R^2 > 0.85) (Xuan et al., 2025), such systems typically rely on multispectral sensors and computationally intensive frameworks. These requirements substantially increase hardware and operational costs, thereby limiting scalability. In contrast, the present study clearly addresses this trade-off by delivering robust performance using cost-effective RGB imagery while maintaining computational efficiency suitable for edge deployment.

The performance advantage of the proposed architecture can be attributed to several methodological innovations that overcome limitations observed in conventional deep learning pipelines. First, the integration of an enhanced attention mechanism facilitated more discriminative feature extraction compared with generic attention modules commonly employed in standard ViTs. For example, Ghazal et al. (2024) reported an R^2 of 0.71 using a fine-tuned ViT model for SPAD-related prediction tasks, indicating comparatively lower performance than the present framework. The improvement from R^2 = 0.7099 (baseline configuration) to R^2 = 0.7345 (proposed ensemble-attention framework) further highlights the importance of advanced inference strategies. This gain aligns with broader evidence from ensemble learning research in precision agriculture, where model averaging and test-time augmentation have been shown to reduce variance and enhance generalization performance under field variability (Gonzalo-Martín et al., 2021; Xuan et al., 2025).

Importantly, the proposed model achieves a favorable balance in the long-standing accuracy-efficiency trade-off that constrains many agricultural artificial intelligence applications. High-resolution UAV-based multispectral systems, such as those described by Yin et al. (2023), and complex hybrid deep learning frameworks (Xuan et al., 2025) can achieve superior predictive accuracy. However, their dependence on specialized sensors, large-scale training infrastructure, and high computational overhead restricts adoption, particularly in smallholder and resource-limited farming contexts. By contrast, the present RGB-based framework supports real-time inference and reduced memory footprint, making it suitable for deployment on low-power edge devices, including smartphones and embedded platforms.

From an agronomic perspective, accurate SPAD prediction is directly linked to nitrogen management optimization, improved chlorophyll assessment, and enhanced crop monitoring efficiency. By lowering the technological and economic barriers associated with advanced phenotyping systems, the proposed approach contributes to democratizing precision agriculture tools. This is particularly relevant for high-throughput phenotyping programs and decentralized decision-support systems, where rapid, scalable, and affordable diagnostics are essential. As highlighted in recent reviews of agricultural computer vision (Junaidi et al., 2025), bridging the gap between laboratory-grade accuracy and field-level deployability remains a central challenge. The present study provides empirical evidence that carefully designed hybrid architectures, supported by ensemble inference strategies, can offer a practical and scalable solution without compromising predictive reliability.

Conclusion

This study presents MoCA-Net, a novel mobile-optimized Vision Transformer framework designed to bridge the critical gap between computational efficiency and predictive accuracy in vision-based SPAD estimation. The principal contributions include the development of a lightweight hybrid architecture, the incorporation of a biologically inspired channel-attention CA module, and a comprehensive multi-scale feature fusion strategy integrated with robust training and regularization methodologies. Rigorous experimental evaluation demonstrated strong predictive performance, achieving an R^2 of 0.73 and a RMSE of 4.2 SPAD units, thereby confirming the model's effectiveness for non-destructive nitrogen estimation under field conditions.

Future research should focus on evaluating the transferability and generalizability of MoCA-Net across diverse crop species with varying leaf morphologies, as well as conducting systematic analyses of environmental confounding factors, including illumination variability, camera calibration differences, and plant water status.

Overall, this work establishes MoCA-Net as a practical and scalable alternative to conventional SPAD meters, contributing to the advancement of accessible, cost-effective, and high-throughput precision agriculture across diverse farming systems worldwide.

Authors' Contribution

RZ, MSF and MHK conceptualized and designed the study. RM conducted field surveys, collected specimens, compiled the dataset. RZ, MSF and RM carried out the validation, data organization, critical revision, and improvement of the manuscript structure and presentation.

Research Funding

This research is partially supported by Pakistan Agriculture Research Board.

Conflict of Interest

The authors declare no conflict of interest.

Sustainable Development Goals Targeted

SDG 2: Zero Hunger
SDG 9: Industry, Innovation, and Infrastructure
SDG 12: Responsible Consumption and Production

References

Berger, K., Verrelst, J., Féret, J.-B., Wang, Z., Wocher, M., Strathmann, M., Danner, M., Mauser, W., Hank, T., 2020. Crop nitrogen monitoring: Recent progress and principal developments in the context of imaging spectroscopy missions. Remote Sensing of Environment 242, 111758.

Borhani, Y., Khoramdel, J., Najafi, E., 2022. A deep learning based approach for automated plant disease classification using vision transformer. Scientific Reports 12, 11554.

Ghazal, S., Kommineni, N., Munir, A., 2024. Comparative analysis of machine learning techniques using RGB imaging for nitrogen stress detection in maize. AI 5, 1286-1300.

Gonzalo-Martín, C., García-Pedrero, A., Lillo-Saavedra, M., 2021. Improving deep learning sorghum head detection through test time augmentation. Computers and Electronics in Agriculture 186, 106179.

Junaidi, A., Hashim, S.Z.M., Othman, M.S.B., Mohamad, M.M., Alhussian, H., Abdulkadir, S.J., Nasser, M., Bena, Y.A., 2025. Deep learning and edge computing in agriculture: A comprehensive review of recent trends and innovations. IEEE Access.Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M., 2022. Transformers in vision: A survey. ACM Computing Surveys 54, 1-41.

Li, G., Wang, Y., Zhao, Q., Yuan, P., Chang, B., 2023. PMVT: A lightweight vision transformer for plant disease identification on mobile devices. Frontiers in Plant Science 14, 1256773.

Li, J., Ge, Y., Puntel, L.A., Heeren, D.M., Bai, G., Balboa, G.R., Gamon, J.A., Arkebauer, T.J., Shi, Y., 2024b. Integrating UAV hyperspectral data and radiative transfer model simulation to quantitatively estimate maize leaf and canopy nitrogen content. International Journal of Applied Earth Observation and Geoinformation 129, 103817.

Li, Z.W., Wang, G.Y., Khan, K., Yang, L., Chi, Y.X., Wang, Y., Zhou, X.B., 2024a. Irrigation combines with nitrogen application to optimize soil carbon and nitrogen, increase maize yield, and nitrogen use efficiency. Plant and Soil 499, 605-620.

Mehta, S., Rastegari, M., 2021. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. arXiv, arXiv:2110.02178.

Saudy, H.S., El-Metwally, I.M., 2023. Effect of irrigation, nitrogen sources, and metribuzin on performance of maize and its weeds. Communications in Soil Science and Plant Analysis 54, 22-35.

Wang, D., Struik, P.C., Liang, L., Yin, X., 2025. Estimating leaf and canopy nitrogen contents in major field crops across the growing season from hyperspectral images using nonparametric regression. Computers and Electronics in Agriculture 233, 110147.

Xuan, F., Su, W., Chen, Z., Huang, X., Zhai, W., Li, X., Zeng, Y., Li, Z., Li, J., Huang, J., 2025. Performance of stacking machine learning and volume model for improving corn above ground biomass prediction. Plant Phenomics 7, 100068.

Yin, Q., Zhang, Y., Li, W., Wang, J., Wang, W., Ahmad, I., Zhou, G., Huo, Z., 2023. Estimation of winter wheat SPAD values based on UAV multispectral remote sensing. Remote Sensing 15, 3595.

Zhao, X., Dong, Q., Han, Y., Zhang, K., Shi, X., Yang, X., Yuan, Y., et al., 2022. Maize/peanut intercropping improves nutrient uptake of side-row maize and system microbial community diversity. BMC Microbiology 22, 14.