Early diagnosis of Sugarcane leaf diseases through CNN and Vision Transformer hybrid model

S Aswani
Koneru Lakshmaiah Education Foundation
Ashish Sinha
dr.aashishsinha@gmail.com
Koneru Lakshmaiah Education Foundation

Research Article

Keywords: Convolution Neural networks (CNN), Vision transformer (VIT), early stage leaf diseases detection, deep learning

Posted Date: August 13th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7232200/v1

Additional Declarations: No competing interests reported.

Abstract

Timely and precise identification of foliar diseases in sugarcane is imperative for yield optimization and disease management. This work proposes a hybrid deep learning framework leveraging Convolutional Neural Networks (CNNs) and Vision Transformers (VITs) for automated multi-class classification of sugarcane leaf diseases, including **healthy**, **yellow rust**, **mosaic**, **rust**, and **red rot**. Initially, baseline CNN architecture was employed to extract spatially localized features, attaining a classification accuracy of **84.3%**. Subsequently, a pre-trained VIT model, capable of modelling long-range dependencies through self-attention mechanisms, was fine-tuned on the same dataset, achieving **93.07%** accuracy. To further enhance feature representation, a hybrid CNN + VIT model was constructed by integrating CNN-based local feature encoders with VIT-based global context modelling. The proposed ensemble architecture achieved a superior accuracy of **97.43%**, demonstrating robust generalization and discriminative power. The results affirm the efficacy of transformer-based architectures in plant disease detection tasks and validate the synergy between convolutional and attention-based models for high-resolution agricultural image analysis.

1 Introduction

A significant industrial crop, Sugarcane (Saccharum spp. hybrid) crop is used to produce sugar, bio energy. A persistent difficulty seen across agricultural industry is the timely detection and effective management of crop diseases, which are essential to minimize Considerable setbacks in productivity. In the case of sugarcane, numerous studies have been conducted to enhance crop productivity. However, On-going strategies for disease diagnosis and mitigation largely depend on the visual identification of disease symptoms, which are often only recognizable after substantial crop damage has occurred. This reliance on manual observation limits early intervention and highlights the need for automated, accurate, and early-stage disease detection techniques.

[1] Walia introduced A modified deep learning CNN approach achieves 98.69% accuracy in detecting sugarcane diseases.[2]Bao This research addresses the challenge of early-stage Sugarcane smut and mosaic are among the most prevalent and destructive biotic stresses affecting sugarcane cultivation crops worldwide. A key fungal disease in sugarcane, smut, is caused by the pathogen Sporisorium scitamineum, manifests as whip-like structures emerging from the plant, while mosaic disease, primarily triggered by the Sugarcane Mosaic Virus (SCMV), leads to chlorotic streaks and mottled leaf patterns that impair photosynthesis. Timely and precise identification of these diseases is essential for effective management and control critical to prevent their rapid spread and to mitigate yield losses. However, traditional diagnostic methods rely heavily on visual inspection by agricultural experts, which is often subjective, time-consuming, and ineffective at early stages of infection. As a result, there is a growing need for automated, image-based disease detection systems powered by deep learning techniques, which can facilitate rapid and reliable diagnosis even before visible symptoms become severe. Which are often difficult to identify due to the absence of visible symptoms? By integrating hyper spectral imaging which captures data beyond the visible spectrum with deep learning techniques, the study

successfully detects disease indicators before symptoms appear. A five-month trial confirmed that deep neural networks could extract relevant spectral features for early disease detection with over 90% accuracy. Additionally, the study introduced and released the first large-scale, high-resolution hyperspectral image dataset of both infected and healthy sugarcane plants, contributing valuable resources for future research.[3]Pascal The study proposes a deep learning method for diagnosing sugarcane leaf diseases using Vision Transformer (VIT) models, specifically DeiT3-Small and DeiT-Tiny. Evaluated on a dataset of 6,748 images across 11 disease classes, these models were also compared with popular CNNs. Results showed that higher model complexity doesn't guarantee better performance. DeiT3-Small achieved the best results with 93.79% accuracy. The research demonstrates that VIT-based models can provide fast and accurate disease diagnosis, offering significant benefits for sugarcane crop management.

[4]Singh This study presents a deep learning-based method to detect weeds in sugarcane fields using UAV imagery. Weeds act as camouflage targets due to their similar colour and texture to sugarcane, making detection difficult. By using pixel-level classification and extracting color and texture features with a Deep Neural Network, the method accurately identifies small weed patches. The proposed approach achieved a detection accuracy of 90.5%, demonstrating improved performance over traditional techniques in complex natural backgrounds.[5] Thite This work introduces the Sugarcane Leaf Dataset, comprising 6,748 The dataset comprises high-quality images labeled across nine disease categories, as well as healthy and dried leaves. Covering major sugarcane diseases like smut, yellow leaf, and pokkah boeng, the dataset Ongoing innovation in the field of machine learning models for accurate and efficient disease detection. Its open availability encourages research collaboration, enabling advancements in automated sugarcane disease diagnosis, monitoring, and management—ultimately promoting better crop health and higher yields.

[6] Upadhye This study proposes a modified Convolutional Neural Network (CNN) approach for accurate and timely detection of sugarcane diseases, which are a major threat to crop yield and farmer income. Motivated by the rapid emergence of new disease types and farmers' limited diagnostic capabilities, the model classifies sugarcane leaf images into healthy and diseased categories. Achieving an accuracy of 98.69% across four disease classes, the approach demonstrates the effectiveness of deep learning techniques in supporting early disease management and reducing financial losses in the sugar industry.

[7]Walia research presents a modified Convolutional Neural Network (CNN)-based deep learning approach for detecting sugarcane diseases with high accuracy. Aimed at addressing the rapid emergence of new disease types and the limited diagnostic skills among farmers, the model classifies sugarcane leaf images into healthy and diseased categories. Using a simple CNN trained on four distinct classes, the method Reached a detection performance of 98.69%, demonstrating strong model effectiveness. This approach supports timely disease identification, helping to prevent crop loss and financial damage in the sugarcane industry.

Among all the developed advanced disease detection techniques, early stage disease detection is vital to mitigate these, but identifying diseases in the initial stages is challenging due to subtle and often indistinguishable symptoms. To mitigate these limitations artificial intelligence and deep learning have emerged as transformative technologies, the strengths of CNN and VITs to create a hybrid model for early stage detection. CNN for feature extraction and VIT for contextual understanding.

2 Back ground

This section discusses the concepts necessary for identifying the early stage diseases identification. Since we have used CNN and Vision Transformer for leaf diseases smut, mosaic, yellow, rust and red rot identification in this study, we briefly review the applications of CNN and Vision Transformer for early diagnosis of sugarcane diseases. The Literature review inspired our methodology and settings for CNN and Vision Transformer.

3. Literature Review

[1] Walia In order to detect sugarcane infections, a significant problem in the sugar business, this study employs a modified deep learning CNN technique. The model achieves a 98.69% illness detection accuracy by classifying sugarcane photos into healthy and unhealthy/diseased groups. This strategy tackles farmers' inability to diagnose diseases and the quick evolution of disease classes. [2]Bao This research uses a modified deep learning CNN approach to identify sugarcane diseases, a major issue in the sugar industry. The model categorises sugarcane images into healthy and unhealthy/diseased groups, achieving 98.69% accuracy for disease detection. This approach addresses the rapid evolution of disease classes and the lack of disease diagnostic skills among farmers.

4. Methodology

The following section presents the methodology adopted for early diagnosis of sugarcane leaf diseases using deep learning techniques—specifically Convolutional Neural Networks (CNN) and Vision Transformers (VIT). The entire workflow includes dataset preparation, pre-processing, model architecture design, training, evaluation, and performance comparison. Figure 1 illustrates the process of collection of data and the functioning of the deep learning model.

Figure2: This diagram presents a hybrid CNN-VIT model for sugarcane disease detection. Input images undergo preparation and pre-processing before being analysed by CNN for local features and VIT for global features. These complementary features are fused and passed to a classifier, resulting in a disease label and confidence score as output.

4.1 Dataset Collection

A custom image dataset was curated comprising five distinct sugarcane leaf classes: Healthy, Yellow rust, Red rot, mosaic, rust. The dataset included a total of 2521 images, distributed as follows: Healthy

(522), Mosaic (462), Red Rot (518), Rust (514), and Yellow Rust (505). These images were collected under various lighting conditions and backgrounds to ensure model generalization.

4.2 Data pre-processing and augmentation

To enhance model robustness and prevent over-fitting, the dataset was pre-processed and augmented using the following techniques: Resizing all images to 128x128 pixels for Convolutional Neural Network and 224x224 pixels for Vision Transformer, Normalization (pixel values scaled to [0,1]) Augmentation: rotation (± 25°), zoom (0.2), and horizontal flipping To facilitate model training and evaluation, the dataset was partitioned into 70% training, 15% validation, and 15% testing sets, subsets while maintaining class balance.

4.3 CNN Model Architecture

To perform multi-class classification of sugarcane leaf diseases, a custom Convolutional Neural Network (CNN) was designed and implemented. The architecture consisted of a sequence of convolutional blocks, each aimed at progressively extracting higher-level spatial features from the input images. The first block included a 2D convolutional layer with 32 Each convolution operation used 3×3 filters and was followed by ReLU activation to enhance model non-linearity, activation function and a 2×2 max pooling layer to down sample the feature maps while retaining the most salient features. The second and third convolutional blocks followed the same structure, increasing the number of filters to 64 and 128 respectively. This hierarchical feature extraction enabled the network to learn low- to high-level representations of disease-specific patterns such as colour distortions, texture irregularities, and spot formations in the leaf images.

After the feature representations generated by the final pooling layer are used for classification were flattened into a one-dimensional vector, which served as the input to the fully connected classification head. A dense (fully connected) layer with 128 neurons and ReLU activation was employed to learn complex feature combinations. To reduce the risk of over-fitting and improve generalization, a dropout layer with a dropout rate of 0.3 was inserted after the dense layer. Finally, a soft max-activated output layer with five neurons was used to produce class probabilities corresponding to the five disease categories: Healthy, Mosaic, Rust, Red Rot, and Yellow Rust.

The model was compiled using the Adam optimizer due to its efficiency and adaptive learning capabilities. The loss function used was categorical cross-entropy, which is suitable for multi-class classification problems. Accuracy was used as the primary evaluation metric during training and validation. This CNN architecture provided a strong baseline for disease classification, achieving a test accuracy of 84.3% before integrating the Vision Transformer for performance enhancement.

4.4 Vision Transformer (VIT) Fine-Tuning

To overcome the limitations of spatial locality in CNNs, a pre-trained Vision Transformer (vit-small-patch16-224) was fine-tuned on the same dataset. The VIT model, based on self-attention mechanisms, enables modelling of long-range dependencies and global image context. Input images were resized to 224×224 and normalized using VIT Feature Extractor. The fine-tuned VIT model achieved a significantly improved test accuracy of **93.07%**, outperforming the CNN model by a large margin. The transformer effectively captured subtle variations in colour and texture, leading to fewer misclassifications, particularly in early-stage symptoms. Additionally, its attention maps highlighted relevant leaf regions even when the symptoms were less visually prominent.

4.5 CNN + Vision Transformer Hybrid Model Architecture

The dataset used in this study is titled "Sugarcane Leaf Disease Dataset", containing five categories: Healthy, Mosaic, Red Rot, Rust, and Yellow. Each class contains RGB images resized to 224×224 pixels. The dataset was split into 80% training and 20% validation data. Pre-processing included resizing and normalization using PyTorch transforms. To enhance the classification accuracy and address limitations of local feature extraction in CNNs, a pre-trained Vision Transformer (VIT) model—specifically vit-small-patch16-224—was fine-tuned using the same sugarcane leaf disease dataset. Vision Transformers are built upon through self-attention, the model can effectively capture long-distance interactions and broader contextual information from images, which is particularly beneficial in detecting subtle and complex patterns associated with diseases across the leaf surface. Hybrid model that consists of two branches: a CNN for extracting local spatial features, and a VIT for learning global contextual information. Both feature vectors are concatenated and passed to a classifier for final prediction.

CNN Pathway: The CNN block captures localized patterns such as disease-specific spots or textures. It consists of two convolutional layers with ReLU activations and Max Pooling, followed by a fully connected layer. VIT Pathway: A pre-trained VIT model is used to extract global image features via the [CLS] token. This vector is projected down to 256 dimensions. Fusion and Classification: The CNN and VIT feature vectors (256 each) are concatenated to form a 512-dimensional feature, passed to a linear classifier with 5 output classes.

Fine-tuning the VIT model on the sugarcane dataset enabled the network to capture fine-grained visual cues and subtle texture variations that often go undetected in convolutional approaches. This resulted in a significant improvement in classification performance, achieving an accuracy of **97.43%**, outperforming the baseline CNN model. The hybrid model showed significant improvement over standalone CNN or VIT implementations by combining both local and global features. The CNN path extracted precise disease textures, while the VIT branch captured structural and contextual information. This synergy resulted in robust classification across multiple disease types.

5. Results and Discussion

The following section presents the experimental outcomes of three deep learning approaches—CNN, VIT, and a hybrid CNN + Vision Transformer (VIT) model—for early diagnosis of sugarcane leaf diseases. The objective was to assess how progressively integrating global contextual awareness via transformers enhances classification accuracy over conventional CNN-based models.

5.1 Results summary

CNN-Based Disease Classification

In the initial phase, a custom Convolutional Neural Network (CNN) was implemented as a baseline for sugarcane leaf disease classification. The model was trained on augmented and pre-processed images resized to 128×128 resolution. It comprised three convolutional layers with increasing filter depths (32, 64, 128), followed by fully connected layers and a soft-max output layer. The CNN model achieved a classification accuracy of **85.4%** on the test set. While this approach effectively captured local features such as lesions, spots, and visible leaf damage, it showed limitations in detecting early-stage symptoms where visual cues are subtler and more spatially distributed. Misclassifications primarily occurred between classes with overlapping textures, such as Mosaic vs. Rust and Yellow Rust vs. Healthy.

5.3CNN + VIT Hybrid Model

Building upon the strengths of both models, a hybrid architecture combining CNN feature extractors and VIT-based global encoders was developed. In this setup, initial convolutional layers extracted local features, which were then passed into a Vision Transformer to learn global context. This hybrid model achieved a peak classification accuracy of **97.43%**, marking the highest performance across all experiments. The improvement is attributed to the synergy between CNN's spatial feature localization and VIT's contextual understanding. It effectively reduced class-wise confusion and was particularly robust in detecting diseases at early stages.

5.4Comparative Summary

Table 1
Accuracy comparsions

Model	Test Accuracy
CNN	84.3%
Vision Transformer (VIT)	93.07%
CNN + VIT (Hybrid)	**97.43%**

Declarations

Funding

Not Applicable.

Author Contribution

Aswani S: Conceptualization,Methodology,Design,Model Implementation,data curation and Manuscript Writting.Ashish Sinha:Supervision,Validation Results.

Acknowledgement

The Authors would like to express their Gratitude to KLEF AP University Staff for their invaluable assistance.

Data Availability

The Dataset used in this paper taken from the kaggle.com .

References

1. Walia, R., Kumar, S.: Advancing sugarcane disease detection through CNN-based deep learning. Int. J. Membr. Sci. Technol. **10**, 1851–1861 (2023)
2. Bao, D., Zhou, J., Bhuiyan, S.A., Adhikari, P., Tuxworth, G., Ford, R., Gao, Y.: Early detection of sugarcane smut and mosaic diseases via hyperspectral imaging and spectral-spatial attention deep neural networks. J. Agric. Food Res. **18**, 101369 (2024)
3. Paçal, İ., Kunduracıoğlu, İ.: Data-efficient vision transformer models for robust classification of sugarcane. *Journal of Soft Computing and Decision Analytics*, *2*(1), 258–271.Bao, D., Zhou, J., Bhuiyan, S. A., Adhikari, P., Tuxworth, G., Ford, R., & Gao, Y. (2024). Early detection of sugarcane smut and mosaic diseases via hyperspectral imaging and spectral-spatial attention deep neural networks. *Journal of Agriculture and Food Research*, *18*, 101369. (2024)
4. Singh, V., Singh, D., Kumar, H.: Efficient application of deep neural networks for identifying small and multiple weed patches using drone images. IEEE Access. (2024)
5. Thite, S., Suryawanshi, Y., Patil, K., Chumchu, P.: Sugarcane leaf dataset: A dataset for disease detection and classification for machine learning applications. Data Brief. **53**, 110268 (2024)
6. Upadhye, S.A., Dhanvijay, M.R., Patil, S.M.: Sugarcane disease detection Using CNN-deep learning method: An Indian perspective. Univers. J. Agricultural Res. **11**(1), 80–97 (2023)
7. Walia, R., Kumar, S.: Advancing sugarcane disease detection through CNN-based deep learning. Int. J. Membr. Sci. Technol. **10**, 1851–1861 (2023)
8. Canata, T.F., Júnior, M.R.B., de Oliveira, R.P., Furlani, C.E.A., da Silva, R.P.: AI-Driven Prediction of Sugarcane Quality Attributes Using Satellite Imagery. Sugar Tech. **26**(3), 741–751 (2024)
9. Bhatt, R., Hossain, A., Majumder, D., Chandra, M. S., Ghimire, R., Faisal, M., … Maitra,S. (2024). Prospects of artificial intelligence for the sustainability of sugarcane production in the modern era of climate change: An overview of related global findings.Journal of Agriculture and Food Research, 101519.

10. Wang, W., Li, C., Wang, K., Tang, L., Ndiluau, P.F., Cao, Y.: Sugarcane stem node detection and localization for cutting using deep learning. Front. Plant Sci. **13**, 1089961 (2022)
11. Chen, W., Ju, C., Li, Y., Hu, S., Qiao, X.: Sugarcane stem node recognition in field by deep learning combining data expansion. Appl. Sci. **11**(18), 8663 (2021)
12. Wang, X., Wu, Q., Zeng, H., Yang, X., Yang, X., Yi, X., … Que, Y. (2024). Digital evolution and twin miracle of sugarcane breeding. Field Crops Research, 318, 109588.
13. Daphal, S.D., Koli, S.M.: Enhancing sugarcane disease classification with ensemble deep learning: A comparative study with transfer learning techniques. Heliyon, **9**(8). (2023)
14. Nangare, Y., Pardeshi, N.G.: Result analysis for sugarcane quality inspection using deep learning. *Open Access International Journal of Science and Technology*, *6*(6). (2021)
15. Chen, M., Xu, Q., Cheng, Q., Xiao, Z., Luo, Y., Huang, Y., Wen, C.: Sugarcane stem node detection based on wavelet analysis. IEEE Access. **9**, 147933–147946 (2021)
16. Alencastre-Miranda, M., Johnson, R.M., Krebs, H.I.: Convolutional neural networks and transfer learning for quality inspection of different sugarcane varieties. IEEE Trans. Industr. Inf. **17**(2), 787–794 (2020)
17. Sun, C., Zhou, X., Zhang, M., Qin, A.: SE-VisionTransformer: Hybrid network for diagnosing sugarcane leaf diseases based on attention mechanism. Sensors. **23**(20), 8529 (2023)
18. Sun, C., Zhou, X., Zhang, M., Qin, A.: SE-VisionTransformer: Hybrid network for diagnosing sugarcane leaf diseases based on attention mechanism. Sensors. **23**(20), 8529 (2023)
19. Hernandez, A.A., Bombasi, J.L., Lagman, A.C.: Classification of sugarcane leaf disease using deep learning algorithms. In *2022 IEEE 13th Control and System Graduate Research Colloquium (ICSGRC)* (pp. 47–50). IEEE. (2022), July
20. Tanwar, V., Lamba, S., Sharma, B., Sharma, A.: Deep learning-based approach for leaf disease of sugarcane classification. In *2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT)* (pp. 176–180). IEEE. (2023), April

Figures

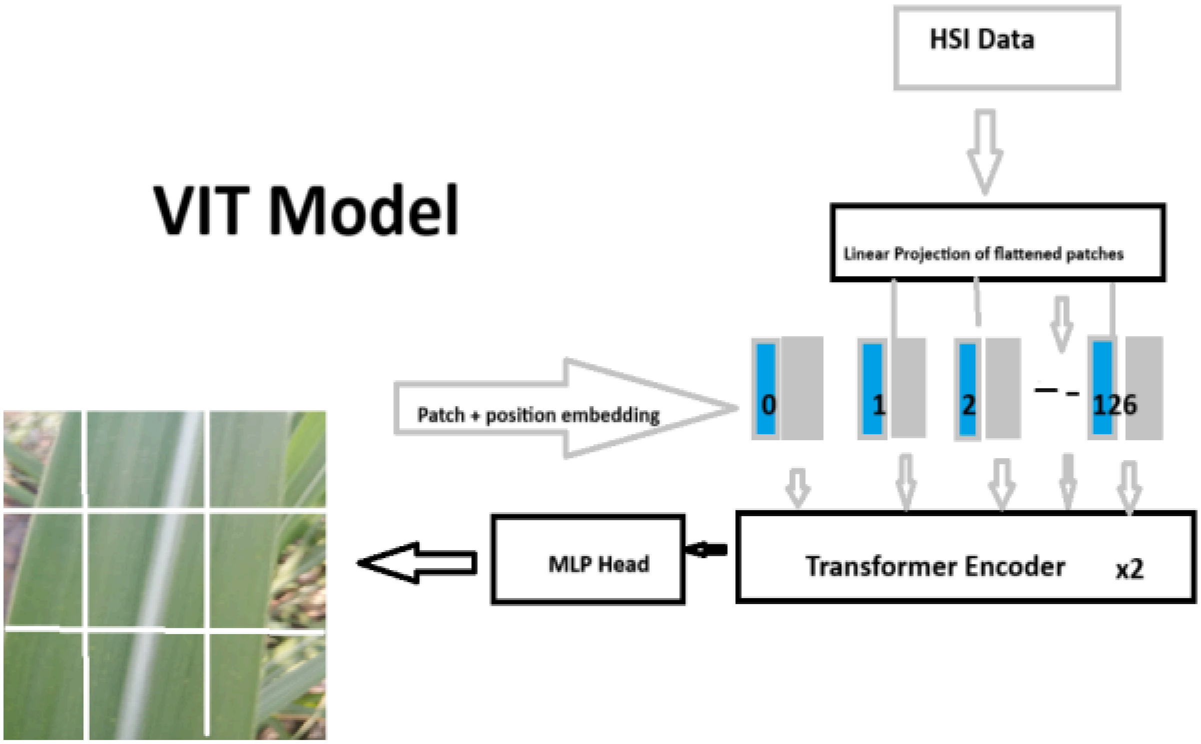

Figure 1

This diagram illustrates the Vision Transformer (VIT) model applied to image data. The image is divided into patches, which are flattened, linearly projected, and embedded with positional information. These embedding's are passed through Transformer encoders and a final MLP head for classification or prediction.

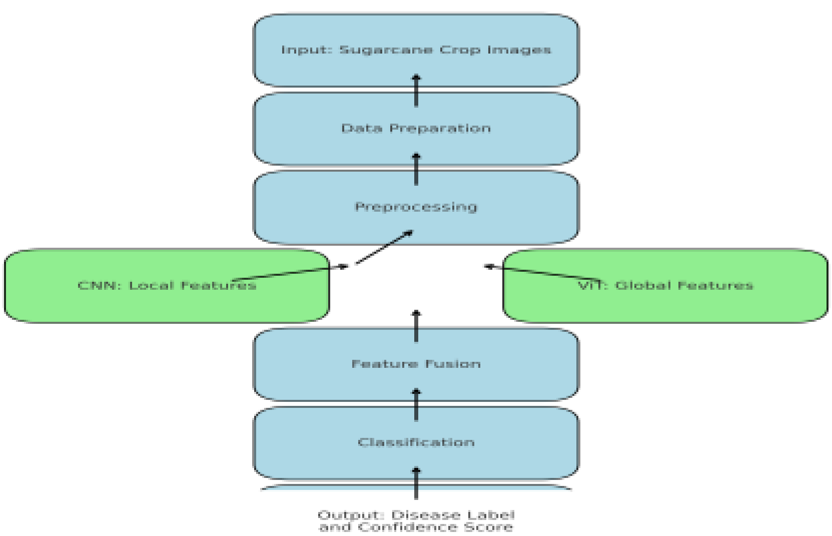

Figure 2

This diagram presents a hybrid CNN-VIT model for sugarcane disease detection. Input images undergo preparation and pre-processing before being analysed by CNN for **local features** and VIT for **global**

features. These complementary features are fused and passed to a classifier, resulting in a disease label and confidence score as output.

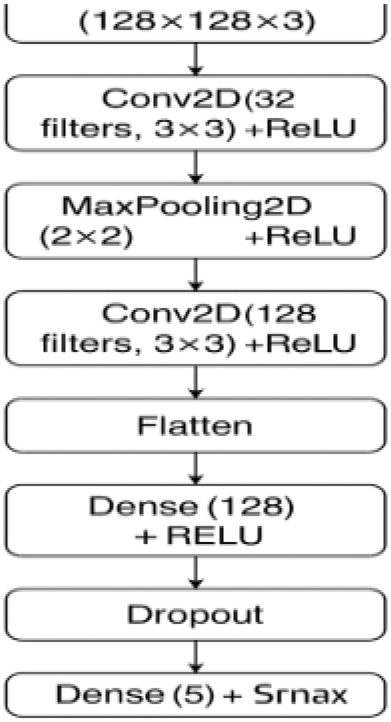

Figure 3

This diagram presents CNN model takes 128×128×3 images as input and applies two Conv2D layers with ReLU activation, followed by Max Pooling to reduce dimensions. The features are flattened and passed through a dense layer with ReLU and a Dropout layer. Finally, a Soft max layer with 5 units performs multi-class classification.

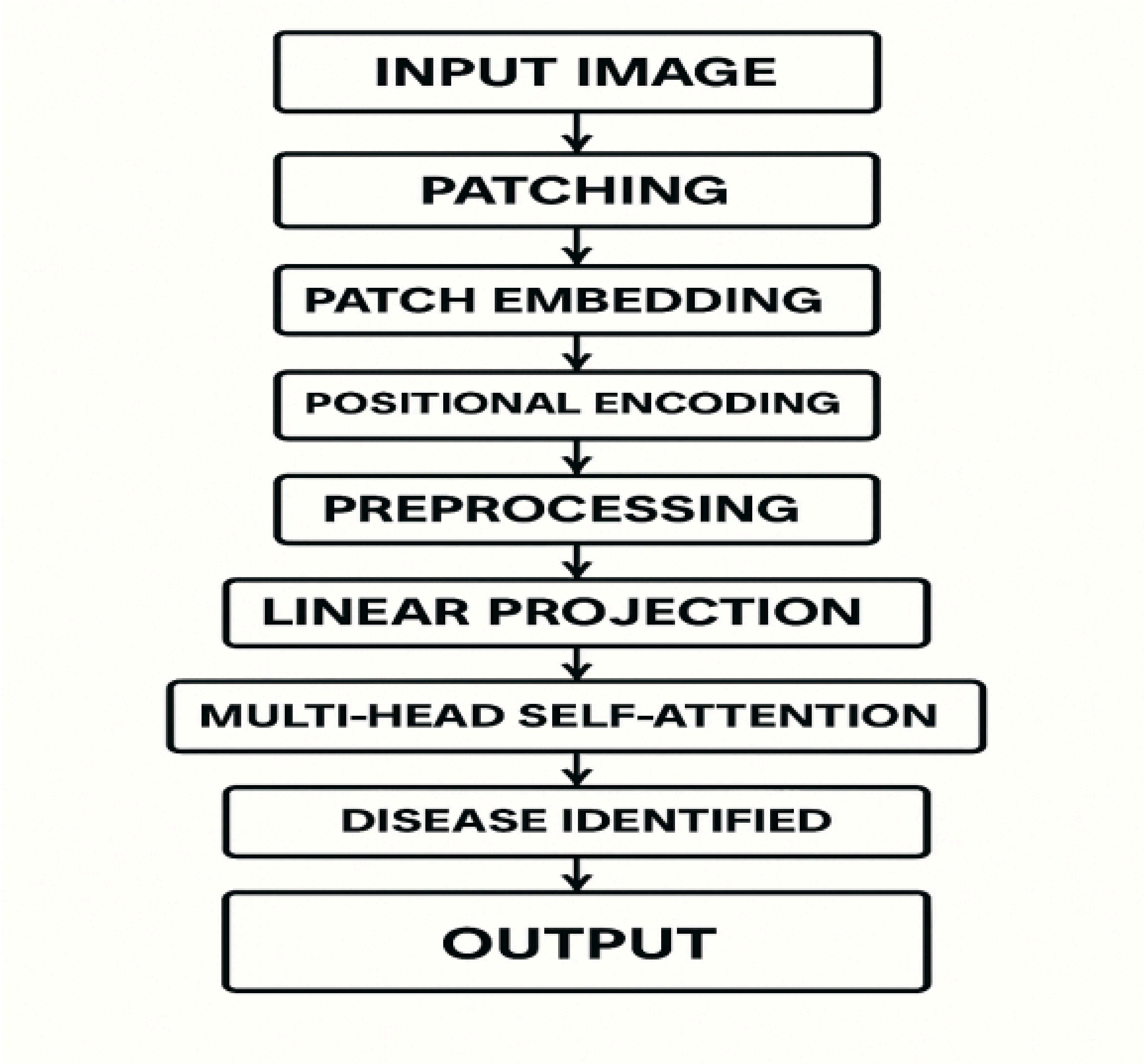

Figure 4

This flowchart represents the Vision Transformer (VIT) process for image-based disease detection. The input image is first divided into patches, which are embedded and combined with positional encoding. After pre-processing and linear projection, the data is processed through multi-head self-attention to extract meaningful features. Finally, the model identifies the disease and produces the output.

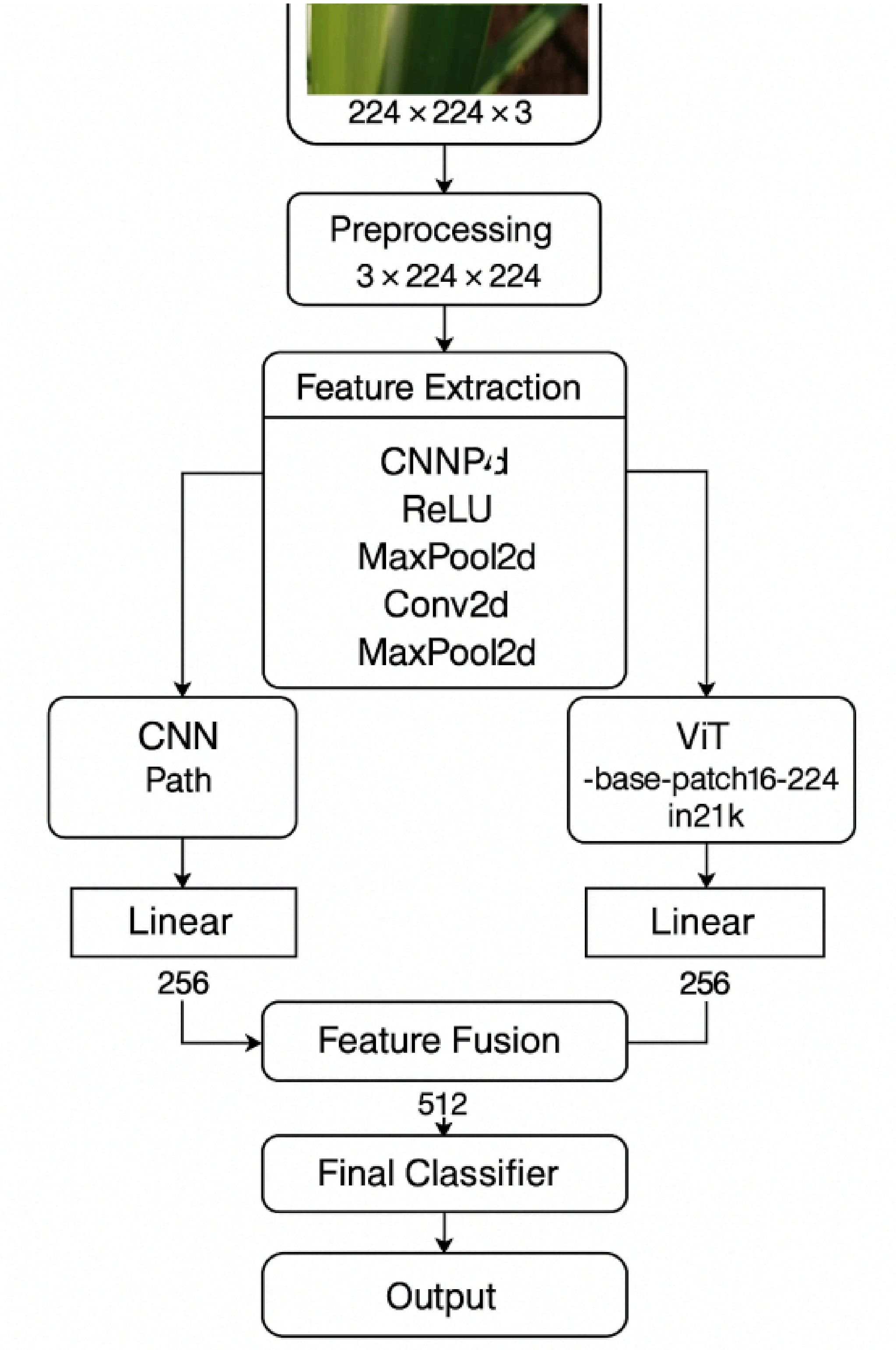

Figure 5

This diagram represents a hybrid CNN–VIT model for image classification. An input image of size 224×224×3 is pre-processed and passed to a feature extraction block with separate CNN and Vision Transformer (VIT) paths. The CNN path extracts local features using convolution, ReLU, and pooling layers, while the VIT path (based on VIT-base-patch16-224-in21k) captures global features. Both feature outputs are linearly transformed and fused into a 512-dimensional vector, which is then passed through a final classifier to generate the output.

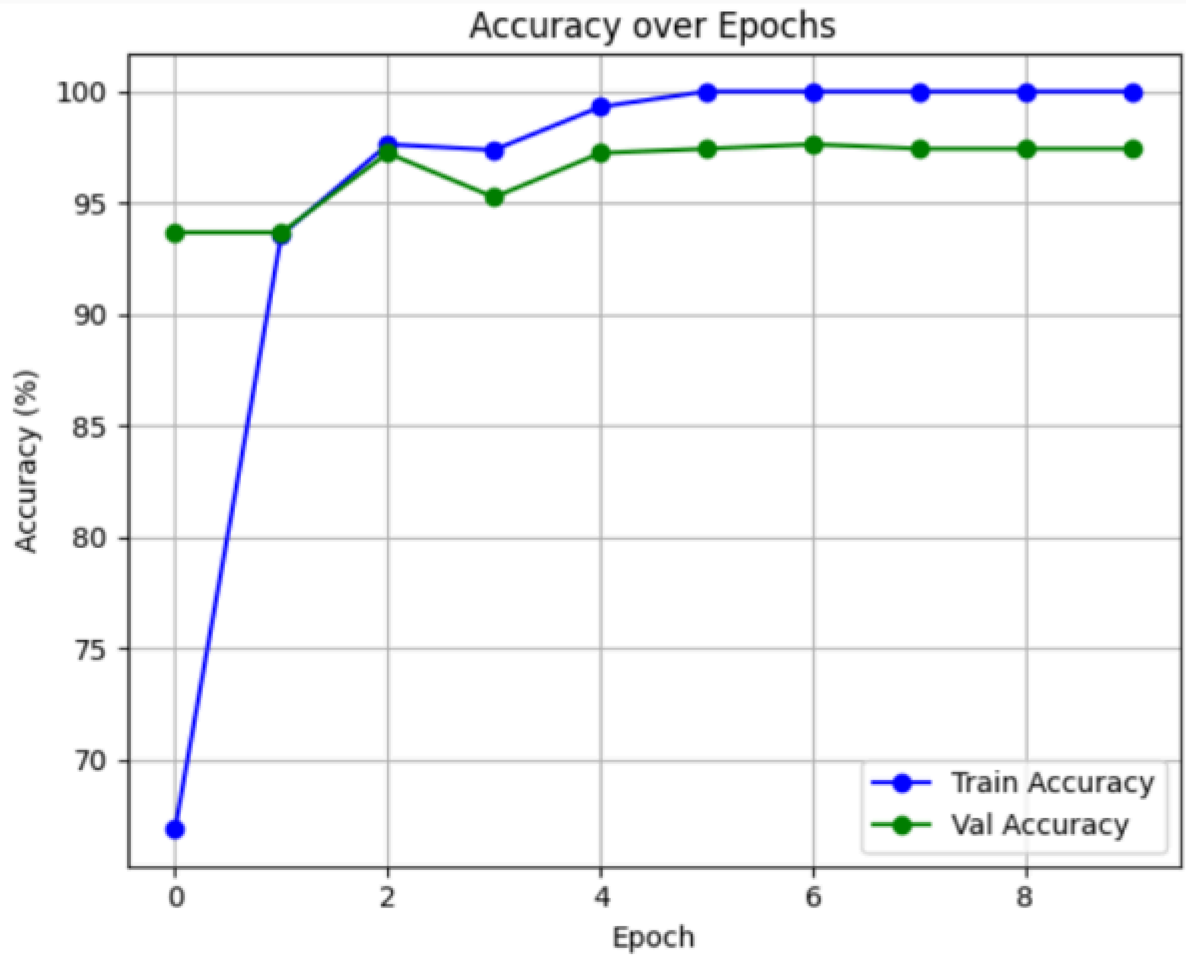

Figure 6

Accuracy of the Model CNN+VIT

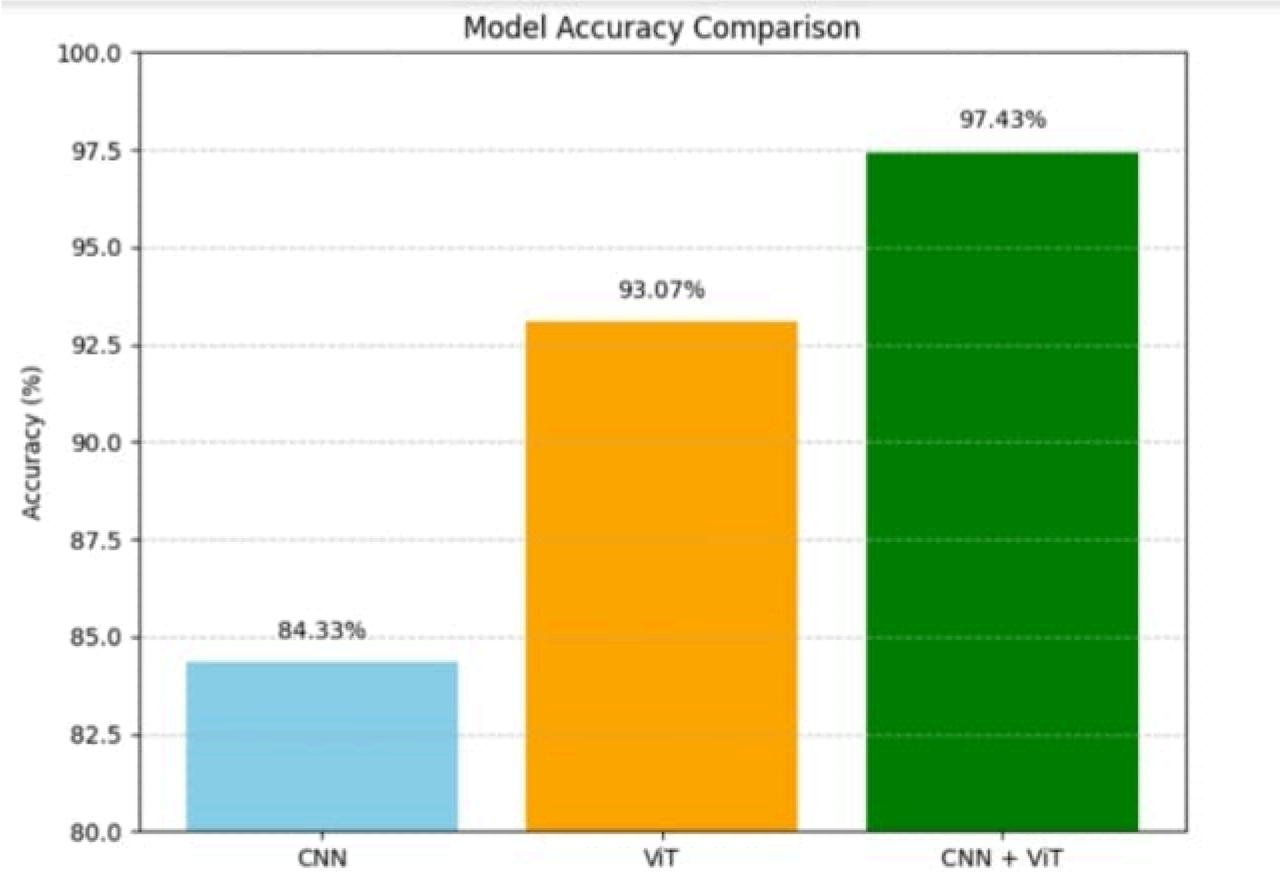

Figure 7

CNN, VIT, CNN+VIT accuracy comparisons.

Figure 8

Actual vs Predicted images of sugarcane Leafs.