International Journal of Computational Methods and Experimental Measurements
Vol. 12, No. 1, March, 2024, pp. 35-43
Journal homepage: http://iieta.org/journals/ijcmem

A Hybrid ViT-CNN Model Premeditated for Rice Leaf Disease Identification

Tharani Pavithra P., Baranidharan B.*

Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur 603203, India

Corresponding Author Email: baranidb@srmist.edu.in

https://doi.org/10.18280/ijcmem.120104

Received: 22 December 2023
Revised: 8 March 2024
Accepted: 20 March 2024
Available online: 31 March 2024

Keywords:
vision transformers, convolutional neural network, rice leaf diseases, attention mechanism, multi-layer perceptron (MLP)

ABSTRACT

Rice, a global staple crop, plays a crucial role in feeding approximately half of the global population. Nevertheless, the persistent spread of diseases poses a significant threat to rice production. Therefore, accurately identifying rice diseases is of paramount practical importance. The proposed approach introduces an innovative hybrid architecture for image classification, harnessing the strengths of both Vision Transformers (ViT) and Convolutional Neural Networks (CNNs). This research investigates five primary diseases affecting rice crops: Blast, Brown Spot, Tungro, False smut, and Bacterial Sheath Blight. Approximately 8000 images of these specific rice leaf diseases were employed for training purposes in the study. What distinguishes this method is its unique integration of a CNN block within the transformer layers, deviating from the traditional ViT architecture. Vision Transformers (ViTs), recognized for their exceptional performance in image classification, excel in providing global insights through attention-based mechanisms. Nevertheless, their model complexity can obscure the decision-making process, and ambiguous attention maps can lead to erroneous correlations among image patches. The incorporation of CNNs in this approach serves to address these challenges by effectively capturing local patterns. This synergistic combination enhances the model's robustness to variations in input data, such as changes in scale, perspective, or context. With the utilization of the proposed hybrid ViT-CNN model architecture, the model achieves remarkable results, boasting 100 percent accuracy and top-5 accuracy, along with a precision of 93.84 percent. Through this hybrid model, we have obtained satisfactory outcomes, surpassing the performance of the latest transformer models in the realm of rice leaf disease identification.

1. INTRODUCTION

The global population is anticipated to expand by around 2 billion people over the next three decades, reaching 9.7 billion by 2050 from the current 7.7 billion and possibly reaching a peak of nearly 11 billion by the year 2100 [1]. Rice stands as a fundamental grain for a large portion of the global population, providing a substantial calorie source for more than half of the Earth's inhabitants [2]. Rice holds a crucial position in India, occupying one-fourth of the total cultivated land. India, ranking second only to China in worldwide rice production, produced a total of 125 million tonnes of rice in the 2022-23 fiscal year. Rice cultivation covered an expansive 45.5 million hectares of land, with an average yield of approximately 4.1 tonnes per hectare. In India, paddy is predominantly grown during the Kharif season and thrives in tropical and sub-tropical regions characterized by hot and humid climates. Nevertheless, rice crops are vulnerable to a variety of diseases, posing a considerable threat to overall agricultural productivity. Paddy diseases can inflict severe damage on rice production and the livelihoods of farmers. Failure to detect these rice plant diseases in a timely manner can have disastrous consequences for food security. The timely anticipation and alert systems fulfil a vital function in mitigating the onset of rice crop diseases and minimizing the unnecessary application of pesticides [3].

In recent years, significant advancements have been made in utilizing deep learning technology for recognition of diseases in plants. Deep learning (DL) technology offers a transparent interface for users, making it accessible even to researchers in the fields of plant protection and statistics who may not have high levels of expertise. Deep learning (DL) has the capability to automatically extract features from images and categorize disease spots on plants, thereby eliminating the labour-intensive processes of feature extraction and classifier design commonly associated with traditional image recognition technology. Furthermore, deep learning can capture the intrinsic characteristics of original images and offers an end-to-end approach. These qualities have garnered widespread attention for deep learning-based technology in the field of plant ailment recognition, making it a prominent and highly researched topic. Numerous research papers have leveraged deep learning techniques to enhance the meticulousness of rice disease recognition [4].

Many current research endeavors in agritech are centered around the application of computer-based deep learning and machine learning based techniques for the detection of diseases in rice leaves. Nonetheless, there is ample room for

enhancements in this domain, particularly concerning decision-support systems aimed at transforming extensive datasets into actionable recommendations. While numerous studies have utilized deep learning methods to enhance the accuracy of rice disease recognition, there exists a significant research gap in the broader field of plant disease recognition. Specifically, there is insufficient investigation into effectively integrating deep learning models within a hybrid architecture tailored specifically for plant disease recognition. Furthermore, despite the transparency and automated feature extraction offered by deep learning technology, there is a crucial need to evaluate the suitability and effectiveness of hybrid architectures in addressing the challenges inherent in plant disease recognition. This involves optimizing classification accuracy while taking into account factors such as dataset characteristics, model interpretability, and scalability across diverse plant species and disease types.

Presently, the literature predominantly focuses on refining the integration of Vision Transformers (ViT) and Convolutional Neural Networks (CNNs). However, a more comprehensive exploration is required to fill the identified gaps and propel advancements in the field of plant disease recognition. The primary goal of this research is to develop a system that leverages state-of-the-art, refined methodologies in integrating ViT and CNNs. The hybrid approach, known as Vit-CNN, aims to combine the detailed, localized feature recognition abilities of CNNs with the comprehensive, contextual understanding offered by ViTs. This system is engineered to autonomously and accurately identify, categorize, and diagnose rice diseases, including their early manifestations, eliminating the need for human intervention.

The remaining part of this document is structured as follows: Starting with an extensive literature review in Section 2, the study lays the groundwork for its contributions by scrutinizing previous research efforts, thereby establishing a contextual framework. In Section 3, we delve into the specifics of the obtained images and basics of the base models used. Section 4 gives an in-depth overview of the proposed approach for recognizing crop diseases and also, we delve into the experimental analyses, where we conduct comprehensive experiments and evaluate the results through comparative analysis. It also demonstrates the potential applications of the ViT-CNN Model in real-world settings, particularly in its role as a decision support system within the field of agriculture. Lastly, Section 5 serves as the conclusion and future work of the paper.

2. LITERATURE REVIEW

Deng et al. [5] and their colleagues with a substantial dataset of 33,026 pictures, encompassed six distinct categories of rice ailments. At the core of their approach was an Ensemble Model integrating multiple sub-models. The process commences with image preprocessing in the methodology developed by Haridasan et al. Subsequently, image-based segmentation is working to pinpoint the areas of the rice crop affected by disease. To accurately recognize and classify specific types of paddy plant diseases, the researchers adopt a hybrid approach. This approach works well where a classifier combining support vector machine with convolutional neural network functions. Activation functions like Rectified Linear Unit (ReLU) and softmax are utilized within these neural networks to ensure precise disease recognition and classification [6]. Sharma et al. [7] introduced the application of computer vision techniques, specifically CNN, combined with traditional machine learning methods. Their primary objective is the identification of diseases affecting plant leaves, with a specific focus on rice and potato plants. The CNN model proposed by the researchers is instrumental in effectively classifying diseases that affect these particular plant species. Latif et al. [8] have put forth an innovative method for the accurate detection and classification of rice leaf diseases by leveraging Deep Convolutional Neural Networks (DCNN) and transfer learning techniques. This refined approach integrates a customized transfer learning methodology grounded in the VGG19 architecture. Through this adapted system, the identification and diagnosis of six distinct disease classes affecting rice leaves can be accomplished.

Santosh Kumar and collaborators present an effective methodology for detecting rice plant diseases through the utilization of CNNs. The primary focus of their research centers on important widely acknowledged rice ailments: foliar smut and brown spot, attributed to fungi, and bacterial leaf blight, caused by bacterial infection. The paper introduces a robust method for the identification and categorization of rice crop ailments based on the characteristics of lesions in leaf images, encompassing aspects such as size, shape, and color. To bolster the precision of disease detection, the suggested model integrates Otsu's thresholding on a global scale for converting images to binary form, proficiently removing circumstantial noise from the images [9]. Chen et al. [10] and colleagues undertook a comprehensive exploration of deep learning techniques, leading to the creation of an ensemble of convolutional networks designed to advance the model's capacity to detect nuanced features in plant lesions. Applying ensemble learning principles, they amalgamated three lightweight CNNs to create an innovative network named "Es-MbNet." This composite network was specifically crafted for the recognition of diverse plant diseases. Zhou et al. [11] and colleagues introduced a unique architecture known as the "residual-distilled transformer." Drawing inspiration from the initial successes of utilizing transformers for computer vision tasks, they integrated a distillation strategy to extract and refine weights and parameters from pre-trained vision transformer models. Subsequently, these extracted features are fed into a multi-layer perceptron (MLP) to make predictions. Sudhesh et al. [12] presented an innovative approach for identifying rice leaf diseases using Dynamic Mode Decomposition (DMD) coupled with attention-driven preprocessing. They focused on four distinct categories of rice leaf diseases which comprises four sets of experiments, evaluated the effectiveness of ten pre-trained DCNN models.

Upadhyay and Kumar [13] devised a straightforward, rapid, and efficient deep learning framework for the early detection of brown spot disease. This method integrates infection severity estimation via image processing techniques. The proposed approach involves two main phases: initially, the dataset containing brown spot-infected leaf images is divided into two subsets, namely early-stage brown spot and developed-stage brown spot. Subsequently, a fully connected CNN architecture is constructed in the second phase to facilitate automatic feature extraction and classification. Aristan and Kusuma [14] utilized a dataset containing 79 different plant classes, sourced from several public domain datasets, which they assessed and compared using four CNN models: MobileNetV3, EfficientNetB0, Mason model, and

ShuffleNetV2. Results from the experiments indicated that the Mason model achieved the highest accuracy among the four. However, all models experienced a slight decrease in accuracy when evaluated on both workstation and mobile devices. In terms of resource consumption, MobileNetV3 exhibited lower consumption compared to the other models overall. Table 1 presents a summary of different models categorized by the algorithms employed, the addressed problems, as well as their respective advantages and disadvantages.

Table 1. Evaluation of various extant models

Ref No	Methodology	Merits	Demerits
[5]	An Ensemble Model integrating: DenseNet-121, SE-ResNet-50, and ResNeSt-50	High accuracy, simplicity, and cost-effectiveness	Many parameters, which may affect the speed of identification
[6]	Support Vector Machine, CNN	CNN surpasses SVM in performance	Accuracy can still be enhanced
[7]	Convolutional Neural Network (CNN)	The CNN model showed superior accuracy, surpassing SVM, KNN and Decision Tree	Selection of hyperparameters can increase the performance of the CNN model
[8]	Deep Convolutional Neural Network (DCNN)-VGG19 -based transfer Model	Attains high accuracy for Non-Normalized and Non-Normalized Augmented data	Still Accuracy should be developed for Normalized augmented data
[9]	Deep Convolutional Neural Network (DCNN) Otsu's global thresholding technique	Decreases both time and model complexities while enhancing performance	It is expected to perform better when the severity of each disease is high
[10]	SE-MobileNet, Mobile-DANet, and MobileNet V2 are combined as ESMbNet	Efficient in recognizing plant diseases on both open-source and local datasets	Misidentifications of samples with highly complex background conditions
[11]	Residual Distilled Transformer Vision Transformer	Effectively highlight the location of rice disease	Demands high computing resources
[12]	10 pre-trained Deep CNN Models (DCNN) Dynamic Mode Decomposition (DMD)	Accuracy, Precision, high in models trained on DMD pre-processed images	Identifying the diseased area amidst complex backgrounds poses significant challenges
[13]	Otsu's thresholding technique Convolutional Neural Network (CNN)	Can easily handle large data sets and its ability to recognize the disease at an early stage	A threshold of 1% might lead to oversensitivity or insufficient segmentation
[14]	MobileNetV3, EfficientNetB0, Mason model, and ShuffleNetV	Mason model is still higher in accuracy than EfficientNetB0 and MobileNetV3 for mobile devices	Overfitting obtained in each model. Mason model requires higher resource consumption

Note: All articles reviewed in the literature focus on the identification of diseases in rice crops.

3. MATERIALS AND METHODS

3.1 Data acquisition

This study examines five major diseases that afflict rice plants: Blast, Brown Spot, Tungro, Falsemut, and Bacterial Sheath Blight. While many contemporary studies prioritize performance metrics derived from publicly available datasets, they often overlook the importance of gathering real-world data from actual planting environments. Many current datasets, such as those from AI Challenger, Plant Village, and research institutions' standard sets, have been sourced from online platforms or curated by research entities [15, 16]. These datasets usually present images with consistent backgrounds and lighting conditions, facilitating high predictive accuracy. However, this uniformity may not be representative of the complexities encountered in genuine farming environments, where varying backgrounds and noise interferences prevail. Therefore, emphasizing data from real-world settings can significantly enhance model robustness.

In December 2021, real-world data comprising both affected and images depicting healthy rice crops were sourced from cultivated lands in Melmaruvathur, Kavaraipettai, and Gummidipondi regions in Tamilnadu.

This dataset, consisting of approximately 1500 images, was captured using Xiaomi and Redmi smartphones boasting a 48-megapixel resolution. However, the authenticity of these open-field captures also meant they were susceptible to environmental noises and distortions. These images, available in JPG format, showcased a diverse range of backdrops. Some depicted other plant leaves or field grass, while others showed varied soil colours. On occasion, inadvertent inclusions like the photographers' fingers were evident. Additionally, inconsistent lighting due to fluctuating weather conditions further complicated the dataset. To augment this collection, supplementary pictures were procured from numerous repositories, including Kaggle's "Rice Leafs Diseases Dataset," UCI Machine Learning Repository's "Rice Leaf Diseases Dataset," and the "Rice Leaf Diseases Image Samples" the dataset curated by Prabira Kumar Sethi, published in Computers and Electronics in Agriculture. One of the inherent challenges with training CNNs is ensuring uniformity in image dimensions, especially when the training dataset contains images of varying sizes. To address this, the study employed data augmentation techniques to resize all images to the CNN's expected 224×224 input dimension. Image stabilization was also achieved to counter gradient propagation problems. Furthermore, image processing techniques such as Erosion, Dilation, Opening, and Closing were utilized to enhance image regions with varying brightness. In total, the study utilized around 8000 images of the aforementioned rice leaf diseases for training. Figure 1 shows the pictorial representation of rice leaf diseases used in this research.

3.2 Base models

3.2.1 Convolutional Neural Network (CNN)

CNN is a category of deep, feed-forward neural networks optimized for interpreting visual data. These networks are

engineered to autonomously and dynamically discern spatial patterns and structures within images. The power of CNNs stems from their capability to autonomously discern spatial patterns in data, eliminating the need for manually designed features.

CNNs are versatile, capable of processing various data types including images, videos, audio, speech, and natural language. In its structural composition, a CNN encompasses multiple layers, initiating with a convolutional layer and advancing through pooling, Relu activation, ultimately culminating in a fully connected layer. As depicted in Figure 2, each input image undergoes several transformations-filtering, reduction, and correction-to eventually be represented as a vector. The essence of a CNN's power resides in its convolutional layers, where it learns the most pertinent filters for specific tasks, such as detection. There's also a cascading effect: the output from one convolutional layer serves as the input for the next. Following the convolutional layers, the pooling layer plays a crucial role. It down samples the data, leading to significant reductions in computational demands, memory needs, and parameter counts

The fully connected layers, true to their name, maintain comprehensive connections to their preceding layers. Ordinarily, these layers utilize functions such as "sigmoid" or "softmax" in the concluding layer to generate predictions regarding classes. Fundamentally, the convolutional layers discern features extracted from the input data, which the pooling layers subsequently condense. Using the high-level features gleaned, the fully connected layers usually classify input data into predefined categories in the final stages. Furthermore, the classification layer not only categorizes data but also extracts features essential for both classification and detection activities [17].

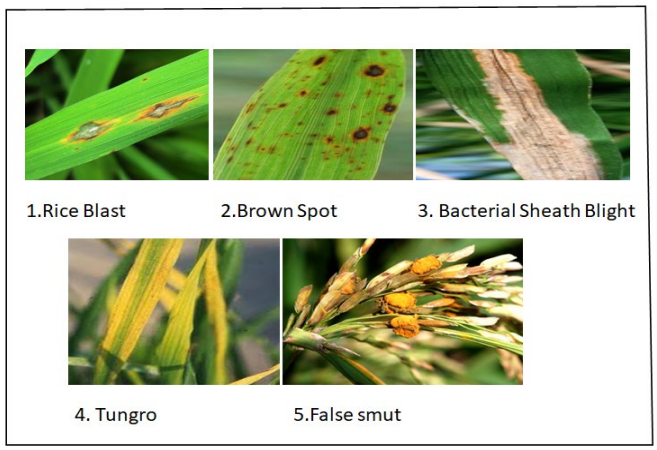

Figure 1. Ailments impacting rice crops

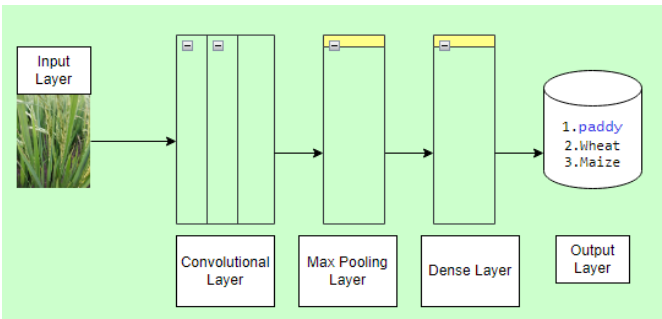

Figure 2. CNN architecture

3.2.2 Visual induced transformer

The Transformer functions as a groundbreaking transduction model, utilizing self-attention exclusively to generate representations for its input and output. This eliminates the necessity for sequence-aligned RNNs or convolution. The architecture involves stacked intra-attention and fully connected layers at individual data points for both the encoder and decoder, illustrated in the left and right sections of Figure 3, respectively [18].

In this process, the encoder receives an input sequence of symbol representations (x1, ..., xn) and converts it into a sequence of continuous representations (z1, ..., zn).

Once we have these continuous representations, the decoder proceeds to produce an output sequence (y1, ..., ym) of symbols individually. The model functions in an auto-regressive fashion. Tugrul et al. [17] uses the symbols generated earlier as extra input at each step when generating the next symbol.

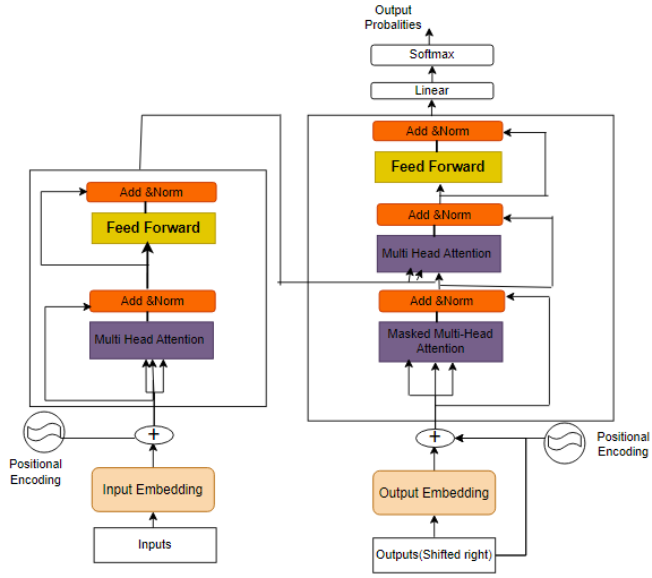

Figure 3. Visual induced transformer architecture

3.2.3 Attention mechanism

The Transformer architecture is centred around the attention mechanism, which comprises three key attention modules: self-attention, masked attention, and cross-sequence attention [19].

Self-attention is a versatile mechanism widely applied in the field of visual learning and comprehension. Its primary goal is to capture the inherent relationships within data or features, thereby diminishing the reliance on external information. This mechanism effectively tackles the issue of handling long-range dependencies by computing the mutual influence between various image patches.

When applied to an image X, the self-attention mechanism can be described as follows. The queries (Q), keys (K), and values (V) are generated through transformations of the input. A commonly used formulation for Q, K, and V is denoted by Eq. (1) [20].

$$K = W^K X, Q = W^Q X, V = W^V X \quad (1)$$

The Eq. (2) gives the scaled dot-product attention as

$$Attention(Q,K,V) = \text{softmax}\frac{\left(QK^{T}\right)}{\sqrt{d_k}}V \quad (2)$$

Instead of relying on a single attention function solely for queries, multi-head attention (MHA) improves the scaled dot-product attention model. It introduces the concept of employing multiple attention functions concurrently to capture diverse information from distinct representation subspaces. In multi-head attention, there are 'h' parallel 'heads,' each representing an independent scaled dot-product attention function. The combined attended features, denoted as F in the multi-head attention functions, can be expressed as

$$F = MHA(Q,K,V) = Concat(h1,\ldots,hh)W^{O} \quad (3)$$

$$h_i = A\left(QW_i^{Q}, KW_i^{K}, VW_i^{V}\right) \quad (4)$$

where, QW_i^{Q}, KW_i^{K} ,$VW_i^{V} \in R^{dxdh}$ are the projection matrices of the i^{th} head. $W^{O} \in R^{h*dhxd}$ is the output projection matrix that aggregates the information from different heads [21].

To preserve positional information, Position embeddings are incorporated into the patch embeddings. These position embeddings are typically standard, learnable 1D position embeddings, as there hasn't been a significant improvement in performance observed by employing more sophisticated 2D-aware position embeddings. The resultant sequence of embedded vectors is then used as input for the encoder [22].

4. PROPOSED METHOD

4.1 Hybrid ViT-CNN architecture for image classification

4.1.1 Enhancing vision transformers with CNN blocks: significance and benefits

The proposed approach employs a novel hybrid architecture for image classification, combining the strengths of both ViT and CNNs. What sets this method apart is its incorporation of a CNN block within the transformer layers, a departure from the conventional ViT architecture. This design involves dividing the image into patches, leveraging Transformer blocks to capture inter-patch relationships, and ultimately making accurate image classifications among different categories. The effectiveness of this model is contingent on the seamless integration of ViT and CNN components, tailored to the characteristics of the dataset and the specific classification task at hand.

By integrating CNNs with ViTs within a hybrid architecture, the combined model harnesses the complementary attributes of each, yielding a host of advantages:

i) Streamlined Handling of Local and Global Features:

CNNs excel in extracting localized, hierarchical features through their convolutional structure. Their proficiency lies in grasping spatial hierarchies within images, where higher layers build upon details from lower layers.

ViTs, in contrast, demonstrate proficiency in capturing global dependencies and contextual information across the entirety of an image, facilitated by their self-attention mechanisms.

Integrating both CNNs and ViTs in a hybrid model harnesses the advantages of each, facilitating the efficient processing of both local and global features present in an image.

ii) Enhanced Generalization:

Vision Transformers (ViTs) typically demand substantial training data for effective learning, owing to their fully-connected design and absence of inherent biases, a characteristic inherent in CNNs.

Introducing CNN layers can enhance the model's ability to generalize, particularly with smaller datasets, by introducing these advantageous biases.

iii) Decreased Computational Complexity:

Vision Transformers (ViTs) often incur a greater computational burden, notably with larger images, as they process images as sequences of patches and compute global self-attention across all patches.

CNNs can alleviate this burden by condensing the image into a collection of higher-level features before feeding them into the Transformer layers.

This compression of sequence length can mitigate computational demands and memory requirements, thereby enhancing the model's efficiency.

iv) Improved Early-Stage Feature Extraction:

CNNs excel in swiftly processing raw pixels to identify fundamental features such as edges and textures, a task that may be more time-consuming for a Transformer to master independently.

Leveraging these pre-processed features can furnish the Transformer layers with a more informative foundation, potentially resulting in more intricate and nuanced feature representations

v) Versatility and Adjustability

Hybrid models exhibit greater adaptability across diverse image data and tasks. For example, in domains where local features hold greater significance, such as medical imaging or fine-grained classification, the CNN's impact can be emphasized.

Conversely, in contexts requiring comprehension of broader contexts or larger scenes, such as image captioning or scene understanding, the Transformer's contribution becomes more prominent.

vi) Resilience to Variances

Transformers, when used alone, may occasionally prioritize global dependencies excessively, potentially overlooking local intricacies. CNNs serve to address this by ensuring effective capture of local patterns as well.

This fusion enhances the model's resilience to input data variations, such as alterations in scale, viewpoint, or context

4.1.2 Hybrid ViT-CNN architecture: image classification methodology

In the experimental setup the proposed model used a dataset of around 8000 images, which were pre-processed by resizing to 224x224 pixels. The experiment was conducted using TensorFlow library functions and Keras on Google Collab. The model underwent training for 50 epochs, employing the Adam optimizer with a learning rate set at 0.001 and a batch size of 32. The dataset contained 6 classes, including 5 rice diseases (Blast, Brown spot, Bacterial Sheath Blight, Falsemut, Tungro) and a sixth class for identifying healthy rice leaves.

Image normalization was applied to the input images to mitigate gradient propagation concerns. A variety of image processing techniques, including Erosion, Dilation, Opening, and Closing, were utilized to enhance regions with varying brightness levels. Positional Data Augmentation techniques

such as random crop, center crop, Random Vertical Flip, Random Rotation, Resize, and Random Affine were then executed. Following this, color Augmentation techniques such as Brightness, Contrast, and Saturation adjustment were applied. Lastly, the augmented images were divided into patches.

The Vision Transformer (ViT) model processes input images of rice leaf diseases by first dividing them into patches that are fixed in size and non-overlapping. These patches are then transformed into lower-dimensional vectors using learned linear projections, effectively turning them into input tokens for the subsequent Transformer model. To account for the spatial layout of the patches within the image, positional information is incorporated into the embeddings.

The patch embeddings, combined with positional encodings, are passed through a multi-layer Transformer architecture. The self-attention mechanism within the Transformer enables the model to capture long-range dependencies between patches and learn the spatial relationships within the image.

After the ViT layer, a CNN layer is applied to the output tensor. This CNN layer is configured with specific parameters, including the number of filters, a 3×3 kernel size, a stride of 1 for the convolution operation, and a dropout rate of 0.1. The dropout rate indicates that during training, 10% of the output neurons will be randomly set to zero as a regularization technique.

The output from the CNN layer will have a shape that depends on the input image and the convolution parameters. To produce the final predictions, a Multi-layer Perceptron (MLP) classification head is added at the end of the model. This classification head takes the output from the preceding layers and generates the predictions for the rice leaf disease classification task. The overall flow of the model is depicted in Figure 4.

CNNs are adept at extracting local hierarchical features by virtue of their convolutional nature. They specialize in comprehending spatial hierarchies within images, where finer details in upper layers rely on information from lower layers. In contrast, ViTs shine in capturing global dependencies and contextual information across the entirety of an image, facilitated by their self-attention mechanisms. A hybrid model seamlessly merges these strengths, enabling the efficient processing of both local and global image features.

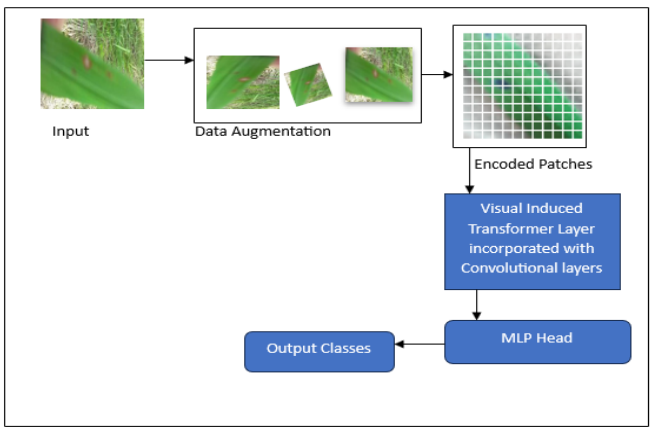

Figure 4. Dataflow of the hybrid ViT-CNN model

Vision Transformers (ViTs) typically demand a substantial volume of data for effective training, primarily because of their fully-connected architecture and absence of inherent inductive biases, which are naturally present in CNNs. The inclusion of CNN layers can enhance the model's ability to generalize, particularly when working with smaller datasets, as these layers introduce these advantageous biases.

Transformers operating in isolation may sometimes overly emphasize global dependencies, potentially overlooking local intricacies. The incorporation of CNNs serves to counterbalance this by ensuring the effective capture of local patterns. This synergy can bolster the model's resilience to variations in input data, including shifts in scale, perspective, or context.

4.2 Empirical results

The results of this research are estimated by Evaluation metrics like Accuracy, Top-5 accuracy and precision. Where Accuracy evaluates the classifier's ability to correctly classify the entire dataset, taking into account both positive and negative cases. In contrast, precision specifically assesses the amount of properly recognized positive samples amid all instances classified as positive by the classifier. It's a metric that provides insights into the classifier's accuracy in identifying positive cases [23]. Top-5 accuracy is a widely employed performance metric in multi-class classification, notably in domains like image recognition and natural language processing. It assesses the classifier's prediction accuracy by determining whether the correct class falls within the top 5 predicted classes for a given sample. This metric proves valuable when assessing models for tasks involving class uncertainty or when numerous correct answers are possible.

$$Accuracy = \frac{TN + TP}{TN + FN + TP + FP} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Top - 5\ Accuracy = \frac{(\text{The count of correct predictions within the top 5})}{\text{Total Number of Samples}} \tag{7}$$

(1). True Positive (TP): The instances that the classifier correctly identified as positive.

(2). True Negative (TN): The instances that the classifier correctly identified as negative.

(3). False Negative (FN): The instances that are actually positive but were incorrectly classified as negative by the classifier.

(4). False Positive (FP): The instances that are actually negative but were incorrectly classified as positive by the classifier.

Utilizing the proposed hybrid ViT-CNN model architecture, the model attains 100 percent accuracy and top-5 accuracy, along with a precision of 93.84 percent. Through this hybrid model, we have achieved satisfactory results, surpassing the performance of the latest transformer models in the identification of rice leaf diseases. This serves as proof that the synthetically extracted features exhibit stronger feature representation capabilities. The accuracy for the hybrid ViT-CNN model is presented in Figure 5 and Figure 6 gives a comparative chart of Accuracy, Precision and Top-5 accuracy of ViT-CNN Model.

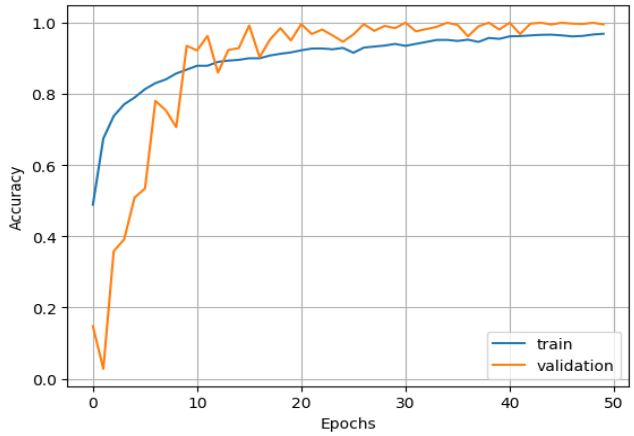

Figure 5. The accuracy for the hybrid ViT-CNN model

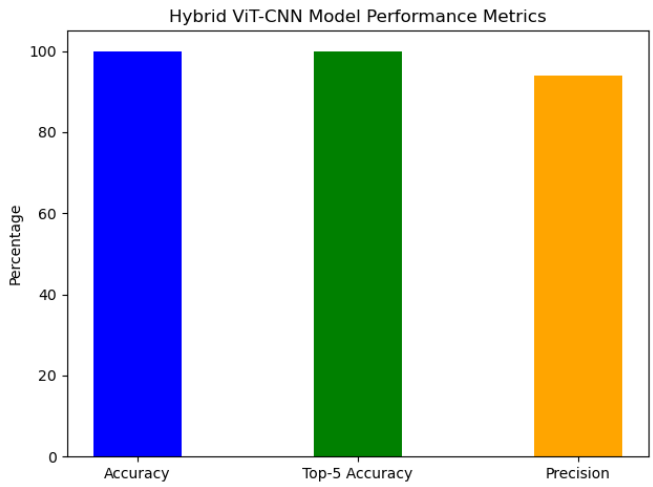

Figure 6. Comparative chart on evaluation metrics for hybrid ViT-CNN model

Table 2. Comparative results of Hybrid ViT-CNN Model with various base models

Models	Accuracy	Top-5 Accuracy	Precision
DenseNet121	94.19	98	93.75
ResNet50	72	98.1	82
InceptionV3	77	91.2	83.2
ViT	95.41	97	86.33
ViT_CNN Model	100	100	93.84

The proposed approach is pitted against conventional neural network models, including DenseNet121, ResNet50, InceptionV3 and ViT alone. The outcomes are presented in Table 2, revealing the favourable performance of the proposed method.

The hybrid approach (Vit-CNN) strives to merge the intricate, local feature recognition capabilities of CNNs with the holistic, contextual understanding provided by ViTs. This fusion aims to harmonize the inductive biases and efficiency inherent in CNNs with the expressive power and scalability of ViTs. The ultimate goal is to create models, similar to DenseNet121, ResNet50, InceptionV3, and EfcientNet, that strike a balance between potency and practicality, exhibiting effectiveness across a diverse array of vision-related tasks.

The Vit-CNN hybrid approach seeks to integrate the nuanced, local feature recognition strengths found in CNNs with the comprehensive, contextual comprehension offered by Vision Transformers (ViTs), achieving a remarkable 100 percent accuracy. This synthesis aims to bring together the inherent inductive biases and efficiency of CNNs with the expressive capabilities and scalability of ViTs. Figure 7 gives a comparative chart on results for Hybrid ViT-CNN Model with various base models.

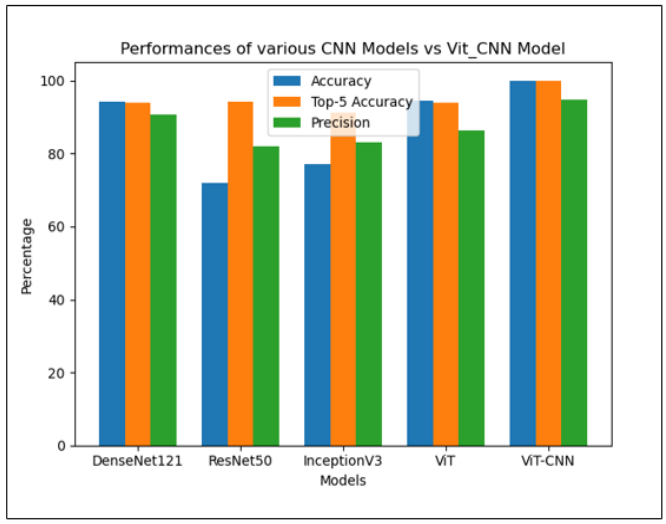

Figure 7. Evaluation metrics for hybrid ViT-CNN model verses various deep learning models

The dataset used for training the Hybrid ViT-CNN model consisted of images obtained from diverse sources, including various repositories and real-world data. These images, captured in JPG format, presented a wide range of backgrounds, including other plant leaves, field grass, and varying soil colours. Environmental noises and distortions were inevitable due to the open-field captures, and occasional inclusions such as photographers' fingers were observed. Furthermore, inconsistent lighting caused by fluctuating weather conditions added complexity to the dataset.

Despite these challenges, the Hybrid ViT-CNN model, trained on images with diverse lighting conditions and backgrounds, demonstrated efficiency in accurately identifying real-world images, showcasing its potential for generalizability across different environmental settings.

4.3 Enhancing agricultural decision support with ViT-CNN model: Real world applications

i) Pre-symptomatic Disease Identification:

ViT-CNN effectively recognizes prevalent rice leaf diseases in images taken with smartphones or drones. This timely identification empowers farmers to promptly implement necessary measures.

ii) Pathological Evaluation:

ViT-CNN assists farmers in prioritizing management strategies and efficiently allocating resources by concentrating interventions on regions experiencing the greatest disease pressure.

iii) Precision Treatment Application:

When integrated with precision spraying systems, ViT-CNN directs the precise application of fungicides or biological control agents solely to the regions affected by diseases.

iv) Pathogen Surveillance and Oversight:

ViT-CNN provides continuous monitoring of rice fields for indications of disease progression, delivering instant updates to farmers via mobile applications or web-based platforms.

5. CONCLUSION AND FUTURE WORK

The early identification of rice leaf diseases is crucial for mitigating widespread outbreaks and substantial economic losses. This study explores five primary ailments affecting rice crops: Blast, Brown Spot, Tungro, Falsemut, and Bacterial Sheath Blight. The images utilized in this research maintain consistent backgrounds and lighting conditions, contributing to high predictive accuracy. However, this uniformity may not fully capture the complexities present in real farming environments, where diverse backgrounds and interference are common. Therefore, focusing on data from authentic farming settings can significantly enhance the model's robustness. By integrating Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) within a hybrid architecture, this unified model capitalizes on the complementary strengths of both, offering various advantages. This synergistic combination improves the model's resilience to variations in input data, such as changes in scale, perspective, or context. Through the adoption of the proposed hybrid ViT-CNN model architecture, the model achieves exceptional results, boasting 100 percent accuracy and top-5 accuracy, alongside a precision rate of 93.84 percent. Through the utilization of this hybrid model, we have obtained satisfactory results that outperform the performance of the latest transformer models in the identification of rice leaf diseases. In the future, ViT-CNN could be expanded to accurately recognize and categorize a multitude of crop diseases. Through integration with drone-based monitoring systems, farmers could gain access to real-time insights, empowering them to make well-informed decisions.

ACKNOWLEDGMENT

We are grateful to “SRM Institute of Science and Technology”, Kattankulathur, India for their motivation and immense support rendered for this research.

REFERENCES

[1] United Nations. Peace, dignity and equality on a healthy planet. https://www.un.org/en/global-issues/population#:~:text=The%20world%20in%202100,surrounding%20these%20latest%20population%20projections.

[2] Chen, J., Zhang, D., Zeb, A., Nanehkaran, Y.A. (2021). Identification of rice plant diseases using lightweight attention networks. Expert Systems with Applications, 169: 114514. https://doi.org/10.1016/j.eswa.2020.114514

[3] Deepika M. (2023). 11 Paddy diseases: Understanding the causes, symptoms, and treatment options. Bighaat. https://kisanvedika.bighaat.com/crop/major-diseases-of-paddy/.

[4] Li, L., Zhang, S., Wang, B. (2021). Plant disease detection and classification by deep learning-A review. IEEE Access, 9: 56683-56698. https://doi.org/10.1109/ACCESS.2021.3069646

[5] Deng, R., Tao, M., Xing, H., Yang, X., Liao, K., Qi, L. (2021). Automatic diagnosis of rice diseases using deep learning. Frontiers in Plant Science, 12: 701038. https://doi.org/10.3389/fpls.2021.701038

[6] Haridasan, A., Thomas, J., Raj, E.D. (2023). Deep learning system for paddy plant disease detection and classification. Environmental Monitoring and Assessment, 195(1): 120. https://doi.org/10.1007/s10661-022-10656-x

[7] Sharma, R., Singh, A., Jhanjhi, N.Z., Masud, M., Jaha, E.S., Verma, S. (2022). Plant disease diagnosis and image classification using deep learning. Computers, Materials & Continua, 71(2). http://dx.doi.org/10.32604/cmc.2022.020017

[8] Latif, G., Abdelhamid, S.E., Mallouhy, R.E., Alghazo, J., Kazimi, Z.A. (2022). Deep learning utilization in agriculture: Detection of rice plant diseases using an improved CNN model. Plants, 11(17): 2230. https://doi.org/10.3390/plants11172230

[9] Upadhyay, S.K., Kumar, A. (2022). A novel approach for rice plant diseases classification with deep convolutional neural network. International Journal of Information Technology, 14(1): 185-199. https://doi.org/10.1007/s41870-021-00817-5

[10] Chen, J., Zeb, A., Nanehkaran, Y.A., Zhang, D. (2023). Stacking ensemble model of deep learning for plant disease recognition. Journal of Ambient Intelligence and Humanized Computing, 14(9): 12359-12372. https://doi.org/10.1007/s12652-022-04334-6

[11] Zhou, C., Zhong, Y., Zhou, S., Song, J., Xiang, W. (2023). Rice leaf disease identification by residual-distilled transformer. Engineering Applications of Artificial Intelligence, 121: 106020. https://doi.org/10.1016/j.engappai.2023.106020

[12] Sudhesh, K.M., Sowmya, V., Kurian, S., Sikha, O.K. (2023). AI based rice leaf disease identification enhanced by Dynamic Mode Decomposition. Engineering Applications of Artificial Intelligence, 120: 105836. https://doi.org/10.1016/j.engappai.2023.105836

[13] Upadhyay, S.K., Kumar, A. (2021). Early-Stage brown spot disease recognition in paddy using image processing and deep learning techniques. Traitement du Signal, 38(6): 1755-1766. https://doi.org/10.18280/ts.380619

[14] Aristan, T., Kusuma, G.P. (2023). Evaluation of CNN models in identifying plant diseases on a mobile device. Revue d'Intelligence Artificielle, 37(2): 441-449. https://doi.org/10.18280/ria.370221

[15] Pothen, M.E., Pai, M.L. (2020). Detection of rice leaf diseases using image processing. In 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, pp. 424-430. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-00080

[16] Sethy, P.K., Barpanda, N.K., Rath, A.K., Behera, S.K. (2020). Image processing techniques for diagnosing rice plant disease: A survey. Procedia Computer Science, 167: 516-530. https://doi.org/10.1016/j.procs.2020.03.308

[17] Tugrul, B., Elfatimi, E., Eryigit, R. (2022). Convolutional neural networks in detection of plant leaf diseases: A review. Agriculture, 12(8): 1192. https://doi.org/10.3390/agriculture12081192

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

[19] Graves, A. (2013). Generating sequences with recurrent neural networks. arXiv Preprint arXiv: 1308.0850. https://doi.org/10.48550/arXiv.1308.0850

[20] Lin, T., Wang, Y., Liu, X., Qiu, X. (2022). A survey of transformers. AI Open, 3: 111-132. https://doi.org/10.1016/j.aiopen.2022.10.001

[21] Yang, Y., Jiao, L., Liu, X., Liu, F., Yang, S., Feng, Z., Tang, X. (2022). Transformers meet visual learning understanding: A comprehensive review. arXiv Preprint arXiv: 2203.12944. https://doi.org/10.48550/arXiv.2203.12944

[22] Yu, J., Li, J., Yu, Z., Huang, Q. (2019). Multimodal transformer with multi-view visual representation for image captioning. IEEE Transactions on Circuits and Systems for Video Technology, 30(12): 4467-4480. https://doi.org/10.1109/TCSVT.2019.2947482

[23] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. arXiv Preprint arXiv: 2010.11929. https://doi.org/10.48550/arXiv.2010.11929