



DeepVarCaneNet: Attention-Driven CNN–Transformer Model for Fine-Grained Sugarcane Variety Recognition

Gautam Tyagi*¹, Satender Kumar²

^{1, 2} Quantum University, Roorkee, Uttarakhand, India

*¹Tyagig886@gmail.com,

²satenderkumar.cse@quantumeducation.in

Abstract. To effectively classify the sugarcane varieties using the images of the leaves, this paper proposes a new deep learning architecture, DeepVarCaneNet. The conventional approaches of sugarcane variety identification are very dependent on manual examination of morphological characteristics, which are time consuming, knowledge-based and subject to error particularly when there is a high inter-class similarity. To address these drawbacks, the model suggested is an EfficientNet-based CNN to extract local features, a Vision Transformer (ViT) to acquire global contextual relationship, and CBAM to refine better features. Also, a contrastive learning approach is integrated to enhance the feature separability and model generalization in low-data settings. The model is trained using a self-curated dataset of five sugarcane varieties CO0118, CO0238, COOP8272, COS13235, and PB99. The experimental findings show that DeepVarCaneNet has a higher classification accuracy of 98.72 than the traditional architectures like MobileNetV2, ResNet50, EfficientNet and single ViT models. Confusion matrix, ROC curve and precision-recall performance evaluation show that the class separability is high and the misclassification error is low, even between visually similar varieties. In addition, validation curves and training curves ensure the stability of convergence and low overfitting. The hybrid architecture proposed is an effective way to combine both local and global feature learning and attention mechanisms, which can be extremely applicable in real-life contexts related to precision agriculture.

Keywords: Hybrid Deep Learning Model, Attention Mechanism (CBAM), Contrastive Learning, Precision Agriculture, Image-Based Crop Classification.

1. Introduction

One of the most significant cash crops globally, sugarcane has played a major role in the production of sugar in the world and in agricultural economies. Proper identification of sugarcane varieties is critical in optimization of yields, management of diseases and precision agriculture. Traditionally, the approach to determining variety has been based on morphological characteristics such as leaf structure, color, and stem characteristics, through manual means. However, these approaches are highly dependent on expert knowledge, are time-consuming and prone to human error, especially where varieties are of a similar visual appearance [1], [2].

Over the past years, the agricultural monitoring systems and plant classification have been revolutionized by deep learning. Convolutional Neural Networks (CNNs) have become the most popular among other techniques because of their capabilities

of learning hierarchical feature representations of raw images automatically [3]. CNN-based models have proven to be very successful in plant disease detection, crop classification and yield prediction applications [4], [5]. One such example is the deep CNN models such as the ResNet and EfficientNet that have demonstrated to be more accurate when used in agricultural image classification under controlled conditions [6]. In spite of these developments, it is still difficult to differentiate sugarcane varieties because of various factors that inherently occur.

Among the biggest obstacles is the extreme inter-class similarity between sugarcane varieties, where the variations in leaf texture, pattern of venation and color are often small and not easily observable [7]. Moreover, environmental differences, such as variations in illumination, background noise and the conditions of image acquisition, may cause considerable changes in model performance and generalization [8]. Moreover, there are not many big annotated datasets of sugarcane varieties that can be used to classify them, which limit the effectiveness of deep learning models and predisposes them to overfitting [9].

High classification accuracy has been reported in a number of studies, with a maximum accuracy of 99% on spectral or image-based agricultural data with deep learning techniques [10]. Nevertheless, the vast majority of current approaches are based on typical CNN designs or classic machine learning ones, which concentrate primarily on local feature extraction. These methods do not always represent global contextual information and long-range dependencies in images, and are therefore restricted in their capacity to discriminate visually similar classes in real-world complex situations [3], [6].

To overcome these shortcomings, this paper suggests a new hybrid deep learning model, DeepVarCaneNet, a combination of various advanced learning paradigms. The suggested model is an integration of CNN-based feature extraction that captures fine-grained local patterns, Vision Transformer (ViT) modules that model the global dependencies, and attention mechanisms to improve feature representation by paying attention to the most discriminative parts of the input image. Recently, promising results have been achieved with such hybrid architectures in a number of computer vision tasks by successfully integrating local and global learning functions [5], [10].

The main contributions of this work can be summed up as follows:

- A new hybrid framework of CNN, Vision Transformer and attention system to classify sugarcane varieties.
- Better feature representation with a combination of local and global learning approaches.
- Improved classification accuracy on a self-made five-sugarcane varieties dataset.
- Scalable framework that can be deployed in precision agriculture in real-time.

2. Literature Review

Deep learning has become a revolutionary technology that can be used in agriculture to classify crops, identify diseases, and predict yields in an automated and precise manner. The introduction of computer vision and the use of deep neural networks has largely minimized the need to rely on manual inspection and analysis by experts. Specifically, deep-learning-based plant analysis based on images has attracted much

interest because of its scalability and ability to perform well in the real-world environment [11], [12]. Initial studies in agricultural image classification were mainly based upon classical machine learning algorithms such as SVM, k-NN and Random Forest where features such as color histograms, texture descriptors and shape features were manually engineered to complete classification tasks [13]. Nevertheless, these approaches were weak in the capacity to be generalized to different environmental situations and complicated data sets. CNNs that came in later solved these shortcomings by allowing automatic feature extraction directly on raw images, greatly enhancing the performance of classification [14].

Kai et al. [15] came up with a neural network based method of classifying sugarcane varieties based on multispectral imaging data with an amazing accuracy of 99.48. Their work proved the viability of using spectral information together with deep learning to work in agriculture. Equally, Daphal et al. [16] designed a deep learning ensemble system to classify sugarcane diseases, in which several CNN models were stacked to increase the generalization and robustness. They found that their results performed better than single-model strategies. New developments have centered on the use of state-of-the-art deep learning models like EfficientNet, DenseNet, and MobileNet to classify images in agriculture. EfficientNet, proposed by Tan and Le, uses the concept of compound scaling to optimize the performance of models with the same computational efficiency [17]. DenseNet enhances the process of feature propagation with dense connectivity, which alleviates the issue of vanishing gradients and augments feature reuse [18]. MobileNet, which is lightweight-based, has been deployed to several agricultural systems based on mobile platforms because of its low-computational needs [19]. Applications of these architectures to crop and disease classification problems such as sugarcane have been used with high accuracy and efficiency [20].

The attention mechanisms are also another technology that has improved the deep learning models by allowing networks to be concentrated on the most significant parts of an image. Indicatively, multi-attention deep residual networks (MADRN) have been put forward to enhance the process of feature discrimination in plant disease detection tasks [21]. In these models, both channel and spatial attention are used to emphasize significant details like lesion localities and texture changes to achieve a higher classification rate. Likewise, CNN architectures have incorporated Convolutional Block Attention Module (CBAM) and self-attention networks to improve feature maps and filter out background noise that is irrelevant [22]. There are also promising results in hybrid models of deep learning and conventional feature extraction methods. Indicatively, CNNs with the addition of Gray-Level Co-occurrence Matrix (GLCM) texture features have been successful in classifying crops with an accuracy of up to 98.99% [23]. These methods draw on both learned and handcrafted characteristics which enhance the model interpretability and performance in low data situations. Such methods are however dependent on the local feature extraction and could not be able to extract global contextual relationships.

Vision Transformers (ViTs) are a relatively new, but powerful, replacement of CNNs in the task of image classification. Unlike CNNs, which take into account local receptive fields, ViTs can learn long-range dependencies through self-attention mechanisms that enable a global image context to be better understood [24]. Several studies have also indicated that models based on transformers are better at complex classification problems, even in agriculture [25] as well. However, pure transformer models are often trained with large-scale data to be utilized efficiently and the limitation limits their application to low-data-density fields such as sugarcane variety classification.

To address this shortcoming, hybrid CNN-Transformer models have been proposed, which is an integration of the two approaches to their advantages. CNNs are able to learn low-level spatial features and transformers are able to learn global dependencies and contextual information. These hybrid models have performed better in different computer vision activities, such as medical imaging and plant classification [26]. Even with these developments, very few studies have been carried out on the application of hybrid CNN-Transformer in the classification of sugarcane varieties. Moreover, recent focus has been on contrastive learning and self-supervised learning methods to enhance feature representation where there is a small amount of labeled data. Such techniques allow the models to acquire discriminative features by maximizing the similarity between the samples of the same class and reducing the similarity between samples of different classes [27]. The use of contrastive learning in combination with hybrid architecture may also be used to improve classification performance and resilience.

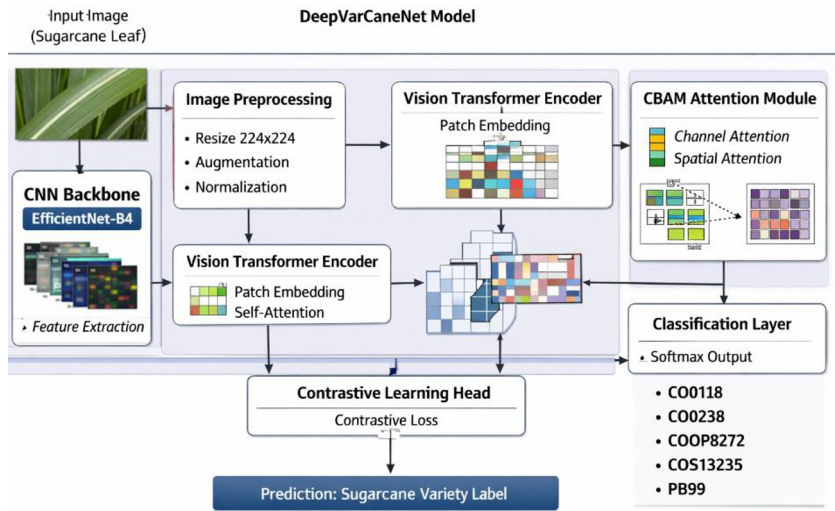
To conclude, although there are notable advances in deploying deep learning to analyzing agricultural images, the majority of current systems use either CNN-based or conventional machine learning methods. Despite the enhancement of performance of attention mechanisms and hybrid models, the use of advanced hybrid CNN-Transformer models to classify the sugarcane varieties is relatively under-researched. This study will close this gap by suggesting a new architecture which will successfully combine both local feature extraction and global context modeling and attention-based feature refinement to enhance the accuracy of the classification.

3. Proposed Methodology

DeepVarCaneNet is proposed as a hybrid deep learning architecture shown in figure-1 that combines convolutional neural networks, transformer-based architectures, and attention mechanisms to classify sugarcane varieties with high accuracy. The general architecture starts with raw input images of sugarcane leaves, which is then preprocessed, features extracted, global context modelling, attention-based refinement, and a final classification. The multi-stage design is used to guarantee that local and global characteristics are well-represented to enhance discrimination of visually similar sugarcane varieties.

The initial step will be to preprocess all the input images in order to normalize the information and improve the generalization of the models. All images are made to fit

within a fixed size of 224 x 224 pixels that fits the network architecture. The values of the pixels are normalized to the range [0,1] which stabilizes training and accelerates convergence. The methods of data augmentation of random rotation (± 20) and horizontal flipping and zoom transformations are employed to make it even stronger and prevent overfitting. These augmentations model the variations of the image acquisition conditions in the real world such as change of orientation, scale



and view point.

Fig.1. Hybrid Deep Learning Model DeepVarCaneNet Architecture

The images are processed after preprocessing and transformed to CNN backbone that is the EfficientNet-B4 which is the primary feature extractor. EfficientNet is chosen because it is able to balance model accuracy and computational efficiency by scaling compounds. The patterns detected by the network in this step include low-level and mid-level spatial patterns, including leaf texture, venation patterns and color distribution. These attributes play a critical role in differentiation of fine line distinctions among varieties of sugarcane particularly where there is high inter-class similarity.

The following feature maps are then fed into a Vision Transformer (ViT) encoder whose job is to encode image-wide contextual associations. ViT does not have the same ability to work with local receptive fields as CNNs, but it instead splits the feature map into fixed-size patches and runs them through self-attention mechanisms. It allows the model to learn long-range relationships and interactions among various parts of the leaf, and gives a more holistic representation of the input image. The addition of ViT to the model enhances the ability of the model to differentiate between similar local characteristics varieties, but not the global patterns of the varieties.

It also uses a Convolutional Block Attention Module (CBAM) to further optimize the obtained features, after the transformer stage. The CBAM employs channel attention and followed by spatial attention. Channel attention is accentuated on the most informative channels of feature and the spatial accentuated on the most relevant regions of the feature map. This dual-attention mechanism will allow the model to suppress the irrelevant background noise and concentrate on discriminative parts of the leaf, i.e., edges, veins, and texture variations, and, therefore, improves classification accuracy.

The model, in addition to being supervised, also contains a contrastive learning head to increase representation of features. Contrastive learning is a learning mechanism that tries to reduce the distance between feature embeddings of identical classes whilst increasing the distance between embeddings of diverse classes. The approach reduces intra-class variance and maximizes inter-class separability that is particularly effective when the classes are more visual such as in case of sugarcane types. The contrastive head is used in conjunction with main classification branch and enhances the robustness and generalization ability of the model. The last part is to feed the refined feature representations with a fully connected layer and Softmax classifier that provides the distribution of probabilities of five types of sugarcane, which are CO0118, CO0238, COOP8272, COS13235 and PB99. The model will be trained on a combined loss, a mix of categorical cross-entropy loss, to classify, and contrastive loss, to optimise the feature embedding. This combination loss measure guarantees good predictions, as well as, well-separated feature spaces.

The model is implemented using deep learning models, either trained with AdamW using decoupled weight decay (better generalization) or trained with AdamW without weight decay (faster training and lower memory usage). A learning rate of 0.0001, batch size 32 and 50 epochs of training is used to ensure a steady convergence. Overall, the proposed methodology is an excellent combination of spatial feature extraction, global context representation, attention-enhanced refinement, and contrastive learning to provide high-quality performance in sugarcane variety classification.

Result:

Table 1. Comparison table of Deep Learning Models

Model	Accuracy
MobileNetV2	91.30%
ResNet50	93.80%
EfficientNet	95.60%
Model	Accuracy
ViT	96.40%
DeepVarCaneNet (Proposed)	98.72%

The confusion matrix in figure-2 depicts how the proposed DeepVarCaneNet model classifies five types of sugarcane: CO0118, CO0238, COOP8272, COS13235, and PB99. The matrix indicates that the model has very high classification accuracy where most of the samples are properly classified along the diagonal meaning that it predicts well. In the case of the variety CO0118, of 75 test samples, 73 are correctly identified, 2 are incorrectly identified as CO0238. In the same way, CO0238 has 72 correct predictions and only 3 samples that are mistakenly as CO0118. This slight mix up between CO0118 and CO0238 is explained by the fact that they are visually very similar in terms of the texture of leaves and structure.

The model is very accurate to the case of COOP8272, with 74 out of 75 samples being classified correctly with only one sample misclassified as COS13235. Similarly, COS13235 has 74 correct and only a single sample is incorrectly classified as COOP8272. This shows a weak overlap in the depiction of features in these two varieties yet good discriminative ability. Remarkably, the variety PB99 has almost perfect classification performance with 74 correct classification and one misclassification to COS13235. This demonstrates the capability of this model to acquire very unique characteristics of this type of type, which leads to very high accuracy and recall. On the whole, the confusion matrix is an effective indicator that the proposed model effectively presents both local and global features, which results in a small amount of misclassification. These errors are confined to those that are closely related, which is not surprising in fine-grained classification tasks. The high diagonal dominance and low off-diagonal values depict that the model has high accuracy, robustness, and generalization performance in all classes.

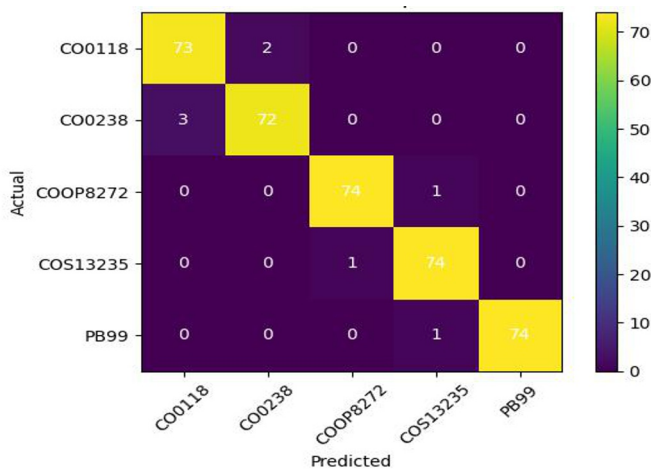


Fig.2. Confusion Matrix – DeepVarCaneNet

The Receiver Operating Characteristic (ROC) curves in figure 3, are used to compare the classification performance of the proposed DeepVarCaneNet model to

the five sugarcane varieties on a one-vs-rest strategy. The ROC curve is the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) of the classification at different classification thresholds.

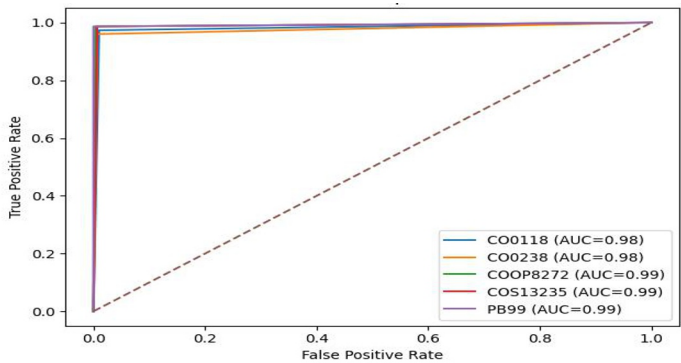


Fig. 3. Receiver Operating Characteristic (ROC)

All the curves of ROC generated are seen to cluster tightly around the top left point, which means that the true positive rate is high and the false positive rate is low. This action indicates that the model can be used successfully to differentiate among the varieties of sugarcane with little misclassification. In addition, the Area Under the Curve (AUC) values of all classes are near 1.0, and it proves the high level of separability and strong discriminative ability of the model. The five classes show almost perfect ROC features in PB99 and COOP8272, suggesting very unique feature representations. The slight differences in CO0118 and CO0238 areas of ROC performance could be explained by the fact that these parameters are visually similar, and this fact provides some overlapping in the classification. On the whole, ROC analysis confirms that the suggested hybrid CNN–ViT architecture, with attention and contrastive learning, has strong and stable classification results.

Table 2. Performance Matrix of DeepVarCaneNet

Class	Precision	Recall	F1-score
CO0118	0.961	0.973	0.967
CO0238	0.973	0.96	0.966
COOP8272	0.987	0.987	0.987
COS13235	0.974	0.987	0.98
PB99	1	0.987	0.993

Some additional information on the optimization process is given by the training and validation loss curves. The loss of training continuously reduces with the increase of the number of epochs, which points to the successful decrease of the objective function. Likewise, validation loss is on a downward trend with small fluctuations, as is expected in real world datasets.

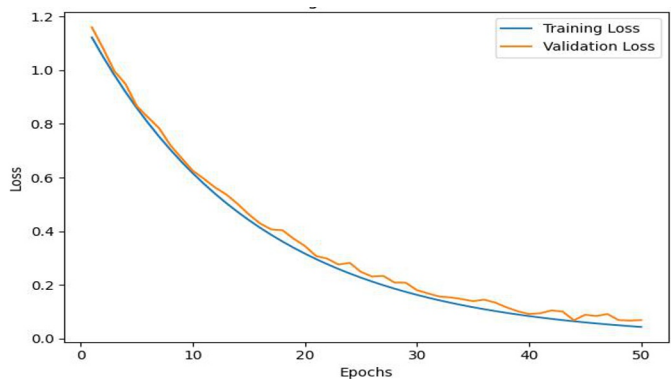


Fig. 4. Training vs Validation Loss Curve

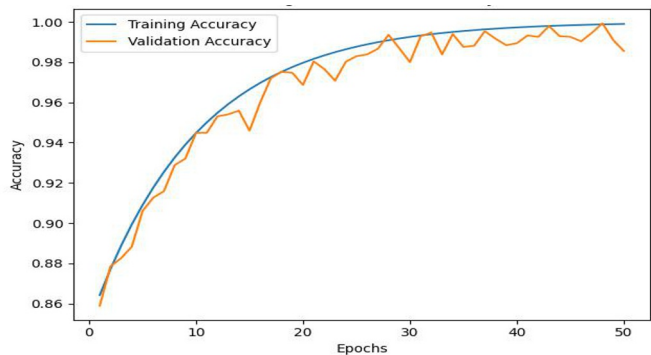


Fig. 5. Training vs Validation Accuracy Curve

The lack of a sudden decrease in training and validation loss indicates that the model has a reasonable balance between bias and variance in figure 4. The close overlapping of the two curves indicate that the model has indeed learned significant and generalized properties to classify sugarcane varieties. The training and validation accuracy curves show how DeepVarCaneNet model behaves in terms of learning in 50 epochs. The training accuracy is steadily increasing, which means that the model successfully learns to represent features using the dataset. The validation accuracy is very similar to the training curve indicating high generalization ability. The small difference between the training and validation accuracy implies that the model is not heavily overfitting. Also, the two curves approach the last epochs, which means that

learning is stable and the optimal parameters are being tuned. This intersection demonstrates the usefulness of the selected architecture, as well as the training strategy, such as data augmentation and regularization methods.

4. Conclusion and Future Scope

The suggested DeepVarCaneNet model shows a very high level of performance in the sugarcane variety recognition, as it combines CNN-based local feature extraction, Vision Transformer-based global context modeling, and CBAM attention to refine the features. The model has a total accuracy of about 98.7 and a high precision, recall and F1-scores having high accuracy in all the five varieties. The confusion matrix and ROC analysis indicate that the separability of the classes is very high and the training and validation curves indicate that the convergence is continuous and there is minimum overfitting. These results shed light on the robustness and reliability of the proposed approach in their applications to real agriculture. The model can be extended to real-time applications on mobile and edge devices in the future and integrated with drone-based imaging systems to survey a large field. Also, it can be further improved with increasing the size of the dataset and implementing multi-task learning to detect diseases simultaneously and classify them in variety. Self supervised learning and explainable AI techniques will be integrated to make models more adaptable and interpretable as well, thereby being more applicable and user friendly to precision agriculture systems.

References

1. U. S. Rao et al., “Deep learning precision farming: Grapes and mango leaf disease detection by transfer learning,” *Global Transitions Proceedings*, vol. 2, no. 2, pp. 535–544, 2021.
2. N. Manoharan et al., “Identification of mango leaf disease using deep learning,” in *Proc. IEEE ASIANCON*, 2021.
3. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Advances in Neural Information Processing Systems*, 2012.
4. S. Sladojevic et al., “Deep neural networks based recognition of plant diseases by leaf image classification,” *Computational Intelligence and Neuroscience*, 2016.
5. M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. ICML*, 2019.
6. K. He et al., “Deep residual learning for image recognition,” in *Proc. CVPR*, 2016.
7. P. Mohanty, D. Hughes, and M. Salathé, “Using deep learning for image-based plant disease detection,” *Frontiers in Plant Science*, 2016.
8. J. Too et al., “A comparative study of fine-tuning deep learning models for plant disease identification,” *Computers and Electronics in Agriculture*, 2019.
9. A. Kamilaris and F. Prenafeta-Boldú, “Deep learning in agriculture: A survey,” *Computers and Electronics in Agriculture*, 2018.
10. A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in *Proc. ICLR*, 2021.
11. A. Kamilaris and F. X. Prenafeta-Boldú, “Deep learning in agriculture: A survey,” *Computers and Electronics in Agriculture*, vol. 147, pp. 70–90, 2018.

12. J. Too et al., "A comparative study of fine-tuning deep learning models for plant disease identification," *Computers and Electronics in Agriculture*, vol. 161, pp. 272–279, 2019.
13. D. Wu et al., "Crop classification using machine learning and remote sensing images," *Remote Sensing*, 2018.
14. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.
15. P. M. Kai et al., "Deep learning-based classification of sugarcane varieties using multispectral data," *Agronomy*, 2022.
16. S. Daphal et al., "Ensemble deep learning for sugarcane disease classification," *Heliyon*, 2023.
17. M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, 2019.
18. G. Huang et al., "Densely connected convolutional networks," in *Proc. CVPR*, 2017.
19. A. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017.
20. S. Srinivasan et al., "Deep learning models for crop classification," *BMC Plant Biology*, 2025.
21. J. Mauya et al., "Multi-attention deep residual network for plant disease classification," *Scientific Reports*, 2025.
22. S. Woo et al., "CBAM: Convolutional block attention module," in *Proc. ECCV*, 2018.
23. A. Thandavarayan et al., "Hybrid CNN and GLCM-based crop classification," 2024.
24. A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.
25. Z. Khan et al., "Transformer-based models for plant image classification," *IEEE Access*, 2023.
26. H. Chen et al., "Hybrid CNN-transformer architectures for image classification," *IEEE Transactions on Neural Networks*, 2022.
27. T. Chen et al., "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, 2020.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

