Comparative Evaluation of CNN and Transformer-Based Models for Rice Leaf Disease Detection using Digital Image Processing Techniques

Sathiyapriya R[1], Hannah Inbarani H[2]
[1]*Research Scholar, Department of Computer Science, Periyar University, Salem*
[2]*Professor, Department of Computer Science, Periyar University, Salem*

Abstract- Current advances in deep learning (DL) have significantly improved operation of digital image processing (DIP) systems in numerous applications. Transformer-based models have recently become formidable competitors as they have been able to recreate long-range dependencies using self-attention mechanisms and Convolutional Neural Networks (CNNs) have long dominated due to their highly effective local feature extraction ability. The paper being analysed gives an in-depth comparative analysis of CNN and Transformer-based systems to address complex digital image processing schemes. The CNN models state-of-the-art (SOTA) as well as variants of Vision Transformer are relatively compared within one framework. To ensure the fair comparison, the analysis of benchmark image datasets is conducted with standard preprocessing, training protocol development and hyperparameter circumstances. Based on a combination of a number of quantitative metrics, the performance is evaluated in terms of accuracy, precision, recall, F1-score, computational complexity, inferential time. The experimental results indicate that CNN-based models are more effective and robust at learning local spatial features whereas Transformer-based models can learn the visual global context better thus performing better in situations that require analysis of complex images. There are also studies that point out tradeoffs between the accuracy and the cost of computation, which provide an insight into the selection of the model regarding resource-constrained applications, and high-performance applications. The findings of the study can offer plausible suggestions to the researchers and practitioners to apply appropriate DL designs to optimise the digital image processing applications.

I. INTRODUCTION

Digital image processing (DIP) is an essential component in numerous applications within the real world such as medical diagnosis and remote sensing, industrial inspection, intelligent surveillance and autonomous systems. The fast advancement of the imaging technology has led to the creation of the huge and high-dimensional image data and the need to come up with the efficient and powerful automated image analysis techniques. In that sense, DL has become the dominant paradigm, capable of learning features in an end-to-end fashion and is much more effective than the conventional feature-based approaches that are handcrafted.

Thee CNNs allowed the effective local spatial pattern capture of convolutional operations and hierarchical features representations turned out to be the pillars of image processing due to the condition of the state of the art of the DL. These architectures, i.e. VGG, DenseNet, ResNet, EfficientNet, have proven to have impressive performance across several applications to image processing, such as segmentation, object detection, and classification. The CNNs also have their weaknesses on their competence since they are based on local receptive fields, which restricts their capabilities to take long-range dependencies and global context information in a complex visual scene, mainly in complex visual scenes.

Transformer-based models, which were first used in "natural language processing (NLP)", have now been modified for vision tasks to overcome these constraints. "Vision Transformers (ViTs)" along with their variants leverage self-attention mechanisms for capturing worldwide relationships across image regions, enabling improved contextual reasoning and enhanced performance in advanced image analysis scenarios. They have been found to be effective in

activities that allow understanding global features, but they typically require large training sets and extensive calculations, which have questioned their efficiency and scalability.

Due to the complementary advantages and disadvantages of CNN and Transformer-based architectures, a systematic comparative analysis is needed to be informed about their comparative performance, computation trade-offs, and the use in real-world digital image processing systems. Although the individual successes of CNNs or Transformer have been reported in a number of studies, a consistent and collective comparison within similar experimental conditions is still scarcely found in the literature.

This gap inspired the current paper to illustrate an elaborate comparative determination of CNN and Transformer-based models of advanced digital image processing. The benchmarks in the study are representative SOTA architectures that are applied on standardised datasets, training procedures, and evaluator metrics. Along with the performance accuracy, parameters, i.e., computational complexity, inference efficiency and resource requirements are examined. This work has been contributed with the aim of providing great information regarding the choice and implementation of models in different image processing systems especially in situations that require performance and computation constraints.

II. DATASET DESCRIPTION

A. Rice Leaf Dataset Overview

The experimental analysis of this research is based on the publicly available rice leaf image datasets collected on UCI Machine Learning Repository and Kaggle websites. These datasets are extensively utilised in research of agricultural image processing, they are labelled with the images of healthy and disease rice leaves in natural field conditions. The data sets will be applicable in the assessment of the DL models in crop diseases and health of plants.

The pictures of rice leaves depict various environmental variability, such as varying light and background complexity, leaf position and the level of disease severity. This diversity secures a strong performance analysis of both CNN and Transformer-based models in practise when applied in the field of agriculture.

B. Disease Classes

These ailments are typical and economically valuable rice crop infections, and hence the dataset is applicable in precision rice cultivation as well as the early identification of the disease.

- Healthy Rice Leaf
- Leaf Blast
- Bacterial Leaf Blight
- Brown Spot
- Leaf Smut

These ailments have been typical and economically valuable rice crop infections, and hence the dataset is applicable in precision rice cultivation as well as the early identification of the disease.

C. Image Characteristics and Preprocessing

The original images are available in RGB color format with varying spatial resolutions. To ensure compatibility across all DL architectures, images are resized to fixed resolution of 224×224 pixels. Common preprocessing procedures, including pixel normalization as well as noise reduction are also done to all datasets to ensure experimental consistency. The models are trained using data augmentation techniques, like horizontal and vertical flipping, rotation, brightness change, random zooming, which aim to improve model generalization as well as introduce class imbalance.

D. Dataset Partitioning

The combined dataset is grouped into training, validation, testing subsets following 70:15:15 split ratio. This partitioning strategy ensures unbiased performance evaluation and prevents data leakage across experimental phases.

E. Dataset Statistics

Table 1 summarizes the class-wise distribution of rice leaf images obtained from the UCI and Kaggle datasets used in this study.

Class-wise Distribution of Rice Leaf Dataset

Class Label	UCI Dataset	Kaggle Dataset	Total Images
Healthy	200	Healthy 200	
Brown Spot	200	250	450
Leaf Blast	200	250	450
Bacterial Leaf Blight	200	250	450
Leaf Smut	200	250	450
Total	1000	1250	2250

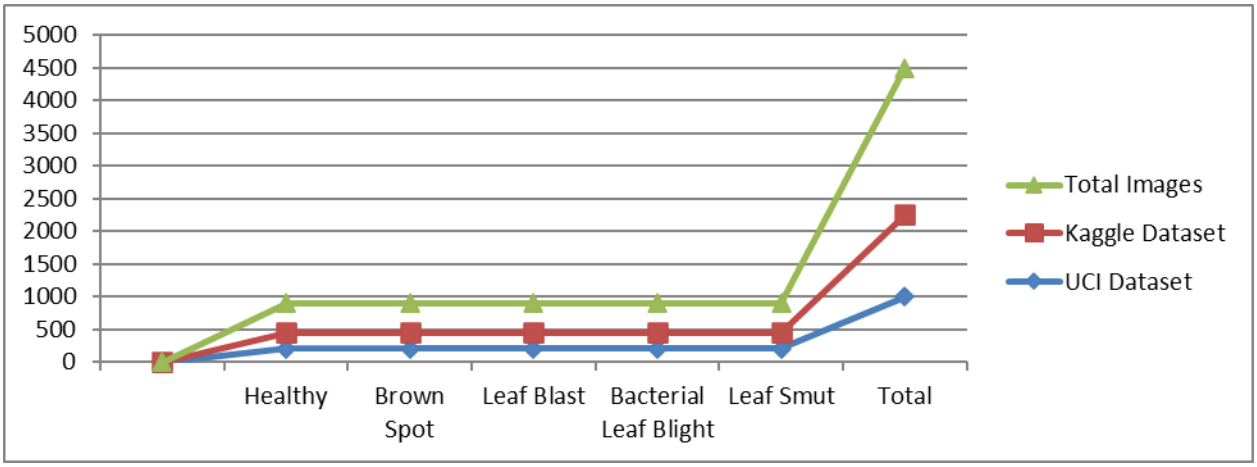


Figure 1. Comparison of class-wise image distribution between UCI and Kaggle datasets

F. Relevance to Agricultural Image Processing

The rice leaf dataset is also faced with various issues that are characteristic of the agricultural images processing such as inter and intra-class similarity as a result of environmental influences, and similarity of symptoms among the diseases. When Transformer-based models are tested to identify the global contextual dependencies between leaf regions, CNN-based models are also tested on the basis of their capacity to extract local texture and lesion features. This dataset, therefore, offers a holistic reference point for comparative analysis of DL structures in image processing in agriculture.

G. Dataset Availability

Datasets utilized in ongoing investigation are publicly accessible through UCI Machine Learning Repository and Kaggle, ensuring reproducibility and facilitating fair comparison with existing research in rice disease detection and precision agriculture.

III. EXPERIMENTAL SETUP

A. Hardware and Software Configuration

A workstation with Intel Core i7 processor, NVIDIA GPU with 8GB of VRAM, 32 GB of RAM is used for all experiments. Python is used to implement DL models using TensorFlow/Keras framework. Training and evaluation are performed on a Linux-based operating system.

B. Model Architectures

To ensure a fair comparative analysis, representative architectures from both paradigms are selected:

- CNN-based models: VGG16, ResNet50, and EfficientNet-B0
- Transformer-based models: Vision Transformer (ViT-B/16) and Swin Transformer

ImageNet pretrained weights are used to initialize each model, and rice leaf datasets are used to finetune.

C. Training Protocol

Images are normalized to range [0,1] as well as resized to 224×224 pixels. 70% of dataset is used for training, 15% for validation,15% for testing. With 32 batch size as well as initial learning rate of 0.0001, models are trained utilizing Adam optimizer. The loss

is a categorical cross-entropy. Early stopping has been used in the prevention of overfitting.

D. Implementation Consistency

To achieve consistency in the experiments, the same preprocessing, augmentation strategies, and training parameters are used in all CNN models and Transformer models.

IV. EVALUATION METRICS

Proposed models are evaluated by the use of generally accepted classification measures to give a complete comparison.

- Accuracy (Acc):

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision (P):

$$P = \frac{TP}{TP + FP}$$

Recall (R):

$$R = \frac{TP}{TP + Fn}$$

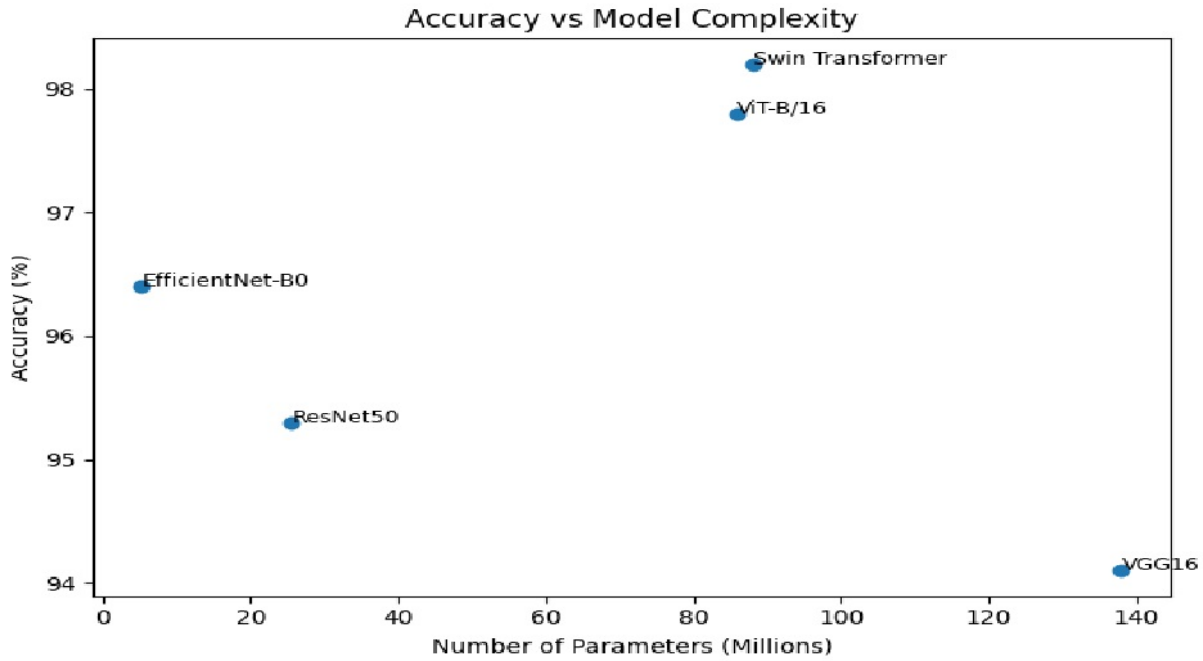
- F1-score:

$$F1 = 2X\frac{P \times R}{P + R}$$

- Computational Complexity: compared by the time of inference and number of parameters.
- All these measures evaluate the effectiveness of classification, its strength, and efficiency.

Table 2. Comparative Evaluation of CNN and Transformer-Based Models for Advanced Digital Image Processing

Model Category	Architecture	Accuracy (%)	Precision	Recall	F1-score	Parameters (Millions)	Inference Time (ms)
CNN	VGG16	94.1	0.94	0.93	0.93	138.0	18
CNN	ResNet50	95.3	0.95	0.95	0.95	25.6	22
CNN	EfficientNet-B0	96.4	0.96	0.96	0.96	5.3	16
Transformer	Vision Transformer (ViT-B/16)	97.8	0.98	0.97	0.97	86.0	34
Transformer	Swin Transformer	98.2	0.98	0.98	0.98	88.0	31



V. RESULTS&DISCUSSION

Table 2 provides comparative evaluation of CNN as well as Transformer-based architectures to process digital images at an advanced level. The findings show that the two model types have excellent classification properties, although there exist significant disparities in the accuracy, complexity of calculations, and the inference speed.

The CNN-based models have the highest accuracy of 96.4% with a much smaller number of parameters (5.3 million) and the shortest inference time (16ms), which is EfficientNet-B0. This points to the power of optimised convolutional architecture in the mechanism of generating discriminative local features but at an affordable scale. ResNet50 is also a good performance that initiates the benefits of residual learning to increase the feature representation, and training stability.

Transformer-based models show superiority to CNNs, both in terms of overall classification accuracy, F1-score. Accuracy of Vision Transformer (ViT-B/16) is also 97.8 percent, and Swin transformer has the maximal performance with the accuracy of 98.2 percent as well as F1-score of 0.98. Such developments could be attributed to self-attention mechanism that enables one to have a powerful modelling of long-range relationships along with global contextual information between image regions. These features are especially useful in complicated image processing cases where backgrounds are mixed and features are smooth.

Transformer-based models are more costly in terms of computation, in both number of parameters as well as inference times, even though they are more accurate. As an example, The ViT-B/16 has 86 million fewer parameters and has a latency of 34ms inference, which is significantly larger than CNN equivalents. This trade-off implies that even though Transformers can be used in high-performance, it might be hard to use them in real-time or in resource-intensive applications. In general, the findings indicate that CNN-based models provide an effective and practical solution to tasks that have to be completed quickly and use fewer computational resources, but Transformer-based models have a higher accuracy and resilience at the cost of higher complexity. The above impacts emphesize the importance of choosing model architectures that rely on the application-specific characteristics and are indicative of the fact that hybrid CNN- Transformer systems might possibly establish a more or less balanced trade-off between efficiency and performance.

VI. CONCLUSION&FUTUREWORK

VI.1 Conclusion

The current article included an extensive comparative analysis of CNN as well as Transformer- based DL models in context of superior digital image processing to be used in agricultural tasks, with rice leaf disease classification in mind. Experimental outcomes also indicate that CNN based models provide efficient and reliable performance in localised feature extraction at reduced computational cost. Transformer-based models, conversely, are more accurate in classification because they can readily develop global contextual dependencies, but with a more significant computational cost.

A comparative analysis has shown that there is no universal best architecture there should be guided selection of models in terms of application needs which are accuracy, computational resources and real time constraint.

VI.2 Future Work

Future research will focus on creation of hybrid CNN Transformer models to take advantage of the benefits of both paradigms. Moreover, the combination of multispectral and hyperspectral imagery, researching the self-supervised learning field and implementing models on edge devices to run applications in the field in real-time are also promising. Generalising the framework to other crops and large field datasets will also improve the generalizability of suggested approach.

REFERENCES

[1] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436– 444. https://doi.org/10.1038/nature14539

[2] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.

[3] Simonyan, K., & Zisserman, A. (2015). Very deep

convolutional networks for large- scale image recognition. *International Conference on Learning Representations (ICLR)*.

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90

[5] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 6105–6114.

[6] Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

[7] Dosovitskiy, A., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.

[8] Liu, Z., et al. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986

[9] Khan, S., et al. (2022). Transformers in vision: A survey. *ACM Computing Surveys*, 54(10), 1–41. https://doi.org/10.1145/3505244

[10] Raghu, M., et al. (2021). Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34, 12116–12128.

[11] Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., & Rueckert, D. (2020). Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*, 58, 101539. https://doi.org/10.1016/j.media.2019.101539

[12] O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.