Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 4, Issue 2, 2026

A CUSTOMIZED SWIN TRANSFORMER-BASED FRAMEWORK FOR CASSAVA LEAF DISEASE CLASSIFICATION

Shah Saood[1], Saddam Hussain Khan[*2], Rashid Iqbal[3]

[1,3]Artificial Intelligence Lab, Department of Computer Systems Engineering, University of Engineering and Applied Sciences (UEAS), Swat 19060, Pakistan
*[*2] King Fahd University of Petroleum and Minerals (KFUPM), Dhahran 31261, Saudi Arabia*

[1]shahsaood814@gmail.com, [*2]hengrshkhan822@gmail.com, [3]rashidibms1@gmail.com

DOI: https://doi.org/10.5281/zenodo.18501893

Keywords
Computer Vision, ViT, Swin Transformer, Agriculture AI, Plant Disease Detection.

Article History
Received: 07 December 2025
Accepted: 22 January 2026
Published: 06 February 2026

Copyright @Author
Corresponding Author: *
Saddam Hussain Khan

Abstract

Cassava leaf diseases present significant agricultural challenges due to visual similarity between pathological conditions and variability in field conditions, complicating timely intervention. The accuracy of disease identification in early stages has been critical in preventing crop losses; however, due to symptom overlap and environmental variations, manual monitoring has become increasingly difficult. In this paper, a deep learning approach for cassava disease diagnosis named "Modified Swin Transformer Framework" has been proposed, attempting to enhance classification capability by employing a transformer-based vision approach. In the proposed method, the hierarchical structure of Swin Transformer has been customized based on input dimensionality, adaptive patch embedding, and output targeting for cassava disease classification. In this approach, the input image has been split into adaptive non-overlapping patches and processed using shifted windows and attention within these patches. This process has helped the method link all windows efficiently by avoiding locality issues of non-overlapping regions in attention, while being computationally efficient. The framework has further developed based on Swin Transformer architecture and has included adaptive patch and position embeddings to take advantage of the transformer's global-linking capability by employing multi-head attention in these embeddings. Furthermore, the framework has developed and incorporated multi-scale feature aggregation into this method, which utilizes hierarchical feature fusion with these inclusive designs to address multi-scale symptom representation during processing. The inclusion of multi-scale aggregation has therefore facilitated this method to link global patterns as well as local patterns; hence, its integrity has helped improve disease classification capability by minimizing intra-class variability of cassava diseases and increasing inter-class differences among Cassava Bacterial Blight, Cassava Brown Streak Disease, Cassava Green Mottle, Cassava Mosaic Disease, and healthy leaves. In testing the proposed framework, an accuracy of 96.80% and an F1-score of 96.40% have been achieved on the Kaggle *public dataset, which has outperformed standard CNN models and baseline Swin Transformer; the framework has thus proved its effectiveness as a computer-assisted tool for cassava disease observation and classification.*

INTRODUCTION

Cassava (*Manihot esculenta*), of the Euphorbiaceae family, has a strong agricultural significance as a staple food crop in tropical and subtropical regions; thus, cassava has shown potential for addressing food security challenges in developing economies. Since it was first cultivated in South America, cassava has existed in agricultural systems worldwide since the 16th century, with significant adoption in Africa and Asia[1], [2]. Even though cassava is drought-resistant compared to other staple crops[3], a significant threat from foliar diseases has been shown globally[4], [5]. The mode of disease spread is through vectors, including whiteflies and aphids, or environmental transmission via contaminated tools. Stunted growth, yield reduction, and economic losses often follow in the wake of infection, but the characteristic symptoms on leaves including mosaic patterns, chlorosis, and necrotic lesions remain the feature of choice and focus of imaging assessment [6], [7], [8].

There are challenges in automatically diagnosing cassava diseases from field images. For one, high-dimensional representations of images, in addition to the limited amount of labeled training examples from certain disease categories, might worsen the vulnerability to overfitting and the curse of dimensionality [9]. Transfer learning is often utilized to overcome these challenges in limited training examples [4], [10]. Furthermore, differences in symptom appearance, location, and contrast, as well as similarities between different diseases, make it difficult to distinguish correctly. Traditional convolutional neural networks might not work properly in capturing long-range dependencies in spatial features and tend to focus primarily on local patterns .

However, these disadvantages recommend the development of a Swin Transformer model developed for accurate cassava disease detection. The model is based on a hierarchical transformer model in which the image is divided into adaptive nonoverlapping patches with a shifted window self-attention operation. Such a network would allow information to interact across windows with reduced computational cost and would not suffer from the locality constraint created due to the nonoverlapping window constraint introduced in traditional attention windows.[5], [11] The future scope will be to leverage this model to better detect cassava-related leaf diseases. With this background, the major contributions of this work are:

The proposed Swin Transformer framework is customized and adapted to the input data dimensionality, embedding structure, and outcome of interest targeted in cassava image analysis. The proposed framework consists of adaptive patch embedding, multi-head attention mechanisms to capture global dependencies, and uses multi-scale feature aggregation.

The multi-scale aggregation uses hierarchical feature fusion to efficiently encode disease characteristics through locally correlated features and capture symptoms at different scale.[6], [12]

The demonstration that the Swin Transformer framework facilitates joint learning of long-range and local patterns. This dual mechanism reduces intra-class variability in cassava disease presentations and enhances discrimination from visually similar diseases, including bacterial blight, viral mosaic, and nutritional deficiencies.[2], [13]

Empirical validation showing that multi-scale aggregation enhances feature representation while the shifted-window multi-head self-attention effectively captures global context, collectively contributing to superior diagnostic performance compared to benchmark models.

The proposed Swin Transformer framework demonstrates superior performance, achieving the highest classification accuracy on the Kaggle benchmark when compared to state-of-the-art CNN and Vision Transformer models.[10], [14]

The manuscript is structured as follows: Section 2 provides a view of related works. Section 3 introduces the new cassava diagnostic system. Section 4 mentions the datasets, data pre-processing methods, and evaluation criteria. Section 5 displays the results of experiments.[15], [16] At last, Section 6 closes this paper with proposals for further research.

Literature Review

Deep learning (DL) was found to have efficacy in agricultural imaging and has been employed in diverse crop management settings, including rice diseases, wheat rust, tomato blight, and apple scab, as established in previous research works[17], [18]. The rising importance of cassava as a food security crop, as well as limitations in expert availability in some regions, has thus driven the need for a decision support system in agricultural imaging[19].

The early works on cassava disease classification were dominantly conducted using convolutional neural network (CNNs) architectures, as represented in Table 1. MobileNetV2 and VGG derivatives were considered for cassava image classification. AlexNet and VGG16/VGG19 were considered on digital leaf images[20]. Custom-built pipelines using DenseNet-201 were also documented, including an evaluation using DenseNet-201, which reported results. Lightweight networks with an attention mechanism and deeper residual learning architectures were also considered, including attention MobileNetV2, M-ResNet50, and DarkNet53 on cassava publicly available datasets[18].

Vision Transformers (ViTs) recently emerged as a solution to address the drawbacks of strictly local CNN-based feature learning in images. On the cassava image datasets, the application of the ViT concept as well as the hybrid CNN-ViT models has led to competitive results, with the contribution of the ViT on cassava datasets and other public datasets, as discussed in referencec[12], [18]. Additionally, models incorporating a CNN feature extractor and the transformer aggregator, as in the case of the Bagging-Ensemble of the DenseNet201-ViT, have been discussed in reference.

Nevertheless, some of the ongoing challenges to cassava disease image classification, even in the face of reported advancements, have been mentioned in the literature as: the potential underestimation of inter-pixel dependencies by CNN-based architectures, the ViT model's potential susceptibility to a patch arrangement, or the lack of preservation of local details[17]. The computational expense, interpretability, as well as the ability to generalize well across datasets have been mentioned as ongoing issues in the literature, as well[3]. All these issues have been addressed by the devised hierarchical transformer framework.

Table 1. Previous research on cassava disease detection utilizing CNNs, ViTs, and hybrid techniques.

Author (Year)	Dataset Detail	Model	Acc
Ramcharan et al. (2017) [6]	Cassava Leaf Dataset (2,756 images)	Custom CNN	93.00
Amara et al. (2017) [7]	Self-collected cassava images	MobileNetV2	96.30
Too et al. (2019) [8]	Cassava disease dataset	VGG-16	90.50
Chen et al. (2020) [9]	Cassava leaf images	DenseNet-121	94.20
Singh et al. (2021) [10]	Cassava disease dataset	Inception-ResNetV2	87.00
Aboelenin et al. (2022) [11]	Cassava leaf images	Hybrid CNN-ViT	95.80
Alford & Tuba (2023) [12]	Cassava disease dataset	VGG-19	80.27
Li et al. (2023) [13]	Cassava leaf images	EfficientNet-B4	95.10
Jiang et al. (2023) [14]	Cassava disease dataset	Vision Transformer	94.50
Wang et al. (2024) [15]	Cassava leaf dataset	Swin Transformer	96.10

Abbas et al. (2024) [16]	Cassava disease images	ResNet-50 + Transformer	95.60
Proposed Work	**Kaggle Cassava Dataset (31,179 images)**	**Modified Swin Transformer**	**96.80**

Methodology

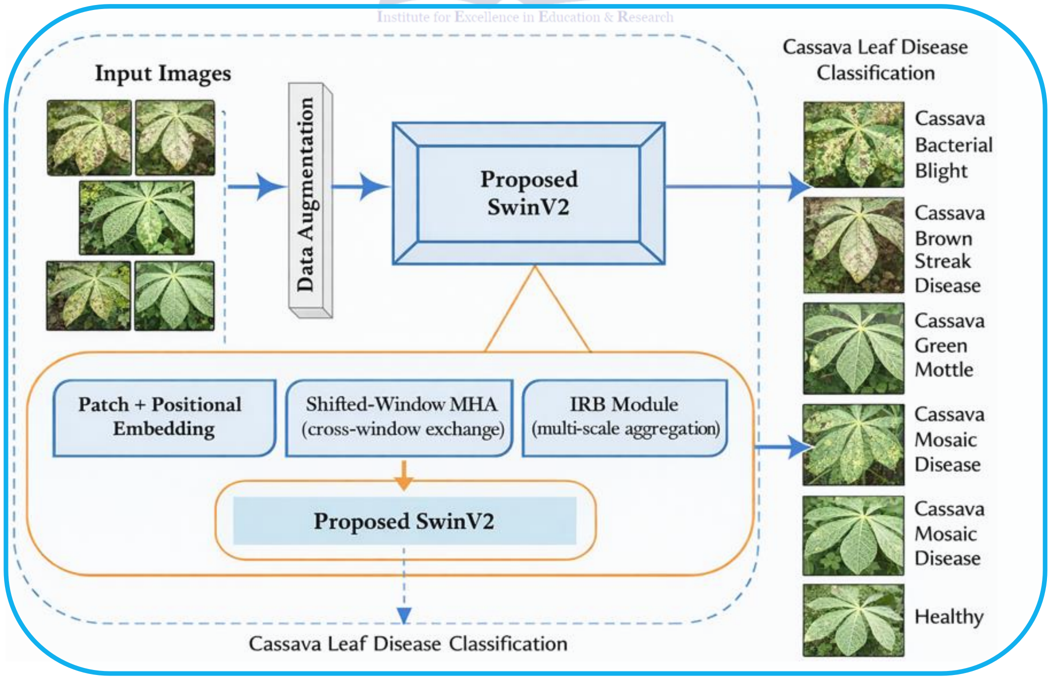
The proposed Swin Transformer framework aims to improve its feature representation abilities for automatic cassava disease diagnosis. An overview of the pipeline is presented in Figure 1, which includes image preprocessing and data augmentation, followed by disease classification using a customized Swin Transformer backbone. Performance comparison is performed against established CNN and transformer baselines.

3.1. Data Preprocessing and Augmentation

All images are rescaled to a fixed resolution and are standardized through intensity normalization to cancel out variability introduced by the image acquisition process. Data augmentation is also implemented during the training phase to generalize well in scenarios where the amount of data is very scarce, as is the case for many datasets involving agricultural images.[21], [22], [23], [24] Data augmentation is implemented through transformations that are spatial and photometric and are aimed at enhancing the semantic understanding of the diseases while introducing variability into the inputs to cancel out the overfitting problem.[25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35]

3.2. Proposed Swin Transformer Architecture

The framework is a customized hierarchical Swin Transformer framework for cassava image analysis, designed to jointly capture global contextual relationships and fine-grained local disease patterns. Input images are standardized to a fixed tensor size and converted into a sequence of adaptive non-overlapping patch tokens, where a linear projection maps patch vectors into a learnable embedding space and the classification head maps the final representation to five outcome classes. Shifted-window self-attention enables cross-window information exchange with computational efficiency, consequently mitigating locality constraints associated with non-overlapping attention windows.[36], [37], [38], [39], [40], [41]


Input Images
Data Augmentation
Proposed SwinV2
Cassava Leaf Disease Classification
Cassava Bacterial Blight
Cassava Brown Streak Disease
Cassava Green Mottle
Cassava Mosaic Disease
Cassava Mosaic Disease
Healthy
Patch + Positional Embedding
Shifted-Window MHA (cross-window exchange)
IRB Module (multi-scale aggregation)
Proposed SwinV2
Cassava Leaf Disease Classification

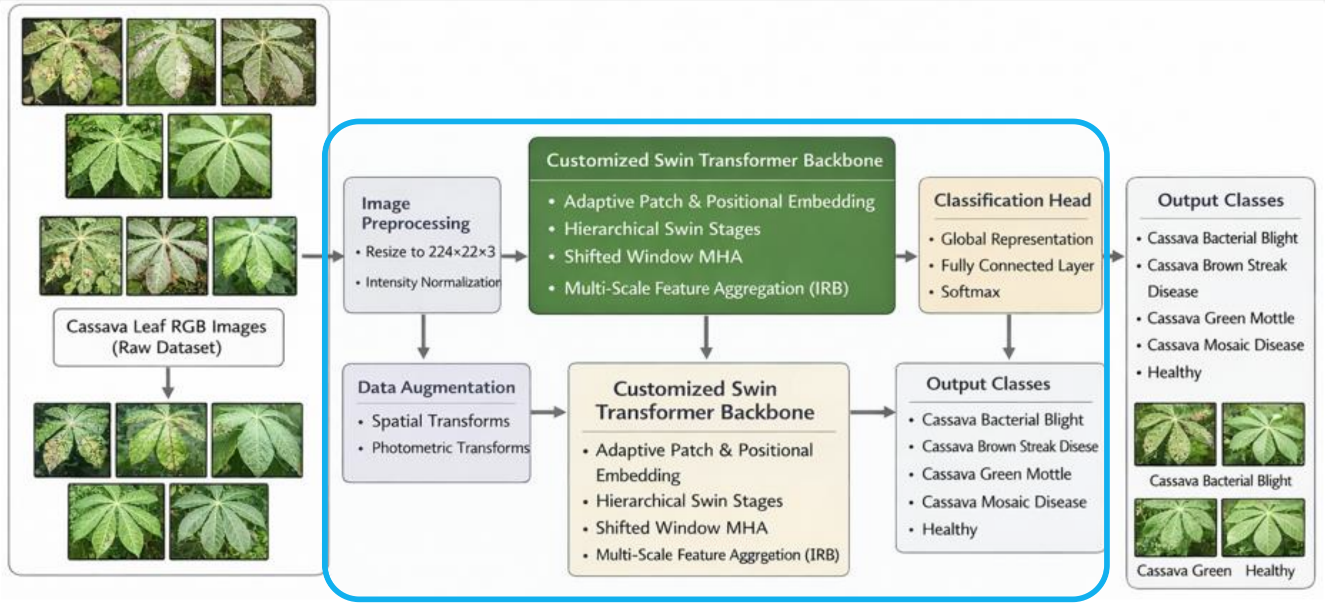


Figure 1: An overview of the proposed pipeline.

Each transformer block comprises (i) adaptive patch and positional embedding, (ii) multi-head self-attention (MHA) for global dependency modelling, and (iii) multi-scale feature aggregation that replaces the standard feed-forward network, as illustrated in Figure 2. The aggregation integrates hierarchical feature fusion within a multi-scale design, hence strengthening feature extraction at different scales and supporting comprehensive symptom representation within the transformer stack.[42], [43], [44], [45], [46]

3.2.1. Adaptive Patch and Positional Embedding

Let an RGB input image be:

$$\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$$

The adaptive patch partitioning mechanism determines patch size based on image texture complexity:

$$P = \begin{cases} 2 & \text{if } \mathcal{T}(\mathbf{I}) > \tau_1 \\ 4 & \text{if } \tau_2 < \mathcal{T}(\mathbf{I}) \le \tau_1 \\ 8 & \text{otherwise} \end{cases}$$

where $\mathcal{T}(\mathbf{I}) = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} \mid \nabla \mathbf{I}(i,j) \mid$ computes the average gradient magnitude as texture complexity, and $\tau_1 = 0.3$, $\tau_2 = 0.1$ are empirically determined thresholds.

In the experimental configuration, images are resized to 224 x 224 x 3 and adaptively partitioned into patches based on texture complexity. Patch vectors are linearly projected into a d-dimensional embedding space represented in equation The image is partitioned into N non-overlapping patches of size $P \times P$:

$$N = \frac{H}{P} \times \frac{W}{P}$$

Each patch is flattened and linearly projected to embedding dimension C:

$$\mathbf{X}_{\text{patches}} \in \mathbb{R}^{N \times 3P^2} = \text{Flatten}(\text{Partition}(\mathbf{I}, P))$$
$$\mathbf{Z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{X}_{\text{patches}} \mathbf{W}_E] + \mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1) \times C}$$

where:

- $\mathbf{W}_E \in \mathbb{R}^{3P^2 \times C}$ = learnable patch embedding matrix
- $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1) \times C}$ = learnable positional embeddings
- $\mathbf{x}_{\text{class}} \in \mathbb{R}^{C}$ = learnable classification token
- $[\cdot;\cdot]$ denotes concatenation along the sequence dimension.

3.2.2. Window-based Multi-head Self-attention with Shifted Windows

Within each block, MHA computes attention across multiple subspaces, consequently improving representational capacity relative to single-head attention. For an input token matrix (Z), query, key, and value projections are defined in equation For input features $\mathbf{Z} \in \mathbb{R}^{N\times C}$, the enhanced window-based attention mechanism is defined as:

Query, Key, Value Projections:

$$\mathbf{Q} = \mathbf{Z}\mathbf{W}_Q, \mathbf{K} = \mathbf{Z}\mathbf{W}_K, \mathbf{V} = \mathbf{Z}\mathbf{W}_V$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{C\times d_k}$ are projection matrices, and $d_k = \frac{C}{h}$ with h attention heads scaled dot-product self-attention is processed in equation 5 as

$$\mathbf{Z}_{\text{windowed}} = \text{WindowPartition}(\mathbf{Z}, M) \in \mathbb{R}^{\frac{N}{M^2}\times M^2 \times C}$$
(5)

Window-based attention is applied within local windows, and shifted windows are alternated across successive blocks, which enables cross-window interactions without full global quadratic cost, as described in the Swin formulation in equation 6. Residual learning is retained through the attention sub-layer:

$$\mathbf{Z}_{\text{shifted}} = \text{Roll}(\mathbf{Z}, \left\lfloor \frac{M}{2} \right\rfloor, \left\lfloor \frac{M}{2} \right\rfloor)$$
(6)

3.2.3. Multi-Scale Feature Aggregation

The standard transformer feedforward network is enhanced by multi-scale feature aggregation to strengthen pattern extraction at different symptom scales. The conventional FFN in transformer models is commonly expressed using two linear layers through GELU activation functions:

$$\text{MSA}(\mathbf{Z}) = \text{LayerNorm}\left(\mathbf{Z} + \sum_{s=1}^{3} \alpha_s \cdot \mathcal{F}_s(\mathbf{Z})\right)$$
(7)

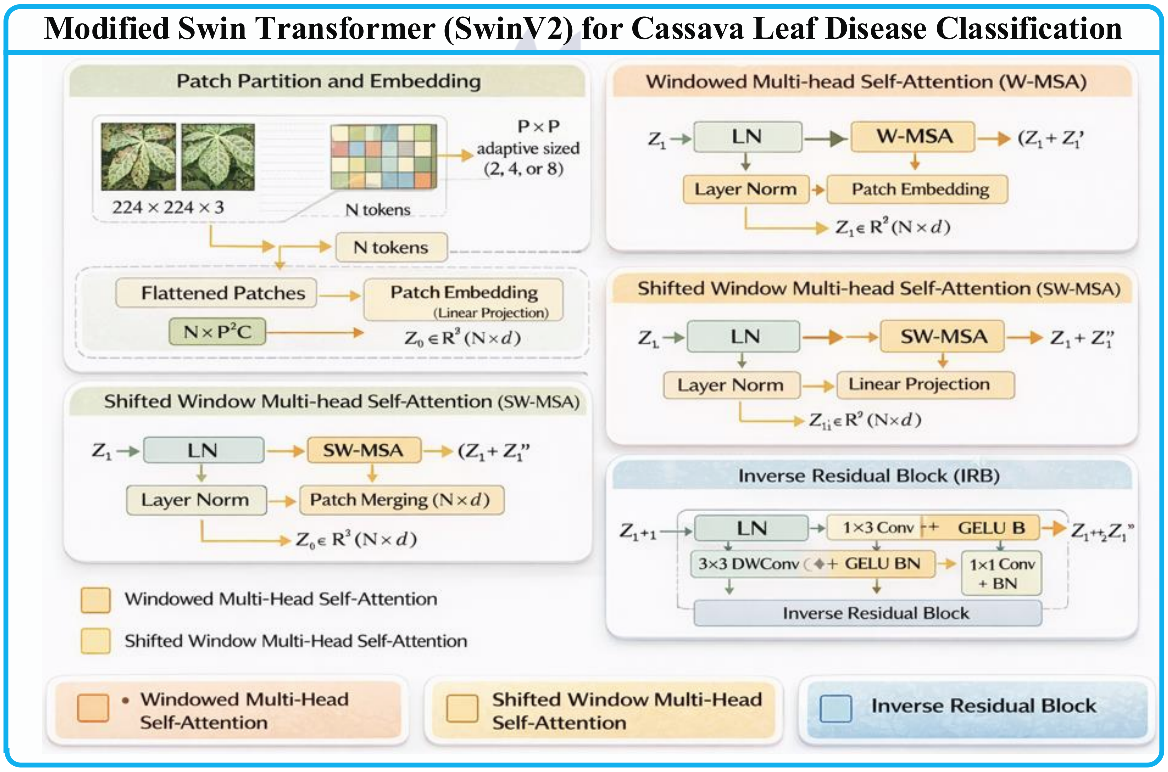


Figure 2. Customized Swin Transformer Framework.

The aggregation adopts a hierarchical structure with feature fusion across scales for comprehensive symptom representation. Expansion and projection operations are implemented through multi-scale transforms around depth-wise operations, and skip connections are retained to support optimisation stability in deep networks in equation 8. A compact formulation is:

$$\mathcal{F}_s(\mathbf{Z}) = \mathrm{Conv}_{1\times1}^{(s,2)}\left(\mathrm{GELU}\left(\mathrm{Conv}_{1\times1}^{(s,1)}\left(\mathrm{DWConv}_{3\times3}^{(d_s)}(\mathbf{Z})\right)\right)\right) \quad (8)$$

with dilation rates $d_1 = 1, d_2 = 2, d_3 = 3$ for multi-scale processing.

Where in equation 8 Φ(.) denotes feature transformation denotes aggregation back to the original dimensionality. The block output is formed with layer normalisation and a residual path This design enables the simultaneous modeling of global dependencies via a shifted-window MHA and multi-scale symptom patterns via the

aggregation pathway, thereby supporting multi-class discrimination under visually similar disease conditions, as depicted in Equation 9.
9.

Experimental Setup:

Dataset Details

This study utilizes a publicly accessible Kaggle dataset of agricultural images, professionally annotated for multi-class classification. The set of images is divided into five groups, which are Cassava Bacterial Blight, Cassava Brown Streak Disease ,Cassava Green Mottle, Cassava Mosaic Disease, and Healthy. The sample images of each class are shown in Figure 3. The proportion of the images represented in each class, described in Table 2, simulates a real-world setup, which also shows class imbalance. This diversity and scale support the robust training and validation of DL models for the differential diagnosis of cassava diseases from visually similar conditions.

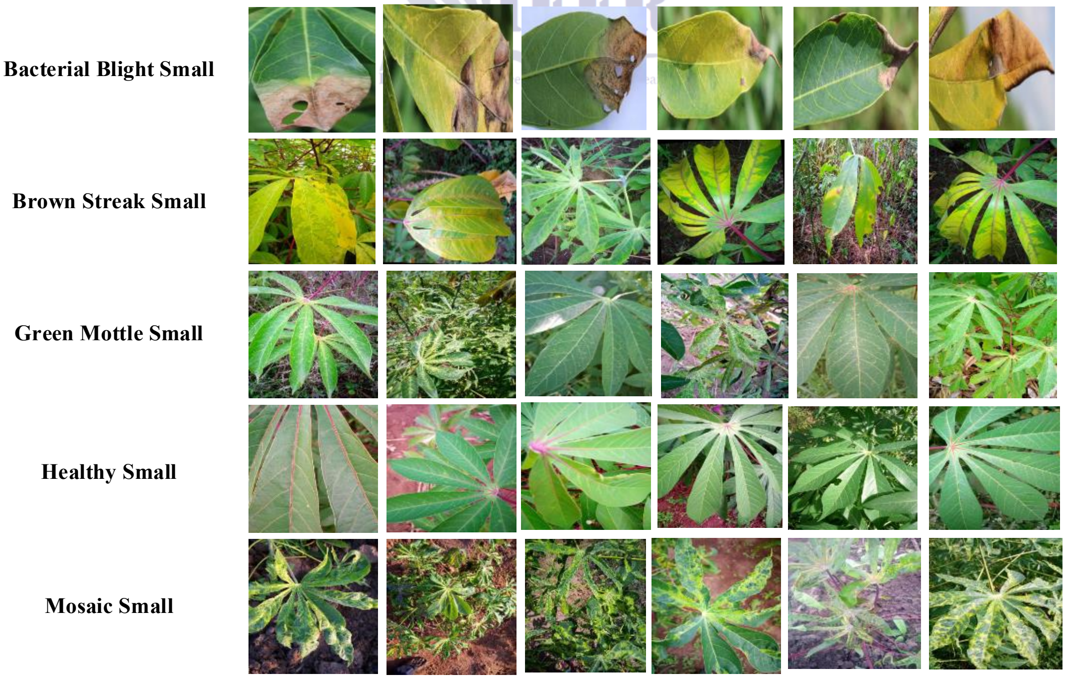


Figure 3. Five diagnostic categories: Bacterial Blight Small, Brown Streak Small, Green Mottle Small, Mosaic Small, and Healthy small cassava leaf Images.

Table 2. Composition of the cassava leaf disease image dataset.

Characteristic	Specification
Total	31179 Samples
Bacterial Bligh Small	7322 Samples
Brown Streak Small	6695 Samples
Green Mottle Small	7018 Samples
Healthy small	6330 Samples
Mosaic Small	3814 Samples
Train (81%)	(25441)
Test (19%)	(5738)
Picture Dimension	224 x 224 x 3
Total	31179 Samples

Experimental Setup

Training configuration for the Swin Transformer framework employed Adam optimizer with a starting learning rate of 10^-3^, a weight decay parameter of 0.04, and a scheduled decay factor of 0.85 applied at 20-epoch intervals. The cross-entropy loss function was selected to manage inter-class imbalance. A batch size of 16 and a final-layer dropout of 0.3 were implemented to reduce overfitting risk. All experiments were coded in Python with TensorFlow and executed on hardware featuring an Intel Core i9-12^th^ Gen CPU, 64 GB of RAM, and a NVIDIA GeForce RTX 4070 Ti GPU.

Evaluation Protocol

A hold-out validation protocol allocated 20% of the total data as a fixed test set. Model efficacy was quantified using conventional diagnostic metrics: Accuracy, Precision, Sensitivity, the F1-score, and the area under the receiver operating characteristic (AUC-ROC) and precision-recall curves (AUC-PR). These metrics are defined by conventional formulations, equations 10-13. Since there was paramount interest in the detection of the cassava disease cases, there was significant focus on maximizing the Sensitivity or Recall values. The standard error of Sensitivity was also determined, hence allowing the calculation of the 95% Confidence Interval using the z-test of 1.96.

$Acc = \frac{TP + TN}{Total} \times 100$ (10)
$Sen = \frac{TP}{TP + FN} \times 100$ (11)
$Pre = \frac{TP}{TP + FP} \times 100$ (12)
$F - score = 2 \times \frac{Pre + Sen}{Pre + Sen}$ (13)

Results and Discussion

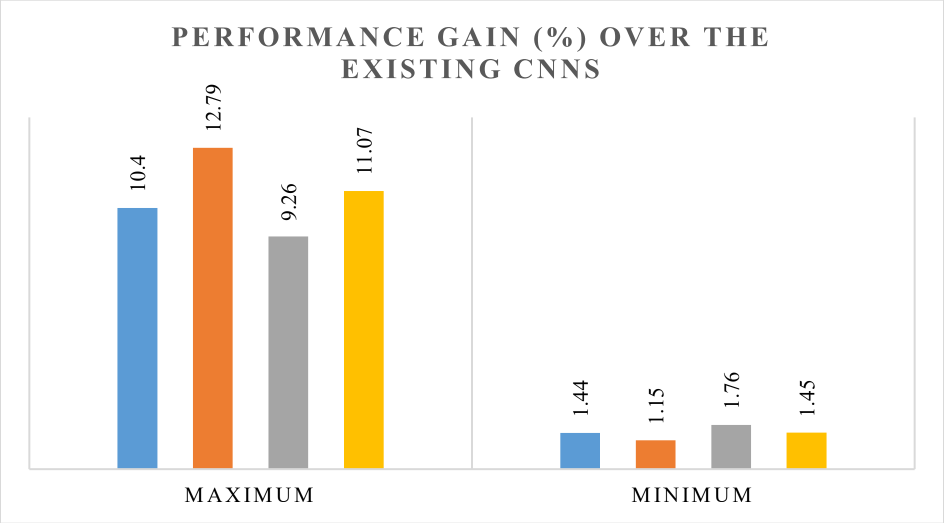
This section presents an analysis of the experimental outcomes and compares the efficiency of the proposed Swin Transformer framework. Results will be compared to the advanced CNN, Vision Transformer, and hybrid CNN/Vision Transformer-based networks on the Kaggle dataset. Table 3 and Figures 4-5 illustrate the main assessment metrics, which involve Accuracy, Precision, Sensitivity, F1_Score, and AUC. Multi-class classification is carried out on the five classifications: CBB, CBSD, CGM, CMD, and Healthy. The proposed framework produces the maximum accuracy of 96.80%, outperforming the CNN accuracy (95.45%) and the baseline Swin Transformer model (96.10%). The ROC-AUC and PR-AUC scores are 0.990 for ranking the classes accurately. The confusion matrix in Figure 4 demonstrates that the error differences are mostly around CMD and CBB, stating that the visual similarity between the two classes is the major point of confusion here. Analysis on a per-class basis reveals that 96.80% of cases are properly classified overall, implying a misclassification error of 3.2%. The computational efficiency can be further noted based on comparisons regarding training complexity. The results revealed that the framework has lower training complexity compared to the employed CNN, ViT, and hybrid models, as presented in Table 3. The proposed approach has faster convergence compared to other alternatives, while hybrid models of CNN and ViT present an oscillatory nature during the optimization process.

Table 1. Model performance metrics for the proposed implementation configuration.

Model	Acc %	Sen	Pre	F1 score
Efficient_Net_B0.	94.21	0.938	0.942	0.940
MobileNetV2.	94.85	0.9455	0.948	0.947
ResNet_50.	95.12	0.948	0.951	0.950
DenseNet_121	95.45	0.951	0.954	0.953
Vision_Transformer	94.78	0.943	0.947	0.945
Swin_Transformer_Baseline	96.10	0.957	0.960	0.959
Hybrid_CNN_ViT	96.15	0.958	0.961	0.960
Proposed_Swin_Framework	**96.80**	**0.964**	**0.968**	**0.966**

Table 2. Performance comparison with prior research

Models	Accuracy	Sensitivity	F1 Score
Existing_CNN's			
MobileNetV2 [21]	96.30	0.960	0.960
Custom CNN [37]	93.00	0.930	0.930
VGG-16 [38]	90.50	0.905	0.905
DenseNet-121 [39]	94.20	0.942	0.942
InceptionResNetV2 [23]	87.00	0.870	0.870
VGG-19 [41]	80.27	0.803	0.803
EfficientNet-B4 [25]	95.10	0.951	0.951
Existing_ViT's			
Vision Transformer [31]	94.50	0.945	0.945
Hybrid_Techniques			
Hybrid CNN-ViT [40]	95.80	0.958	0.958
ResNet-50 + Transformer [42]	95.60	-0.956	-0.956
Swin Transformer [32]	96.10	0.961	0.959



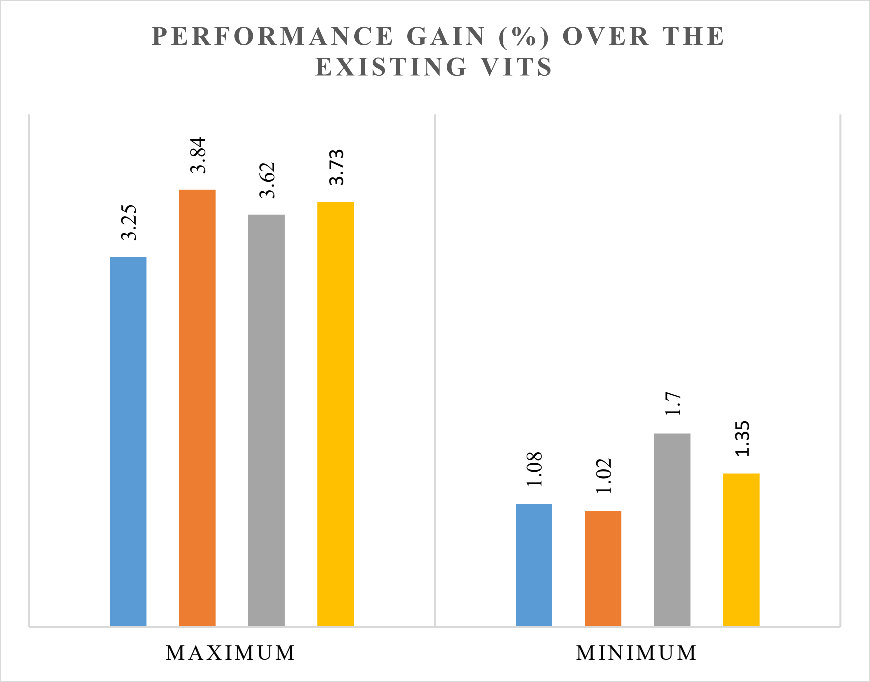


Figure 5. Performance improvements of the proposed Swin Transformer framework compared to established CNN and ViT.

Conclusion

This work introduced a Swin Transformer framework designed to address key challenges in cassava disease image classification. The framework is designed with adaptations of a customized Swin Transformer framework regarding data dimension, embeddings, and a target-oriented result task applied for cassava image interpretation. It integrates adaptive patch and positional embeddings along with a multi-head attention method, enabling global dependency modeling with a newly developed multi-scale feature aggregation. The proposed model applies a hierarchical transformer-based architecture, dividing images into adaptive non-overlapping spatial patches. A sliding window-based self-attention method sustains interwindow connectivity while maintaining computational efficiency, thus reducing the strict locality requirement of traditional Vision Transformers. The multi-scale aggregation applies hierarchical feature fusion aiming to boost the identification of symptoms at different scales and guarantee comprehensive disease representation. The proposed architecture of the Swin Transformer framework helpfully sustains a balanced integration of global contextual cues relatedness with multi-scale symptom patterns, which significantly suppresses intra-class heterogeneities of cassava diseases while improving inter-class discrimination boundaries against similar leaf conditions. Performance evaluation using a public Kaggle challenge indicates the effectiveness of the developed framework, offering the Swin Transformer framework a high classification accuracy of 96.80% with a high F1 score of 96.40%. The framework performs significantly better than conventional CNN models as well as baseline Swin Transformer models. Future studies will aim towards the customization of such architectures to agricultural imaging areas involving challenging environments of field conditions, high heterogeneity of disease manifestations, and limited availability of labeled training paradigms.

Acknowledgment:

We thank the Artificial Intelligence Lab, Department of Computer Systems Engineering, University of Engineering and Applied Sciences, for providing a

healthy research environment and necessary computational resources.

REFERENCES

[1] J. G. A. Barbedo, "Factors influencing the use of deep learning for plant disease recognition," *Biosyst. Eng.*, vol. 172, pp. 84–91, Aug. 2018, doi: 10.1016/j.biosystemseng.2018.05.013.

[2] A. Ramcharan, K. Baranowski, P. McCloskey, B. Ahmed, J. Legg, and D. P. Hughes, "Deep Learning for Image-Based Cassava Disease Detection," *Front. Plant Sci.*, vol. 8, Oct. 2017, doi: 10.3389/fpls.2017.01852.

[3] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using Deep Learning for Image-Based Plant Disease Detection," *Front. Plant Sci.*, vol. 7, Sep. 2016, doi: 10.3389/fpls.2016.01419.

[4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," 2020, pp. 213–229. doi: 10.1007/978-3-030-58452-8_13.

[5] H. Si *et al.*, "A Dual-Branch Model Integrating CNN and Swin Transformer for Efficient Apple Leaf Disease Classification," *Agriculture*, vol. 14, no. 1, p. 142, Jan. 2024, doi: 10.3390/agriculture14010142.

[6] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Jun. 2021.

[7] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2021, pp. 9992–10002. doi: 10.1109/ICCV48922.2021.00986.

[8] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," *Comput. Electron. Agric.*, vol. 145, pp. 311–318, Feb. 2018, doi: 10.1016/j.compag.2018.01.009.

[9] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," *Comput. Electron. Agric.*, vol. 145, pp. 311–318, Feb. 2018, doi: 10.1016/j.compag.2018.01.009.

[10] P. S. Thakur, S. Chaturvedi, P. Khanna, T. Sheorey, and A. Ojha, "Vision transformer meets convolutional neural network for plant disease classification," *Ecol. Inform.*, vol. 77, p. 102245, Nov. 2023, doi: 10.1016/j.ecoinf.2023.102245.

[11] S. Aboelenin, F. A. Elbasheer, M. M. Eltoukhy, W. M. El-Hady, and K. M. Hosny, "A hybrid Framework for plant leaf disease detection and classification using convolutional neural networks and vision transformer," *Complex & Intelligent Systems*, vol. 11, no. 2, p. 142, Feb. 2025, doi: 10.1007/s40747-024-01764-x.

[12] Z. Rauf, A. Sohail, S. H. Khan, A. Khan, J. Gwak, and M. Maqbool, "Attention-guided multi-scale deep object detection framework for lymphocyte analysis in IHC histological images," *Microscopy*, vol. 72, no. 1, pp. 27–42, Feb. 2023, doi: 10.1093/jmicro/dfac051.

[13] A. Ramcharan, K. Baranowski, P. McCloskey, B. Ahmed, J. Legg, and D. P. Hughes, "Deep Learning for Image-Based Cassava Disease Detection," *Front. Plant Sci.*, vol. 8, Oct. 2017, doi: 10.3389/fpls.2017.01852.

[14] W. Wang *et al.*, "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2021, pp. 548–558. doi: 10.1109/ICCV48922.2021.00061.

[15] J. G. A. Barbedo, "Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification," *Comput. Electron. Agric.*, vol. 153, pp. 46–53, Oct. 2018, doi: 10.1016/j.compag.2018.08.013.

[16] Z. Liu *et al.*, "Swin Transformer V2: Scaling Up Capacity and Resolution," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2022, pp. 11999–12009. doi: 10.1109/CVPR52688.2022.01170.

[17] M. Arsenovic, M. Karanovic, S. Sladojevic, A. Anderla, and D. Stefanovic, “Solving Current Limitations of Deep Learning Based Approaches for Plant Disease Detection,” *Symmetry (Basel).*, vol. 11, no. 7, p. 939, Jul. 2019, doi: 10.3390/sym11070939.

[18] S. H. Khan *et al.*, “A new deep boosted CNN and ensemble learning based IoT malware detection,” *Comput. Secur.*, vol. 133, p. 103385, Oct. 2023, doi: 10.1016/j.cose.2023.103385.

[19] J. Chen, J. Chen, D. Zhang, Y. Sun, and Y. A. Nanehkaran, “Using deep transfer learning for image-based plant disease identification,” *Comput. Electron. Agric.*, vol. 173, p. 105393, Jun. 2020, doi: 10.1016/j.compag.2020.105393.

[20] S. Aboelenin, F. A. Elbasheer, M. M. Eltoukhy, W. M. El-Hady, and K. M. Hosny, “A hybrid Framework for plant leaf disease detection and classification using convolutional neural networks and vision transformer,” *Complex & Intelligent Systems*, vol. 11, no. 2, p. 142, Feb. 2025, doi: 10.1007/s40747-024-01764-x.

[21] S. H. Khan, A. Sohail, A. Khan, and Y. S. Lee, “Classification and Region Analysis of COVID-19 Infection using Lung CT Images and Deep Convolutional Neural Networks,” Sep. 2020.

[22] S. H. Khan, M. H. Yousaf, F. Murtaza, and S. Velastin, “PASSENGER DETECTION AND COUNTING FOR PUBLIC TRANSPORT SYSTEM,” *NED University Journal of Research*, vol. XVII, no. 2, pp. 35–46, Mar. 2020, doi: 10.35453/NEDJR-ASCN-2019-0016.

[23] S. H. Khan, A. Khan, Y. S. Lee, M. Hassan, and W. K. Jeong, “Segmentation of shoulder muscle MRI using a new Region and Edge based Deep Auto-Encoder,” *Multimed. Tools Appl.*, vol. 82, no. 10, pp. 14963–14984, Apr. 2023, doi: 10.1007/s11042-022-14061-x.

[24] S. H. Khan *et al.*, “COVID-19 detection and analysis from lung CT images using novel channel boosted CNNs,” *Expert Syst. Appl.*, vol. 229, p. 120477, Nov. 2023, doi: 10.1016/j.eswa.2023.120477.

[25] M. A. Shah, M. M. Alam, and S. H. Khan, “A Tumor Aware DenseNet Swin Hybrid Learning with Boosted and Hierarchical Feature Spaces for Large-Scale Brain MRI Classification,” Jan. 2026.

[26] M. H. A. Khan, M. J. Arshad, M. M. Zahoor, and S. H. Khan, “An Improved Predictive Model for Assessing the Impact of Deforestation and CO_2 Emissions on Flood Hazards in Pakistan,” Apr. 03, 2025. doi: 10.21203/rs.3.rs-6363085/v1.

[27] W. Ullah, Y. N. Khalid, and S. H. Khan, “A Novel Deep Hybrid Framework with Ensemble-Based Feature Optimization for Robust Real-Time Human Activity Recognition,” Jan. 2026.

[28] W. Ahmad, A. R. Shahzad, M. A. Amin, W. H. Bangyal, T. J. Alahmadi, and S. H. Khan, “Machine learning driven dashboard for chronic myeloid leukemia prediction using protein sequences,” *PLoS One*, vol. 20, no. 6, p. e0321761, Jun. 2025, doi: 10.1371/journal.pone.0321761.

[29] S. H. Khan and R. Iqbal, “RS-FME-SwinT: A Novel Feature Map Enhancement Framework Integrating Customized SwinT with Residual and Spatial CNN for Monkeypox Diagnosis,” Oct. 2024.

[30] S. H. Khan and R. Iqbal, “A Comprehensive Survey on Architectural Advances in Deep CNNs: Challenges, Applications, and Emerging Research Directions,” Mar. 2025.

[31] R. U. Din, S. Ahmed, S. H. Khan, A. Albanyan, J. Hoxha, and B. Alkhamees, “A novel decision ensemble framework: Attention-customized BiLSTM and XGBoost for speculative stock price forecasting,” *PLoS One*, vol. 20, no. 4, p. e0320089, Apr. 2025, doi: 10.1371/journal.pone.0320089.

[32] R. U. Din, S. Ahmed, and S. H. Khan, “A Novel Decision Ensemble Framework: Customized Attention-BiLSTM and XGBoost for Speculative Stock Price Forecasting,” Jan. 2024.

[33] S. H. Khan, R. Iqbal, and S. Naz, “A Recent Survey of the Advancements in Deep Learning Techniques for Monkeypox Disease Detection,” Nov. 2023.

[34] S. H. Khan *et al.*, “COVID-19 infection analysis framework using novel boosted CNNs and radiological images,” *Sci. Rep.*, vol. 13, no. 1, p. 21837, Dec. 2023, doi: 10.1038/s41598-023-49218-7.

[35] H. M. Asif, S. H. Khan, T. J. Alahmadi, T. Alsahfi, and A. Mahmoud, “Malaria parasitic detection using a new Deep Boosted and Ensemble Learning framework,” *Complex & Intelligent Systems*, vol. 10, no. 4, pp. 4835–4851, Aug. 2024, doi: 10.1007/s40747-024-01406-2.

[36] A. Ullah, G. Qi, S. Hussain, I. Ullah, and Z. Ali, “The Role of LLMs in Sustainable Smart Cities: Applications, Challenges, and Future Directions,” Feb. 2024.

[37] A. Khan, S. H. Khan, M. Saif, A. Batool, A. Sohail, and M. Waleed Khan, “A Survey of Deep Learning Techniques for the Analysis of COVID-19 and their usability for Detecting Omicron,” *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 36, no. 8, pp. 1779–1821, Nov. 2024, doi: 10.1080/0952813X.2023.2165724.

[38] M. M. Zahoor *et al.*, “Brain Tumor MRI Classification Using a Novel Deep Residual and Regional CNN,” *Biomedicines*, vol. 12, no. 7, p. 1395, Jun. 2024, doi: 10.3390/biomedicines12071395.

[39] S. H. Khan, R. Iqbal, and S. Naz, “A Recent Survey of the Advancements in Deep Learning Techniques for Monkeypox Disease Detection,” Nov. 2023.

[40] S. H. Khan, R. Iqbal, and S. Naz, “A Recent Survey of the Advancements in Deep Learning Techniques for Monkeypox Disease Detection,” Nov. 2023.

[41] S. H. Khan and R. Iqbal, “A Comprehensive Survey on Architectural Advances in Deep CNNs: Challenges, Applications, and Emerging Research Directions,” Mar. 2025.

[42] S. H. Khan and R. Iqbal, “A Comprehensive Survey on Architectural Advances in Deep CNNs: Challenges, Applications, and Emerging Research Directions,” Mar. 2025.

[43] M. M. Zahoor *et al.*, “A New Deep Hybrid Boosted and Ensemble Learning-Based Brain Tumor Analysis Using MRI,” *Sensors*, vol. 22, no. 7, p. 2726, Apr. 2022, doi: 10.3390/s22072726.

[44] S. H. Khan, A. Sohail, M. M. Zafar, and A. Khan, “Coronavirus disease analysis using chest X-ray images and a novel deep convolutional neural network,” *Photodiagnosis Photodyn. Ther.*, vol. 35, p. 102473, Sep. 2021, doi: 10.1016/j.pdpdt.2021.102473.

[45] S. H. Khan *et al.*, “COVID-19 detection in chest X-ray images using deep boosted hybrid learning,” *Comput. Biol. Med.*, vol. 137, p. 104816, Oct. 2021, doi: 10.1016/j.compbiomed.2021.104816.

[46] S. H. Khan, A. Sohail, M. M. Zafar, and A. Khan, “Coronavirus disease analysis using chest X-ray images and a novel deep convolutional neural network,” *Photodiagnosis Photodyn. Ther.*, vol. 35, p. 102473, Sep. 2021, doi: 10.1016/j.pdpdt.2021.102473.