YTUP

Journal of Thermal Engineering
https://jten.yildiz.edu.tr/
https://doi.org/10.47481/jten.0064

RESEARCH ARTICLE

Computationally efficient CNN–vision transformer fusion framework for efficient fault diagnosis in solar photovoltaic system

Priyanka Vyas[1,2,*], Sheetal U. Bhandari[3], Pramod R. Sonawane[4]

[1]Department of Electronics and Telecommunication, Pimpri Chinchwad College of Engineering, Savitribai Phule Pune University, Pune 411044, Maharashtra, India
[2]Department of CSE-AI, Nutan Maharashtra Institute of Engineering and Technology, Pune 410507, Maharashtra, India
[3]Department of Electronics and Telecommunication, Pimpri Chinchwad College of Engineering, Savitribai Phule Pune University, Pune 411044, Maharashtra, India
[4]Department of Electronics and Telecommunication, Pimpri Chinchwad College of Engineering, Savitribai Phule Pune University, Pune 411044, Maharashtra, India

Abstract

Solar photovoltaic (PV) power plants are a key pillar of the global transition to renewable energy, and rapid technological development has led to widespread deployment. As deployment increases, reliable fault detection is essential to ensure long-term performance, safety, and financial sustainability of PV systems. Common defects such as cracking, delamination, discoloration, and shadowing can significantly reduce power output and cause irreversible failure of the module unless these defects are detected in the early stages. Current fault-detection systems are mostly based on classical image-processing methods or single deep-learning models, which cannot adapt to changing environmental and operational conditions and cannot cope with complex visual variations. Moreover, the limited variety and quality of datasets hinder the robustness and generalizability of the model. To examine four key PV defects under different lighting and weather conditions, a high-quality RGB dataset was created. The proposed work introduces a new hybrid CNN-ViT architecture that incorporates depth wise separable convolutions to dramatically decrease the computational complexity while keeping discriminative spatial features. Squeeze-and-excitation blocks are also added to improve defect-aware feature learning and provide adaptive channel-wise recalibration, which can be used to effectively highlight photovoltaic defects such as cracks, delamination, and discoloration. Moreover, a multi-scale attention mechanism with different kernel sizes (1×1, 3×3, and 5×5) is proposed to encode defect patterns at different spatial resolutions. To overcome these shortcomings, this paper proposes a hybrid deep learning architecture that combines a customized convolutional neural network (CNN, ResNet-50V2) and a Vision Transformer (ViT). The CNN part captures fine-grained local spatial features, while the ViT uses global self-attention to be long-range contextual dependencies. The experimental performance is high, with an accuracy of 95.85, a precision of 95.62, a recall of 95.90, and an F1-score of 95.76. The model proposed in this research paper is compared with ultramodern hybrid and transformer-based architectures, including DL-LGM, Cat Boost-GB, DeepF-SVM, DeiT, CCT, and Swin Transformer. Overall, the framework offers a powerful, scalable, intelligent PV monitoring solution that enables prompt fault detection and enhances system reliability and efficiency.

Keywords: CNN-vision transformer, solar photovoltaic system, hybrid model, solar panel fault detection and diagnosis, delamination

Cite this article as: Vyas, P., Bhandari, S. U., & Sonawane, P. R. (2026). Computationally efficient CNN–vision transformer fusion framework for efficient fault diagnosis in solar photovoltaic system. Journal of Thermal Engineering, 12(5), 1760–1777. https://doi.org/10.47481/jten.0064

1. Introduction

The energy industry is rapidly moving towards clean, renewable energy sources. Solar PV technology is rapidly expanding and one of the most promising solutions to sustainable energy [1]. "International Energy Agency (IEA)" explains that green power production has increased faster than ever in the past several decades, and is projected to do so in the future, which is projected to increase threefold by 2030 [2]. The widespread use of solar photovoltaic (PV) technology is crucial for fulfilling global electricity requirements and lowering carbon emissions. PV modules are composed of solar cells made of a

*Corresponding Author
E-mail Adress: priyanka06vyas@gmail.comm

Submitted: 11 November 2025; **Accepted:** 19 June 2026

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkılıç

semiconductor material that converts sunlight into electricity. They are meant to last a long time outdoors, but the performance can be influenced by dust, shade, moisture, UV exposure, and thermal cycling. Cracking, delamination, discoloration, and potential-induced degradation (PID) are structural defects that can reduce power generation and can lead to the eventual failure of modules in whole or in part with time [3]. To ensure that PV systems are safe, dependable, and cost-effective, continuous monitoring via automated problem detection is essential to guarantee the economic viability of these systems. Cracking, shadowing, delamination, and discoloration are the most significant problems because they directly affect the system's output power and service life [4]. Changes in temperature, mechanical stress, or production errors can cause cracks that damage electrical connections in the silicon wafer [5]. Dust, bird droppings, trees, or buildings that are nearby can cast shadows that cause hotspots and damage in certain areas. Delamination occurs when layers separate. Delamination accelerates corrosion by degrading insulation [6]. As the chemicals in the encapsulant layers are broken down by UV radiation, discoloration occurs, usually presenting as yellowing or browning. This leads to the reduction in current and loss of light energy [7]. An air-cooled photovoltaic system incorporating circular pin fins and perfected using Response Surface Methodology was developed to reduce thermal losses and improve solar panel efficiency and operational sustainability [8][9][10]. These issues may show that PV modules are aging and, unless replaced in time, may lead to severe problems.

Common approaches for diagnosing PV include visual inspection, electroluminescence (EL) imaging, and infrared thermography (IRT). Nevertheless, EL inspection should be conducted under controlled laboratory conditions, and IRT is prone to thermal noise and high operational costs. This implies that neither technique is particularly suitable for large-scale implementation of PV. Conversely, RGB imaging of the visible spectrum is a straightforward, low-cost, and easily scalable approach to surface defect detection with ground-based or UAV cameras. With the development of computer vision and artificial intelligence (AI), PV system defects can be found automatically. Earlier "machine learning (ML)" methods that used handcrafted features—namely, "Local Binary Patterns (LBP)" and "Histogram of Oriented Gradients (HOG)"— were unreliable when lighting conditions changed. [11]. Deep learning (DL), particularly CNNs, has markedly improved performance through its autonomous feature-learning capabilities. CNN architectures, including AlexNet [12], VGG [13], ResNet [14], Inception [15], and MobileNet [16]. They have done a great job at finding cracks, hot spots, and damage to surfaces. Even though deep learning (DL) is now a useful tool for monitoring, existing CNN-ViT hybrid models often do not detect RGB PV issues in real working conditions. Nevertheless, CNNs are limited in capturing long-range spatial associations and therefore cannot effectively find small, spatially dispersed defects such as microcracks and color gradients. Also, most datasets are small and narrow, concentrating on one or two types of faults, which are difficult to relate to real-world scenarios. A recent breakthrough in computer vision is the development of Vision Transformers (ViTs). To enhance interpretability and deal with complex fault situations, earlier study uses self-attention mechanisms to encode global contextual relationships in image patches [17]. Pure transformer architectures are potentially effective but computationally expensive and require large datasets to prevent overfitting. Two hybrid CNN-Transformer models, SolarViT [18] and Fusion-Solar-Net [19], are developed to combine the power of CNNs and Transformers to enhance PV fault classification. Current CNN-ViT hybrid models do not effectively detect RGB PV faults under real-world operational conditions. To address this gap, this study aims to introduce a hybrid CNN-ViT framework that improves the computational efficiency of the solar panel fault detection system and achieves better diagnostic accuracy, making it highly reliable for edge-device deployment.

Proposed Solution and Novelty: The paper suggests a lightweight hybrid CNN-ViT architecture, which integrates the local feature-extraction capability of a modified ResNet50V2 and the global dependency modelling capability of a Transformer encoder to address these limitations. The suggested architecture helps find small, overlapping, geographically dispersed faults in solar panels, which traditional CNN-only models cannot find. This study stands out among recent papers of the ViT conference [20] in 2025 because it addresses the gaps in their datasets and architectures. In contrast to SolarViT [18], which is based on mostly drone-based infrared imagery, or Fusion-Solar-Net [19], which is based on environmental contaminants such as dust and snow, this study presents a new lightweight CNN-ViT fusion model that is directly designed to detect multi-class RGB surface-level defects. The dataset was collected from several solar plant sites that experience natural environmental fluctuations. A stronger pre-processing and enhancement pipeline, consisting of CLAHE, bilateral filtering, global normalization, and image augmentation, was used to enhance clarity and reduce noise in the images. The experimental outcomes prove that the proposed CNN–ViT fusion model achieves an accuracy of 95.85%, precision of 95.62%, recall of 95.90%, and an F1-score of 95.76%. This is compared with ultramodern hybrid and transformer-based architectures, including DL-LGM, Cat Boost-GB, DeepF-SVM, DeiT, CCT, and Swin Transformer. These results prove that the hybrid strategy is superior and performs well in detecting solar panel faults in real-world settings. The proposed efficient photovoltaic monitoring system offers a practical benefit to utility-scale solar farms and smart energy management platforms. Early detection of faults in PV systems reduces maintenance costs and improves operational efficiency.

Summary of the novel contributions of this study:

1. A novel lightweight CNN-ViT fusion model detects both significant and subtle PV defects by accurately capturing localized defect textures and long-range spatial correlations.
2. A new real-world RGB dataset with 5600 images of four primary field-level PV faults in different operational situations.
3. Developed an enhanced preprocessing and enhance-

ment pipeline that improves the visibility of surface defects and boosts general performance of model.

4. The proposed model outperforms the ultramodern deep learning and transformer-based architectures in performance comparisons, proving superior accuracy, interpretability, and robustness.

The rest of this paper is structured as follows: Part 2 examines current AI-based PV defect-detection algorithms employing ML, CNNs, ViTs, or combinations thereof. Section 3 outlines the proposed method and architecture. Section 4 presents experimental and ablation studies and a performance comparison with the best available models. Section 5 summarizes the article and discusses future research needs.

2. Literature survey

The literature forms four major types of AI-based fault detection and diagnosis (FDD) methods for PV panels. They are standard machine-learning and deep-learning techniques, mainly based on CNNs, transformer-based architectures, and hybrid CNN-ViT models.

2.1. Traditional machine learning algorithms

These methods rely heavily on built features like LBP, HOG, GLCM, and edge descriptors, as well as classifiers like SVM, Random Forests, and k-NN. For instance, [21] used an SVM classifier with hand-crafted edge and texture features for inspection based on EL, while [22] used lightweight machine learning models built into edge devices. Earlier studies, like [23] and [24], showed that machine learning methods worked well in controlled settings, but the methods weren't strong enough to handle the differences in the real world because they had limited feature generalization.

2.2. CNN-based architectures

The most popular method for automatically finding hierarchical spatial features in EL, IR, and RGB images is the use of CNN-based architectures.[25] showed that CNN models were better than ML methods at finding cracks, inactive areas, and hot spots. The system was more stable with [26] proposing an ensemble CNN model of RGB based surface fault diagnostics. CNN models have also been increasingly used to perform UAV-based PV diagnostics, as proved in [27] and [28]. Localization issues were addressed using pixel-level segmentation techniques, including DeepLabV3+ and U-Net. [29] and [30] were able to correctly separate microcracks and color changes. The datasets were balanced using generative models, namely, "Generative Adversarial Networks (GANs)", which exemplifies community efforts to mitigate the issue of inadequate data. However, CNN-only architectures have difficulty modeling long-range contextual relationships. This poses a significant problem for finding discoloration or delamination locations worldwide.

2.3. Transformer-based models

To overcome the constraints of CNNs, transformer-based designs have garnered interest in their capacity to model global spatial linkages via self-attention mechanisms. Vision Transformer (ViT) models [31] and later variations, including DeiT [32], Swin Transformer [33], and ViTAE [34], have proved exceptional performance in image classification, dense prediction, and remote sensing applications. In PV system monitoring, transformer-driven methodologies, as shown by [35], [36] and [37], showed enhanced resilience to environmental fluctuations. However, most ViT implementations require large, annotated datasets and substantial computing power making them less suitable for PV surface problem detection because datasets are generally small, imbalanced, or collected under uncontrolled environmental conditions.

2.4. Hybrid CNN–ViT architectures

Recently, hybrid CNN–ViT architectures have become an effective approach for integrating the best features of both CNNs and Transformers. Other hybrid frameworks such as TransUNet[38] combine the use of convolutional layers to extract local features and transformer blocks to perform global spatial reasoning. [39] introduced a convolution-guided attention scheme that enhances the quality of the tokens and minimizes the semantic ambiguity. It was proved in [40] that multi-branch hybrid architectures are effective in enhancing the accuracy of classification on complex data. Besides, [41] confirmed the usefulness of CNN spatial filters, combined with transformer attention, in complex real-world detection. Despite these improvements, little research has investigated hybrid CNN-ViT solutions for PV inspection, particularly for RGB-based surface-level defects such as discoloration, delamination, and shadowing, for which real-world datasets stay inadequate.

A literature review reveals several major gaps in the studies. First, most work to date employs EL or IR imaging, which requires expensive equipment and carefully controlled settings to obtain images. This complicates scaling of operations for the routine maintenance of PV fields. Second, the models directly based on CNNs cannot interpret the global structural context and thus are less effective at handling large errors or low-contrast features. Third, running pure transformer topologies is expensive and requires large amounts of annotated data, which are not always easy to obtain in PV fault domains. Lastly, current hybrid CNN-ViT models have not been to detect multi-class RGB surface defects in real-world environments, leaving a significant gap in practical, field-deployable solar asset monitoring systems.

Due to these limitations, this work presents a lightweight, customized CNN–ViT fusion framework specifically engineered for RGB-based detection of significant solar panel faults—such as cracking, delamination, discoloration, and shading—and bolstered by a newly created real-world dataset gathered under various plant conditions. The goal of this hybrid design is to combine the local texture sensitivity of CNNs with the global dependence modelling of transformers. This will make next-generation PV condition-monitoring systems more resilient, scalable, and easier to understand.

Table 1 shows that various CNN-based and transfer-learning methods perform well in controlled settings; however, they primarily rely on EL and IR imagery, limiting their scalability in real-world PV settings. Even though recent fusion and transformer-based architectures enhance global feature modelling, their performance is constrained by the scarcity of RGB training data and high computational costs. Thus, a light hybrid model that effectively addresses a variety of RGB surface defects and is deployable in the field stays desirable. It is an expansion of the performance comparison that goes beyond the simple models mobileNetV3 and Resnet50; it includes architectural complexity and experimental validation with more advanced models such as CCT, DeiT, and Swin Transformer.

A "full literature table evaluation of model-based fault detection approaches underlines the need for more research to improve accuracy and flexibility for real-world photovoltaic (PV)systems [65]. Existing deep learning methods for solar panel fault identification are challenging due to imbalanced datasets and environmental influences".

Table 1. Summary of existing deep learning techniques for solar PV fault detection and diagnosis

Sr.No.	Ref.	Technique	Type of Image	Performance
1	[42]	"SPF-Net: InceptionV3 + U-net	RGB	accuracy: 98.34%, Precision: 94%. Recall: 94%. F1 score: 0.94.
2	[39]	CNN	RGB, thermal	Accuracy, precision, recall, F1 score: 94%
3	[43]	EfficientNetB0	IR	Accuracy: 93.93%, F1-score: 89.82%, Precision: 91.50%, Sensitivity: 88.28%
4	[44]	MobileNet, VGG16, Xception, EfficientNetB7, ResNet50	RGB	Accuracy 97%
5	[45]	Transfer learning	EL	InceptionV3: 93%, EfficientB0: 85%, VGG 19: 82%
6	[46]	SVM classifier (custom CNN)	RGB	Accuracy: 93.03%
7	[23]	Support Vector Machine, Fourier image reconstruction	EL	Accuracy: 93.03%, Precision, Recall Rate
8	[47]	U-DenseNet	IR	Accuracy:99.39%, Precision:98.79 Recall: 100%.
9	[48]	DarkNet-53, VGG-19, Xception, EfficientNet	EL	Accuracy: 94.55%
10	[19]	Lightweight CNN -ViT	RGB	Accuracy 95%
11	[49]	Hybrid YOLO8, PSA-CCT	EL	mAP50= 77.9%
12	[50]	ViT	RGB	Accuracy: 93%
13	[33]	Swin transformer	RGB	Accuracy: 87.3%
14	[51]	Ensemble deep neural network (DNN)	RGB dataset Total (600 original images)	Accuracy of 99.68%
15	[52]	custom-designed CNNs	RBG images	91.1% for binary classification and 88.6% for multi-classification.
16	[53]	Ensemble based model	EL image	98.07%, 96.6%,
17	[30]	Vision Transformer (ViT), MobileNet, VGG16, Xception, EfficientNetB7, and ResNet50	2624 electroluminescence (EL) images	Accuracy for monocrystalline: 85.29% Polycrystalline:94.12%
18	[5]	InceptionV3, EfficientNetB0, and customized CNN architectures based on VGG16 blocks	RGB	Accuracy 97%
20	[54]	SVM classifier (custom CNN)	Grayscale	InceptionV3 93%, EfficientNetB0 85%, customized CNN 82%
21	(Sergiu [23])	Support Vector Machine (SVM), pattern recognition, Fourier image reconstruction to detect faults	RGB	0.9857
22	[56]	Mean, pooling and max pooling with a dual-spin pooling mechanism	Electroluminescence (EL)	0.93025
23	[57]	Filter-induced filter augmentation flow (FAI)	EL	0.9697
24	(Pa et al., 2022)	Transfer Learning (Squeeze Net, Dark Net, and Alex Net) CNN Model	Thermal	Accuracy: 97.42%, F1 score 97%
25	(Hassan & Dhimish, 2023a)	Light CNNs architecture	RGB	91.1% for binary classification and 88.6% for multi-classification".
26	(Rahman et al., 2021)	VGG-16, VGG-19, Inception-v3, Inception-v2, ResNet50	EL	98.07%, 96.6%,

In the second part of the literature review, existing datasets used for PV surface defect classification and localization are examined. As summarized in Table 2, most publicly reported studies have employed electroluminescence (EL) and infrared (IR) imaging modalities because of their superior ability to capture subsurface anomalies, such as microcracks and hotspots, under controlled acquisition conditions. Nonetheless, few datasets have cost-effective RGB images, despite their applicability to large-scale field inspections. Moreover, some of the most critical defects, specifically discoloration, delamination, and irregular shading, are underrepresented in existing datasets, and deep learning models do not generalize easily to various real-world lighting, weather, and environmental conditions. This raises a major research gap: the lack of extensive RGB datasets for the surface defects most relevant to performance, which are needed to ensure credible PV health monitoring in operational solar plants.

Table 2. Existing dataset types for faults addressed in various literature

Sr.No.	Author and Reference	Fault addressed	Image-dataset type
1	[58]	"Cracks and hot spots.	IR and visible
2	[13]	Glass breakage, discoloration, delamination, snail trails, and burn mask	IR
3	[59]	shading, soiling, snowing, temperature rise	Thermal
4	[51]	Glass breakage, Burn marks, Snail trail, Discoloration, Delamination	EL
5	[60]	Micro-cracks, Hot spots, Delamination, Deformation	IR
6	[61]	Discoloration, Snail trail, Burn marks, Delamination, Glass breakage	Aerial RGB
7	[28]	Glass breakage, burn marks, discoloration, snail trail, delamination	RGB images from UAV
8	[62]	Cracks (various types), Corrosion, Delamination, Paste spot, Dirtiness, Interconnection, Soiling, Shunt	EL
9	[63]	Glass breakage, burn marks, discoloration, snail trail, delamination	RGB
10	[64]	Glass breakage, snail trails, burn marks, delamination, discoloration	RGB
11	[46]	Glass breakage, burn marks, discoloration, snail trail, delamination"	RGB

The proposed method is deeply rooted in addressing key real operational problems to close the gap between theoretical computational research and real-world field engineering. Physical inspection and defect localization are typically labor-intensive, risky, and prone to human error when conducted in large-scale industrial deployments such as commercial solar PV power plants. This system offers an automated, non-destructive pipeline for real-time visible-fault diagnostics using a high-performance, computationally efficient architecture.

3. Methodology

The proposed methodology includes a customized CNN–ViT model for multiclass fault classification and consists of 4 main stages, as illustrated in Figure 1: (i)data acquisition, (ii)image pre-processing and feature extraction, (iii)hybrid model development integrating convolutional and transformer-based architectures, and(iv)classification of four critical photovoltaic module defects—cracking, shading, discoloration, and delamination.

During data acquisition, two solar power plants are captured as RGB images and complemented with carefully selected web materials (Section 3.1). This is followed by the feature engineering phase, in which the raw images are subjected to contrast enhancement, filtering, resizing, normalization, and augmentation (Section 3.2). Several filter methods are tested, and bilateral filtering is identified as the most effective in improving image quality without destroying defect-relevant structures. The processed images are then used to train the proposed customized hybrid CNN-ViT model during the modelling stage (Section 3.3). Lastly, the trained hybrid model performs multiclass fault classification, assigning one of the four defect labels to each input PV image.

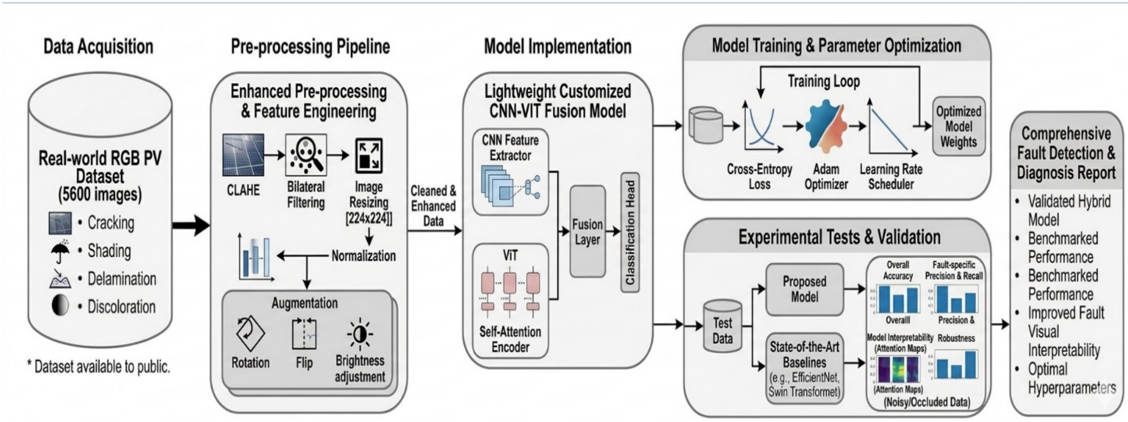


Figure 1. Framework of the proposed CNN–ViT method for panel fault detection

3.1. Data acquisition

RGB-based datasets for studying solar panel faults are not publicly available and are often biased or visually augmented; most datasets concentrate on a narrow range of fault types (e.g., cracks or shading). Minor and subtle defects (e.g., hybrid and early delamination) are often underrepresented, which limits the dataset's representativeness and impairs the model's generalization in real-world settings.

To overcome these limitations, the dataset used in this paper was created from direct field data collected at operational solar power plants and enhanced by the selective inclusion of realistic online RGB images. Field inspections were conducted at two sites: the RMK Solar Power Plant and the Solar Power Plant at the SPPU Research Park Foundation, both of which are in Pune, India, and house outdoor, ground-mounted PV systems that have been operational for more than a decade. Table 3 describes the data collection plan and conditions at each site.

Table 3. Dataset collection details

Location	Camera Model	Lens	Resolution	Lighting Scenarios	Types of Faults Captured
RMK Solar Power Plant, Pune, India	Canon EOS Rebel T7	18–55 mm	24.1 MP	Natural sunlight, overcast, afternoon	Cracks, soiling, shading, bird droppings, delamination, dusting
Solar Panel found at SPPU Research Park Foundation	Canon EOS Rebel T7	18–55 mm	24.1 MP	Natural sunlight, overcast, evening	Discoloration, delamination, cracking, shading

Here, the Canon EOS Rebel T7 DSLR camera is used in the field with an EF-S 18-55 mm kit lens. The 24.1 MP APS-C sensor (6000 × 4000 pixels) in this system has a high pixel density, easing the detection of small errors. This method has enabled acquisition of both the module's general background and small surface defects, including microcracks and localized discoloration, under a wide range of natural lighting conditions. Along with field shots, single-panel RGB images are obtained from sources such as Kaggle, Google Shots, and Adobe Stock. These pictures are chosen because they illustrate the types of faults discussed. Table 4 provides a brief overview of the pooled dataset.

Table 4. Number of images per fault-dataset

Fault Type	Images from Online Sources	Images from Field Visit	Total images
Cracking	859	273	1132
Shading	262	484	746
Delamination	374	699	1073
Discoloration	516	856	1372
Total	**2011**	**2312**	**4323**

Representative examples from each fault class are shown in Figure 2, illustrating the diversity of fault appearance across modules and acquisition conditions. Altogether, the created dataset offered a varied, high-resolution, and field-relevant set of RGB images of solar panel faults, which can be used for both supervised learning and transfer learning. The combination of field-based DSLR imaging and carefully curated web data provides a more balanced, representative dataset, enabling the training and evaluation of deep learning models to detect solar panel faults.

3.2. Image pre-processing and feature engineering

Several pre-processing methods are applied in this work to enhance image quality and extract discriminative fault features for use in later deep learning models. They are "Contrast Limited Adaptive Histogram Equalization (CLAHE)", bilateral filtering, image resizing, normalization, and data augmentation. Raw RGB images often suffer from non-uniform illumination, sensor noise, and background clutter. The pre-processing sequence adopted addresses these issues as follows:

1. CLAHE enhances contrast and highlights subtle fault structures.
2. Bilateral filtering suppresses noise while preserving edges, which is crucial for crack and delamination detection.
3. Resizing standardizes input images to 224 × 224 pixels to ensure compatibility with modern CNN and ViT architectures.
4. Normalization scales pixel values to a standardized range, enabling more stable and faster model convergence.
5. Augmentation (rotation, flipping, and brightness adjustment) increases dataset variability and improves model robustness to real-world variations.

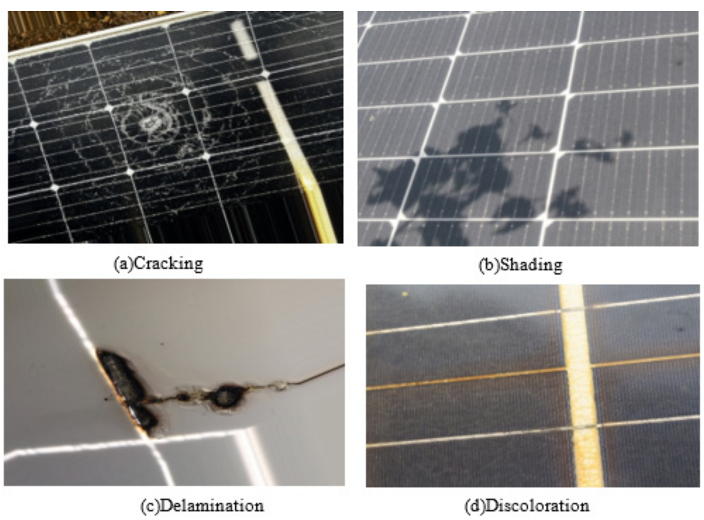


Figure 2. Representative examples of four fault categories: **(a)** cracking, **(b)** shading, **(c)** delamination, and **(d)** discoloration

3.2.1. CLAHE for contrast enhancement

CLAHE is an advanced version of "Adaptive Histogram Equalization (AHE)" that improves image contrast while preventing excessive noise amplification. Moreover, this is particularly effective at enhancing detail in low-contrast images without compromising the natural appearance. These features are obtained by applying histogram equalization to small regions of the image (tiles) rather than to the entire image. Let $P_{new}(x,y)$ refers to enhanced pixel value at the position (x,y), $P_{old}(x,y)$ refers to the original pixel intensity, P_{min} and P_{max} indicates minimum as well as maximum intensity values in local contrast-limited window, and L denotes number of gray levels in image as defined in Equation (1) [65].

$$p_{new}(x,y) = \frac{(p_{old}(x,y) - p_{min}).(L-1)}{p_{max} - p_{min}} \quad (1)$$

This transformation redistributes pixel intensities within each tile, improving the visibility of faint cracks, discoloration gradients, and early-stage delamination.

3.2.2. Bilateral filtering

To decide the best denoising approach, four fundamental filters—mean, median, Gaussian, and bilateral— were applied to images from each fault class. Bilateral filtering was shown to provide the best balance between noise reduction and edge preservation. "Bilateral filtering is a non-linear, edge-preserving smoothing method that combines spatial closeness with intensity similarity. Let $I_{filtered}(x)$ the specify filtered intensity of pixel x; (y) I refers to the intensity of neighboring pixels; W_p defines the normalization factor to ensure the sum of the weights equals $f_r(|I(y) - I(x)|)$ 1; denotes the range kernel, which corresponds to the Gaussian function that preserves edges (as shown in Equation (3)); and $f_s(\| y - x\|)$ refers to the spatial kernel that estimates the Gaussian function controlling the influence of neighboring pixels (as illustrated in Equation (4)).

$$I_{filtered}(x) = \frac{1}{W_p}\sum_{y\in\Omega} I(y) f_r(|I(y) - I(x)|)\, f_s(\| y - x\|) \quad (2)$$

$$f_r(d) = e^{-\frac{d^2}{2\sigma_r^2}} \quad (3)$$

$$f_s(d) = e^{-\frac{d^2}{2\sigma_s^2}} \quad (4)$$

3.2.3. Resizing

To ensure consistency across dataset and compatibility with CNN and ViT architectures, all images are resized to 224 × 224 pixels using bilinear interpolation". The bilinear interpolation method is addressed in Equation (5), "where $I'(x,y)$ specify new pixel intensity of the coordinate (x,y), $I(x_i, y_j)$ defines the neighborhood of pixel intensities, $(1 - |x - x_i|)(1 - |y - y_j|)$ denotes the weights of control interpolation, which depend on the distance from original pixel locations.

$$I'(x,y) = \sum_{i=0}^{1}\sum_{j=0}^{1} I(x_i, y_j)(1 - |x - x_i|)(1 - |y - y_j|) \quad (5)$$

3.2.4. Normalization

The pixel values have been normalized to standard range [0,1], as expressed in Equation (6), where $I_{normalized}(x,y)$ denotes normalized pixel value and $I(x,y)$ is the original intensity, μ indicates the standard deviation.

$$I_{normalized}(x,y) = \frac{I(x,y) - \mu}{\sigma} \quad (6)$$

3.2.5. Data augmentation

Augmentation methods, namely, rotation, flipping, and brightness adjustment, are employed to improve the model's robustness. The rotational transformation is represented in Equation (7), where (x,y) denotes the original coordinates and the new coordinates result from rotation by the angle θ.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad (7)$$

The flipping transformation is estimated by Equation (8), where W denotes the image width for horizontal flipping.

$$I_{flipped}(x, y) = I(W - x, y) \quad (8)$$

The brightness adjustment in data augmentation is defined in Equation (9), where α denotes the brightness scale factor and β denotes the intensity offset.

$$I_{bright}(x, y) = I(x, y) \times \alpha + \beta \quad (9)$$

Overall, the pre-processing pipeline begins with raw images, followed by CLAHE to improve contrast, bilateral filtering to remove noise without losing edges, resizing to normalize image size, normalization to enhance training stability, and augmentation to increase data variability. The steps result in improved input, improved feature discriminability, and" more robust deep learning models in the future.

3.3. Proposed hybrid CNN–ViT model for solar panel fault detection and diagnosis

The suggested hybrid CNN-ViT architecture consists of a modified CNN with ResNet50 at its core and a more sophisticated Vision Transformer (ViT). The feature maps from both the customized CNN and the improved ViT are fused at the final stage and passed to a fully connected classification layer to produce the final prediction.

3.3.1. Customized CNN branch with depth wise separable convolutions, SE-Net, and multi-scale attention

The original ResNet50 backbone is changed into a personalized CNN by incorporating depth wise separable convolutions, "Squeeze-and-Excitation (SE)" blocks, and a multi-scale attention (MSA) mechanism to enhance local feature discrimination and fault localization, as shown in Figure 3.

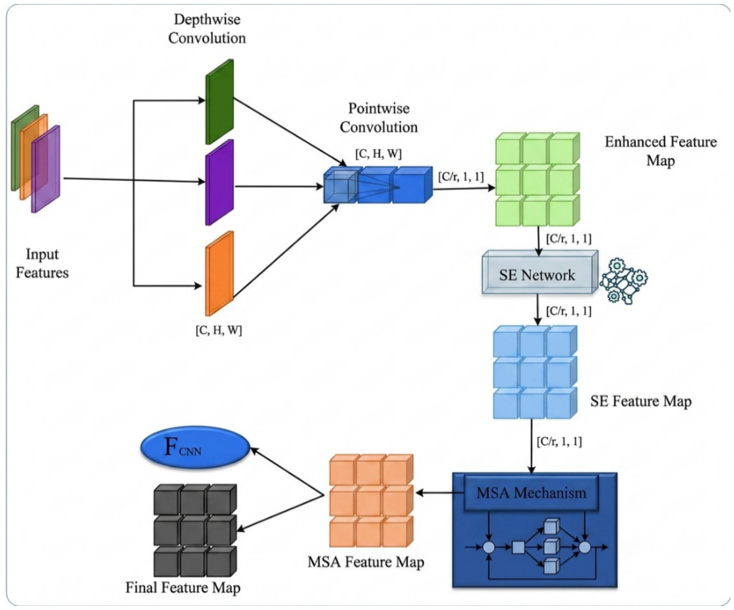


Figure 3. Customized CNN with depth-wise separable convolutions, SE-Net, and MSA

"In the proposed CNN architecture, a depth wise separable convolution is used instead of a conventional convolutional layer to reduce computational cost. Depth wise separable convolution factorizes a standard convolution into two steps: a Depth wise convolution and a pointwise convolution".

Table 5 presents a comparative analysis of generalized and enhanced CNN architectures, highlighting convolution strategy, parameter efficiency, attention mechanisms, and feature representation capability.

Table 5. Comparative analysis between generalized CNN and enhanced CNN architecture

Parameter	Generalized CNN	Enhanced CNN
Convolution Type	Standard convolution (same kernel applied across all channels)	Depthwise convolution (per-channel) + Pointwise convolution (1×1 to mix channels) → Depthwise Separable Conv
Parameter Count	Higher number of parameters	Much fewer parameters — computationally efficient
Feature Attention	No explicit channel attention	Has SE Network → assigns importance to each feature channel
Spatial/Global Dependency	Limited ability to model long-range interactions	MSA (Self-Attention) captures global spatial relationships across the feature map
Output Feature Map	Pure CNN feature map	Enhanced, weighted, globally contextualized feature map
Information Flow	Strong in local patterns, weak in global context	Local features (CNN) + Channel importance (SE) + Global relations (MSA)

Let W_d refer to the depth-wise filter, W_p represent the pointwise filter, C specify the number of channels, k show the kernel size, and X denote the input image. Feature extraction is illustrated in Equations (10) and (11). The final enhanced feature map shown in Equation (12):

$$F_{depth}(x,y,c) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} X(x+i, y+j, c).W_d(i,j,c) \tag{10}$$

$$F_{pointwise}(x,y) = \sum_{c=1}^{C} F_{depth}(x,y,c).W_p(c) \tag{11}$$

$$F_{DSC} = f_{depthwise}(X) + f_{pointwise}(F_{depth}) \tag{12}$$

The second feature is the Squeeze-and-Excitation Network (SE-Net) block. They are used to recalibrate channel-wise responses, thereby emphasizing defect-relevant information, such as cracks, delamination, or discoloration. "SE-Net" applies channel-wise recalibration by using two fully connected layers (W1, W2) with nonlinear activations. The non-linearity is achieved by the ReLU function (δ), while scaling to a normalized range is achieved by the sigmoid function (σ). The recalibrated feature vector is given by Equation (13).

$$F_{SE} = \sigma(W_2 \delta(W_1 F_{CNN})) \tag{13}$$

The Multi-Scale Attention mechanism is introduced; it employs multiple convolutional kernels (1×1, 3×3, and 5×5) to extract features across different fields. MSA enabled the model to detect both fine and coarse patterns. The MSA mechanism highlights attention to less-accessible fields, where A, as shown in Equation (14), denotes the attention weights across scales.

$$F_{MSA} = A_{small} F_{SE} + A_{medium} F_{SE} + A_{large} F_{SE} \tag{14}$$

Let α and β denote the learnable parameters for balancing feature contributions. The final feature map after MSA fusion is given by Equation (15). The output features from the SE and MSA modules are fused via weighted addition, followed by global average pooling and a SoftMax classifier to produce CNN-based predictions.

$$F_{CNN} = \alpha F_{CNN} + \beta F_{MSA} \tag{15}$$

The MSA output is now ready for fusion with the feature maps of vision transformer architecture. Table 6 gives the layer-wise components of CNN.

Table 6. shows the layer-wise components of the CNN

Stage / Module	Type / Operation	Filter / Patch Size	Output Dimension
Input Layer	Input Image	—	75 × 75 × 3
Block-1: Depth wise Separable Convolution	Depthwise Conv	3 × 3	75 × 75 × 3
	Pointwise Conv	1 × 1 → 32	75 × 75 × 32
	MaxPooling	2 × 2 / s2	37 × 37 × 32
Block-2: Depth wise Separable Convolution	Depthwise Conv	3 × 3	37 × 37 × 32
	Pointwise Conv	1 × 1 → 64	37 × 37 × 64
	MaxPooling	2 × 2 / s2	18 × 18 × 64
Block-3: Depth wise Separable Convolution	Depthwise Conv	3 × 3	18 × 18 × 64
	Pointwise Conv	1 × 1 → 128	18 × 18 × 128
SE Attention Module	Squeeze Global Avg Pool	—	1 × 1 × 128
	Excitation FC Layers	128 → 8 → 128	Recalibrated 18 × 18 × 128
Multi-Scale Attention (MSA)	Attention Conv	1×1, 3×3, 5×5	18 × 18 × 128 (combined)
	Feature Fusion	Weighted Sum (α=0.5, β=0.5)	18 × 18 × 128
Global Feature Encoding	Global Average Pool	—	1 × 1 × 128
Classifier Head (CNN Path)	Fully Connected Layer	Dense (128 → Num Classes)	Output: Num Classes

3.3.2. Improved vision transformer

The improved vision transformer architecture is shown in Figure 4. In PV panel fault detection, local and global contextual features are vital for correct classification. To address this issue, the original ViT [41] is incorporated to extract global features.

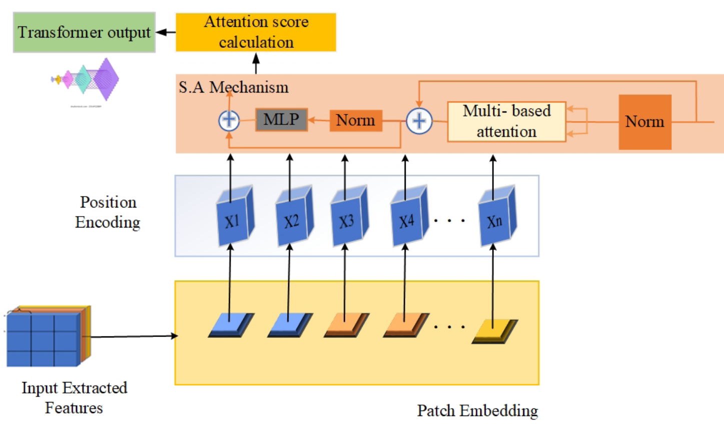


Figure 4. Improved vision transformer architecture

"The input image is 1st pre-processed by dividing it into fixed-size patches, which are then flattened and linearly embedded into feature vectors. Positional encoding is added to keep spatial relationships among patches. The resulting sequence of embeddings is fed into multiple stacked Transformer encoder layers. Each encoder uses a multi-head self-attention mechanism (Q, K, and V projections) along with feedforward networks to capture global contextual information and long-range dependencies across image patches".

The input image is partitioned into P×P patches, which are embedded in a sequential representation. X_p^i Let the patch embeddings be specified; E denotes the embedding E_{pos} matrix, and the positional encoding is evaluated as in Equation (16).

$$Z_0 = [X_p^1 E; X_p^2 E; \ldots; X_p^N E] + E_{pos} \tag{16}$$

The self-attention mechanism calculates attention scores, where Q denotes Z (queries), K denotes $W_Q Z$ (keys), W_K and V W_V denotes $d_k Z$ (values); together, they define the embedding dimension as shown in Equation (17). The transformer output is illustrated in Equation (18):

$$A = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{17}$$

$$F_{ViT} = \mathrm{MLP}(\mathrm{LayerNorm}(A + Z)) \tag{18}$$

An attention-enhanced classification layer reaches the final fault localization W_C output; the classification layer weights are denoted by W, and the output shows the probability distribution of fault classes, $b_c \hat{y}$ as shown in Equation (19).

$$\hat{y} = \mathrm{softmax}(W_c F_{final} + b_c) \tag{19}$$

The modified CNN primarily captures local spatial features, while "Vision Transformer (ViT)" captures global contextual "relationships among image patches. To use the strengths of both architectures, the final feature maps from the CNN and the ViT encoder are combined and fed into a fully connected classification head. This feature-fusion method helps the CNN detect small fault features, such as fracture margins or discoloration boundaries. The upgraded ViT models, on the other hand, capture long-range dependencies, such as large shading patterns across the PV module. The hybrid design yields a stronger, more discriminative feature space for correct solar panel fault classification by combining local and global representations.

3.3.3. Computational and memory efficiency of the proposed customized CNN

The proposed hybrid CNN-ViT framework is expected to effectively learn local and global fault features through the combination of depth wise-separable convolutions, channel attention, multi-scale attention, and multi-scale global modelling using transformers. The ViT captures the long-range contextual relationships among image patches, while the customized CNN extracts fine-grained spatial information (e.g., crack boundaries, fracture margins, and discoloration edges). The combination of CNN and ViT feature maps allows the representation of local defect cues and global structural patterns in a complementary manner, leading to a more discriminative feature space for photovoltaic fault classification. An important mathematical benefit of the proposed CNN is the use of depth wise-separable convolutions instead of standard convolutions. In a conventional 3×3 convolution with input C_{in} channels and output channels, the number of parameters is shown in Equation (16).

$$\mathrm{Params}_{std} = 3 \times 3 \times C_{in} \times C_{out} \tag{16}$$

For example, C_{in} = 3 this C_{out} = 64 results in 1,728 parameters. In contrast, depthwise-separble convolution decomposes this operation into a depthwise convolution followed by a pointwise (1×1) convolution in Equation (17).

$$\mathrm{Params}_{ds} = 3 \times 3 \times C_{in} + 1 \times 1 \times C_{in} \times C_{out} \tag{17}$$

In the same case, only 219 parameters are found, being an 87.3 percent reduction in trainable weights. Consistently as shown in Table 7, the same amount of reduction is clear" throughout convolutional blocks with a range of 85.8 to 88.1 percent parameter savings, showing that there is a significant drop in complexity of model without reduced representational capacity.

Table 7. Per-block parameter counts in proposed model

Block	C_in → C_out	Standard CNN 3×3 params	Depthwise-separable params	Reduction
1	3 → 32	864	123	85.80%
2	32 → 64	18,432	2,336	87.30%
3	64 → 128	73,728	8,768	88.10%

In addition to parameter efficiency, the proposed design requires considerably activation memory. Even though convolutional parameters do not depend on spatial resolution, the memory needed to store intermediate feature maps increases linearly with image size. Processing a high-resolution 4681×3969 image of dimensions generates an activation map of dimensions 4681×3969×32 after the first convolutional block and requires only about 2.21 GB of GPU memory in float32 format. This shows that most hardware platforms cannot support full-resolution image processing in real-time. Thus, as a first step, it is critical to downsize input images to a standard resolution (512x512) to enable model training and inference while keeping important defect data.

Table 8. Standard and proposed CNN MAC calculations

Block (H×W, C_in→ C_out)	Standard MACs	Depthwise-sep MACs	Speedup
4681×3969, 3→32	16,052,160,096	2,285,203,347	7.0×
2340×1984, 32→64	85,571,665,920	10,845,020,160	7.9×
1170×992, 64→128	85,571,665,920	10,176,491,520	8.4×
Total	**187.2B**	**23.3B**	**≈8.0× less compute**

Depthwise-separable convolutions also substantially reduce computational cost. The traditional convolution involves a significant number of multiply-accumulate operations (MACs) since every filter is applied across all input channels. However, separable convolutions decouple the spatial-filtering and channel-mixing processes, and as a result, computational cost decreases significantly. When comparing the three convolutional blocks under analysis, the approximate number of MACs is cut down to 23.3 billion, which results in a total ≈8× speed-up as shown in table 8. This leads to accelerated training, decreased inference latency, and reduced energy usage which are vital for real-time inspection and edge deployment.

In addition to improving efficiency, the proposed architecture enhances feature quality through a built-in attention mechanism. The SE block is an adaptive channel-wise response recalibration block that focuses on fault-relevant feature channels and suppresses redundant information. MSA module combines the responses of multiple receptive fields, allowing detection of defects at various scales, from micro-cracks to large, shaded areas. Global Average Pooling also reduces the representation to a compact, discriminative descriptor.

Finally, the lightweight yet expressive CNN features are fused with ViT embeddings, allowing the transformer to focus on global reasoning with reduced computational burden. This synergy between efficient local feature extraction and transformer-based contextual modelling enables precise and scalable photovoltaic fault detection.

4. Experimental results

4.1. Image pre-processing

The image-processing stage involves image-enhancement operations implemented in Python using the OpenCV library. All images across the different fault categories are subjected to several preprocessing methods to minimize noise and enhance the visibility of fault-related features. Table 9 shows the number of images used in the experiments per class. Four spatial filtering methods are applied to each RGB image as independent variables to effectively suppress noise while preserving features: mean, median, Gaussian, and bilateral filtering.

Table 9. Dataset used for experimentation

Type of faults	Original images	Augmented image	Total images
Cracking	1132	268	1400
Shading	746	654	1400
Delamination	1073	327	1400
Discoloration	1372	28	1400
Total	**4323**	**1277**	**5600**

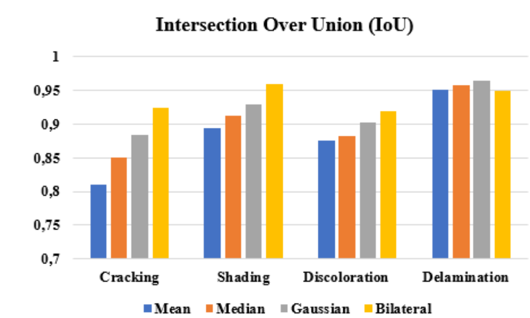
The objective performance of each filtering strategy is measured using "Structural Similarity Index Measure (SSIM)"[66], "Peak Signal-to-Noise Ratio (PSNR)" and "Intersection over Union (IoU)". PSNR measures the quality of image reconstruction compared to the original, while SSIM measures the similarity between two images based on their structure, brightness, and contrast. The IoU measure compares the processed image with its ground truth to assess how well the processed image preserves critical fault areas.

Table 10 presents a numerical comparison of Mean, Median, Gaussian, and Bilateral filtering methods for four types of photovoltaic (PV) faults: cracking, shading, discoloration, and delamination. Gaussian and bilateral filters may be employed to remove cracks and shade. Higher PSNR and SSIM values show that they are more efficient at removing noise and preserving structure. The Bilateral filter has the highest IoU (0.9245) and PSNR (31.38 dB) for cracking defects, showing that it does not blur fine crack edges. Bilateral filtering also achieves the highest IoU (0.9599) and SSIM (0.8983) for shading problems, i.e., it preserves important global shading patterns. Defects such as discoloration and delamination are typically removed using Gaussian filtering; the SSIM values are 0.9028 and 0.9561, respectively. This means that it preserves the smooth variations in the intensity of these problems more effectively. Mean and median filters are used to sharpen images but do not preserve detailed fault structures well. The performance measures show that Bilateral filtering is more effective for faults showing sharp local features, while Gaussian filtering is more effective for faults showing gradual degradation patterns. This makes it easy to extract features that can later be used in group samples.

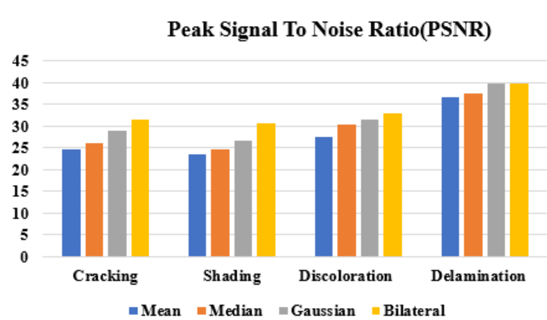
Table 10. Performance of mean, median, gaussian, and bilateral filters on four PV fault types

Fault Type: Cracking				Fault Type: Shading			
Filter	**IoU**	**PSNR(DB)**	**SSIM**	**Filter**	**IoU**	**PSNR (DB)**	**SSIM**
Mean	0.8098	24.6262	0.7713	Mean	0.8938	23.3859	0.728
Median	0.8498	26.1113	0.8331	Median	0.9124	24.5426	0.7681
Gaussian	0.8836	28.831	0.9168	Gaussian	0.9286	26.5652	0.8668
Bilateral	**0.9245**	**31.382**	**0.8878**	**Bilateral**	**0.9599**	**30.6944**	**0.8983**
Fault Type: Discoloration				**Fault Type: Delamination**			
Filter	**IoU**	**PSNR (DB)**	**SSIM**	**Filter**	**IoU**	**PSNR (DB)**	**SSIM**
Mean	0.8748	42.4459	0.9613	Mean	0.9512	36.7996	0.9254
Median	0.8823	30.2912	0.8141	Median	0.9582	37.4008	0.9316
Gaussian	0.903	33.5541	0.9028	Gaussian	0.9641	39.7116	0.9561
Bilateral	**0.9195**	**33.0099**	**0.8374**	**Bilateral**	**0.9498**	**37.4378**	**0.9226**

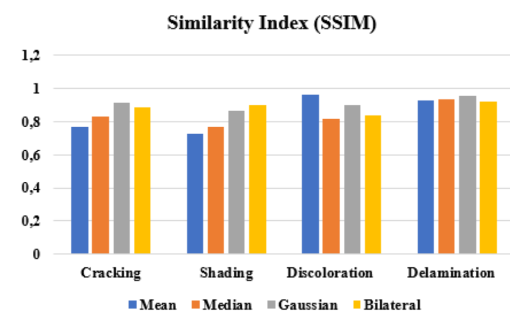
As Figure 5 shows, bilateral and Gaussian filters achieve higher IoU, PSNR, and SSIM values for most PV faults. Bilateral filtering is effective for defects that occur along edges, such as cracking and shading. Gaussian filtering is best used for discoloration and delamination. Generally, the mean and median filters are not as effective at enhancing features.



(a) Intersection Over Union (IoU) of all filter types



(b) Peak Signal to Noise ratio (PSNR)performance of various filters and fault type



(c) Similarity Index of all filters and class type

Figure 5. Performance comparison of filtering methods using IoU, PSNR, and SSIM

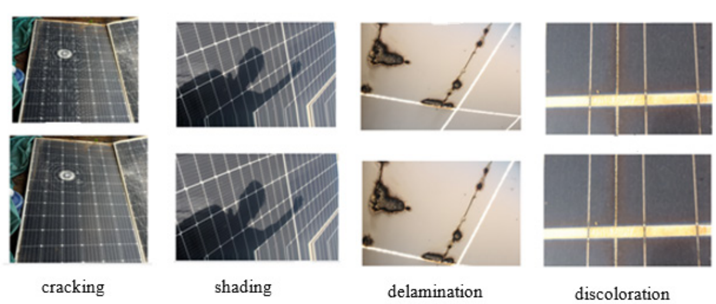


Figure 6. Bilateral filter performance across different fault types

Figure 6 proves that bilateral filtering effectively improves the optical clarity of the photovoltaic fault types—cracks, shading, delamination, and discoloration—while preserving critical edge details. The 1st row shows the original RGB images, while 2nd row proves results of bilateral filtering, which clearly proved fault visibility with negligible structural distortion.

4.2. Results of the proposed hybrid CNN-ViT model

The Hybrid CNN-ViT fault detection and diagnosis model was implemented in Python on a system equipped with an Intel(R) Core (TM) i9-12900 CPU running at 2.40 GHz and with 32.0 GB of RAM. To ensure the reliability of the proposed model, all models were run with the same system configuration and computational parameters. Parameters such as the fixed input image resolution (i.e., 224*224) are given to the model. The previously processed dataset was used for training and testing with fixed ratio as 80: 20 learning rates and the model train using 10 epoch to get quick performance of model which us Adam optimizer. A comparative study was conducted to assess effectiveness of proposed CNN-ViT model against recently developed DL approaches such as DT-LGM [67], CatBoost-GB [68], DeepF-SVM [69], Sparknet[39] Also few recent transformer models (Data-Efficient Image Transformer (DeiT)[17], "Compact Convolutional Transformer (CCT)", and Swin Transformer [33].

The evaluation was based on four performance metrics: accuracy, precision, recall, and F1-score. The dataset, which had 5600 photographs, was divided into 2 parts: 4480 for training and 1120 for testing. This split ensured that the model generalized to unseen data without overfitting. Table 11 shows that the proposed CNN-ViT hybrid model performs better. It achieved an F1-score of 95.76%, which was higher than that of both convolutional DL models and transformer-based models. It achieved 95.85% accuracy, 95.62% precision, and 95.90% recall. The performance improvement results from carefully adjusting hyperparameters, such as the dropout rate and learning rate, which, in turn, improve feature extraction and reduce overfitting. This result shows that the proposed CNN-ViT model is more correct in detecting and classifying multiple fault types in solar PV panels, including cracking, shading, discoloration, and delamination.

The results show that self-attention processes can detect both global and contextual relationships in images of photovoltaic faults. Although they were powerful, their performance was inferior to that of the recommended model.

Table 11. Performance Evaluation of State-of-the-Art Deep Learning Models

Type of Model	Model	Reference	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
DL Hybrid Model	DT-LGM	[67]	89.14	89.2	89.05	89.11
	CatBoost-GB	[68]	88.46	88.36	88.4	88.36
	DeepF-SVM	[69]	85.22	85.37	85.22	85.24
Transformer-Based Model	DeiT	[32]	93.69	93.73	93.69	93.70
	CCT	[49]	93.60	93.85	93.60	93.63
	Swin Transformer	[33]	93.42	93.49	93.42	93.44
Proposed Model	**CNN-ViT**	**Proposed**	**95.85**	**95.62**	**95.90**	**95.76**

Table 12 shows that the proposed CNN-ViT model performs better for both learning and real-time use. DeiT, CCT, and Swin Transformer are all transformer-based systems that perform well, but the proposed method achieves the best balance by combining CNN-driven local feature extraction and Vision Transformer-based global attention. This combination increases the representational power of the model and makes it more computationally intensive. The proposed model has the most competitive overall training time (80.63 minutes) and inference speed (1.70 ms), both the best among independent transformer models. This implies it could be used in real-time fault monitoring, where rapid decisions are needed to keep the functionality and reliability of solar PV systems. The suggested CNN-ViT system provides a more stable and efficient solution for near-accurate, rapid detection of small defects in PV modules, surpassing existing transformer-only methods.

Table 12. Performance efficiency comparison based on training duration and inference latency

Sr.No.	Model Name	Total training time	Average Inference Time
1	DeiT	85.61 minutes	1.84 ms
2	CCT	79.99 minutes	1.96 ms
3	Swin Transformer	67.06 minutes	3.29 ms
4	**Proposed CNN-ViT**	**80.63 Minutes**	**1.70 ms**

The analysis of the confusion matrix shown in Figure 7 proves that the proposed Hybrid CNN-ViT model achieves superior fault-classification performance across all four categories of solar PV defects compared with existing convolutional and transformer-based architectures. The proposed model correctly found 147 samples, with only 4 misclassifications. The same trend was noted in delamination detection, where the proposed model made 160 correct detections and 5 errors. This performed better than the Swin Transformer, which, despite making more correct predictions, made 15 incorrect predictions. The suggested model accurately recognized 146 samples in the slight discoloration group, with only 6 errors. The proposed model had the largest difference in shading faults, with 79 correct and only 7 errors. The findings show that the CNN-based spatial feature extraction and ViT-based global attention greatly reduce the number of false positives, provide stable class-wise discrimination, and make the system less sensitive to small changes in surface quality and lighting. This shows that it applies to reliable, real-time solar PV fault-monitoring applications.

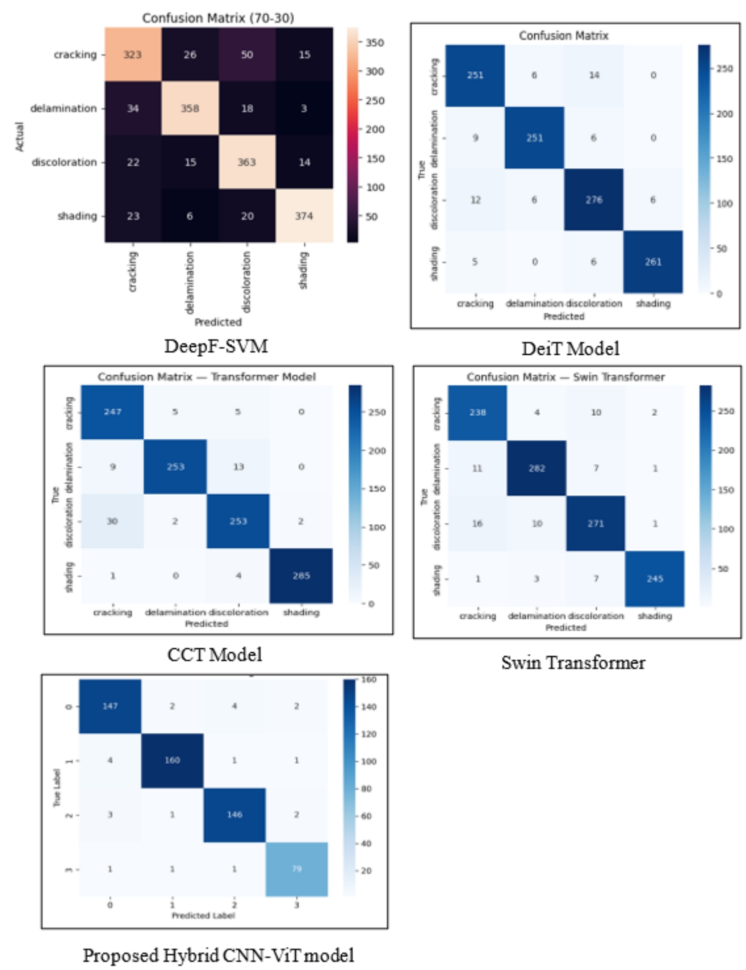


Figure 7. Confusion matrix analysis of state-of-the-art deep learning and transformer models

5. Conclusion

The present study presented highly effective hybrid CNN–ViT-based multiclass PV surface-fault detectors, namely, cracking, delamination, discoloration, and shading, based on visible-spectrum RGB images. An extensive set of 5600 images was created by combining 4323 field images with other augmented samples to ensure good representation of various environmental conditions. A well-developed preprocessing pipeline using CLAHE and bilateral filtering enhanced image quality, as shown by high IoU, SSIM, and PSNR values for various types of faults. This research also introduces substantial advancements beyond the conference version by employing depth-wise separable convolutions, achieving an 87.3% reduction in parameters and 8*0 less computational power compared to a standard layer.

The suggested hybrid architecture, which integrates the local spatial feature extraction of CNNs with the global contextual modelling capabilities of Vision Transformers, worked exceptionally well, with an accuracy of 95.85, a precision of 95.62, a recall of 95.90, and an F1-score of 95.76. It outperforms past standards and ultramodern models, including hybrid DL models, DeiT, CCT, and Swin Transformer models, in terms of robustness and interpretability across a variety of weather and lighting conditions.

The efficiency of operations was also examined, and the model was found to have a competitive training time of 80.63 minutes and an inference time of 1.70 ms per image, making it proper for real-time solar monitoring systems. The confusion matrix was also assessed to support significant reductions in misclassification, particularly for complex defects such as discoloration and shading, and to show greater reliability of fault discrimination under challenging real-world conditions. These results make the CNN-ViT model a practical and scalable intelligent system capable of improving the solar plant status checking system, reducing downtimes, and increasing the overall power-generating efficiency.

The suggested hybrid CNN-ViT model shows excellent performance; however, its limitations, including environmental dependencies, should be acknowledged when considering future applications. Variations in the readings could be caused by ambient temperature, wind speed, and solar radiation in the region of application. This might result in false-positive and false-negative classifications.

Future work will enhance this framework by adding healthy photovoltaic modules to enable comprehensive health diagnostics, expanding the set of defect classifications (e.g., hotspots and potential-induced degradation), conduct extensive inspections by UAVs, and applying active learning and domain adaptation to enable generalization to other installation conditions. This research presents an established, high-performance, and deployment-ready AI model that aims to enhance the sustainable reliability of photovoltaic assets.

Acknowledgment

The authors gratefully acknowledge the Department of Electronics and Telecommunication Engineering at Pimpri Chinchwad College of Engineering, Nigdi, Pune, for providing the facilities, technical support, and supervision during this study. Their ongoing support and the scholarly atmosphere have been vital to the successful execution of this endeavor.

Conflicts of interest

The authors declare that there are no conflicts of interest about the publication of this manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or non-profit sectors.

Author contributions

Priyanka Vyas: Conceptualization; data acquisition; data formulation; model architecture; experimentation; authorship — first draft. Dr. Sheetal U. Bhandari: Oversight; methodological improvement; validation; critique and revision; aid in the study process. Prof. Pramod Sonawane: Methodological design, execution of the CNN–ViT model, data analysis, article revision, and critical evaluation.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability statement

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declaration of generative AI

During the preparation of this manuscript, the authors used an AI-assisted language tool solely to improve the grammar, readability, and language of the manuscript.

Nomenclature

PV	Photovoltaic
RGB	Red, Green and Blue
CNN	Convolutional Neural Network
ViT	Vision transformer
DeiT	Data-efficient Image Transformer
CCT	Compact Convolutional Transformer
DL-LGM	Deep Learning-based Light Gradient Boosting Machine
Cat Boost-GB	Category Boosting Gradient Boosting
DeepF-SVM	Deep Features Support Vector Machine
IEA	International Energy Agency
UV	Ultraviolet
PID	potential-induced degradation
EL	Electroluminescence
IRT	infrared thermography
AI	artificial intelligence
ML	machine learning
LBP	Local Binary Patterns
CLAHE	Contrast Limited Adaptive Histogram Equalization
SVM	support Vector Machine
LBP	Local Binary Patterns
HOG	Histogram of Oriented Gradients
RMK	RMK Solar power Plant
GLCM	Gray-Level Co-occurrence Matrix
SPPU	Savitribai Phule Pune University
DSLR	Digital Single-Lens Reflex
SE Blocks	Squeeze-and-Excitation blocks
MSA	Multi-Scale Attention

References

[1] Patrick Jowett, “IEA forecasts over 4,000 GW of new solar by 2030,” pv magzine.

[2] International Energy Agency, “Renewables 2024: Analysis and Forecast to 2030,” Oct. 2024, International Energy Agency.

[3] J. Van Gompel, D. Spina, and C. Develder, "Cost-effective fault diagnosis of nearby photovoltaic systems using graph neural networks," Energy, vol. 266, p. 126444, 2023, doi: https://doi.org/10.1016/j.energy.2022.126444.

[4] P. R. Sonawane, S. Bhandari, R. B. Patil, and S. Al-Dahidi, "Reliability and Criticality Analysis of a Large-Scale Solar Photovoltaic System Using Fault Tree Analysis Approach," Sustain., vol. 15, no. 5, 2023, doi: https://doi.org/10.3390/su15054609.

[5] B. V. Faassen, J. Serrano, and P. D. Rosero-Montalvo, "Micro-Fracture Detection in Photovoltaic Cells with Hardware-Constrained Devices and Computer Vision," vol. 14, no. 8, pp. 1–9, 2024, [Online]. Available: http://arxiv.org/abs/2403.05694

[6] A. Sinha, O. S. Sastry, and R. Gupta, "Detection and characterisation of delamination in PV modules by active infrared thermography," Nondestruct. Test. Eval., vol. 31, no. 1, pp. 1–16, 2016, doi: https://doi.org/10.1080/10589759.2015.1034717.

[7] A. Sohail, N. Ul Islam, A. Ul Haq, S. Ul Islam, I. Shafi, and J. Park, "Fault detection and computation of power in PV cells under faulty conditions using deep-learning," Energy Reports, vol. 9, pp. 4325–4336, 2023, doi: https://doi.org/10.1016/j.egyr.2023.03.094.

[8] S. Chavali, R. Nanwatkar, and S. Mutalikdesai, "Thermodynamic performance analysis of an air conditioning system integrated with a solid desiccant dehumidifier wheel using solar reactivation," vol. 12, no. 2, pp. 665–683, 2026, doi: https://doi.org/10.14744/thermal.0001109.

[9] A. Hannouch, J. Faraj, R. Taher, M. Mortazavi, and M. Khaled, "A simplified approach to modeling temperature dynamics in photovoltaic systems – Validation, case studies, and parametric analysis," Int. J. Thermofluids, vol. 23, no. July, p. 100767, 2024, doi: https://doi.org/10.1016/j.ijft.2024.100767.

[10] J. Wang et al., "Deep-Learning-Based Automatic Detection of Photovoltaic Cell Defects in Electroluminescence Images," Sensors, vol. 23, no. 1, pp. 1–21, 2023, doi: 10.3390/s23010297.

[11] J. A. Dhanraj et al., "An effective evaluation on fault detection in solar panels," Energies, vol. 14, no. 22, pp. 1–14, 2021, doi: https://doi.org/10.3390/en14227770.

[12] Z. Guo et al., "An Improved Neural Network Model Based on Inception-v3 for Oracle Bone Inscription Character Recognition," Sci. Program., vol. 2022, 2022, doi: https://doi.org/10.1155/2022/7490363.

[13] N. V. Sridharan and V. Sugumaran, "Convolutional Neural Network based Automatic Detection of Visible Faults in a Photovoltaic Module," Energy Sources, Part A Recover. Util. Environ. Eff., vol. 00, no. 00, pp. 1–16, 2021, doi: https://doi.org/10.1080/15567036.2021.1905753.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," CoRR, vol. abs/1512.0, 2015.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 2818–2826, 2016, doi: https://doi.org/10.1109/CVPR.2016.308.

[16] M. Sandler, A. Howard, M. Zhu, and A. Zhmoginov, "Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.pdf," arXiv, pp. 4510–4520, 2018.

[17] S. Kalokhe, F. Khan, A. Shaikh, E. Ansari, S. Shaikh, and N. Jahan, "Computer Vision Advancement with Vision Transformers: A Comprehensive Review," in 2024 5th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks(ICICV), Los Alamitos, CA, USA: IEEE Computer Society, Mar. 2024, pp. 447–453. doi: https://doi.org/10.1109/ICICV62344.2024.00076.

[18] P. Makwane, Y. Kumar, A. Srivastava, M. Gokhale, S. Singh, and V. Sisodiya, "Solar ViT: Vision Transformer for Fault Detection in Solar PV Systems," Int. J. Basic Appl. Sci., vol. 14, no. 2, pp. 535–541, 2025, doi: https://doi.org/10.14419/1h6tfk39.

[19] R. G, R. Rajalakshmi, D. S. Manish, and R. N. A, "Fusion-Solar-Net for solar panel fault detection," Results Eng., vol. 27, no. July, 2025, doi: https://doi.org/10.1016/j.rineng.2025.106513.

[20] P. Vyas, S. Bhandari, and P. Sonowane, "Vision Transformer-Based Fault Diagnosis for Solar Panels Using Visible Image Dataset," 2025, pp. 1–6. doi: https://doi.org/10.1109/ICEPE65965.2025.11139664.

[21] M. A. El-Rashidy, "An efficient and portable solar cell defect detection system," Neural Comput. Appl., vol. 34, no. 21, pp. 18497–18509, 2022, doi: https://doi.org/10.1007/s00521-022-07464-2.

[22] M. Hojabri, S. Kellerhals, G. Upadhyay, and B. Bowler, "IoT-Based PV Array Fault Detection and Classification Using Embedded Supervised Learning Methods," Energies, vol. 15, no. 6, 2022, doi: https://doi.org/10.3390/en15062097.

[23] M. W. Akram et al., "CNN based automatic detection of photovoltaic cell defects in electroluminescence images," Energy, vol. 189, p. 116319, 2019, doi: https://doi.org/10.1016/j.energy.2019.116319.

[24] A. A. Al-Katheri, E. A. Al-Ammar, M. A. Alotaibi, W. Ko, S. Park, and H. J. Choi, "Application of Artificial Intelligence in PV Fault Detection," Sustain., vol. 14, no. 21, 2022, doi: https://doi.org/10.3390/su142113815.

[25] Y. Wang et al., "Adaptive automatic solar cell defect detection and classification based on absolute electroluminescence imaging," Energy, vol. 229, 2021, doi: https://doi.org/10.1016/j.energy.2021.120606.

[26] N. V. Sridharan and V. Sugumaran, "Deep learning-based ensemble model for classification of photovoltaic module visual faults," Energy Sources, Part A Recover. Util. Environ. Eff., vol. 44, no. 2, pp. 5287–5302, 2022, doi: https://doi.org/10.1080/15567036.2022.2083729.

[27] C. F. Jeffrey Kuo, S. H. Chen, and C. Y. Huang, "Automatic detection, classification and localization of defects in large photovoltaic plants using unmanned aerial vehicles (UAV) based

infrared (IR) and RGB imaging," Energy Convers. Manag., vol. 276, no. November 2022, p. 116495, 2023, doi: https://doi.org/10.1016/j.enconman.2022.116495.

[28] J. Jerome Vasanth, S. Naveen Venkatesh, V. Sugumaran, and V. S. Mahamuni, "Enhancing Photovoltaic Module Fault Diagnosis with Unmanned Aerial Vehicles and Deep Learning-Based Image Analysis," Int. J. Photoenergy, vol. 2023, no. iv, 2023, doi: https://doi.org/10.1155/2023/8665729.

[29] L. Pratt, J. Mattheus, and R. Klein, "A benchmark dataset for defect detection and classification in electroluminescence images of PV modules using semantic segmentation," Syst. Soft Comput., vol. 5, no. June 2022, 2023, doi: https://doi.org/10.1016/j.sasc.2023.200048.

[30] D. Dwivedi, K. V. S. M. Babu, P. K. Yemula, P. Chakraborty, and M. Pal, "Identification of surface defects on solar PV panels and wind turbine blades using attention based deep learning model," Eng. Appl. Artif. Intell., vol. 131, 2024, doi: https://doi.org/10.1016/j.engappai.2023.107836.

[31] A. Vaswani, "Attention Is All You Need," no. Nips, 2017.

[32] H. Touvron, M. Cord, M. Douze, F. Massa, and A. Sablayrolles, "Training data-efficient image transformers & distillation through attention," 2020.

[33] Z. Liu et al., "'Swin transformer: Hierarchical vision transformer using shifted windows.' Proceedings of the IEEE/CVF international conference on computer vision.," pp. 1–11, 2021, [Online]. Available: https://github.

[34] Y. Xu, "ViTAE : Vision Transformer Advanced by Exploring Intrinsic Inductive Bias," no. NeurIPS, pp. 1–23, 2021.

[35] H. Tella, A. Hussein, S. Rehman, B. Liu, A. Balghonaim, and M. Mohandes, "Solar photovoltaic panel cells defects classification using deep learning ensemble methods," Case Stud. Therm. Eng., vol. 66, p. 105749, 2025, doi: https://doi.org/10.1016/j.csite.2025.105749.

[36] D. H. Kang and Y. J. Cha, "Efficient attention-based deep encoder and decoder for automatic crack segmentation," Struct. Heal. Monit., vol. 21, no. 5, pp. 2190–2205, 2022, doi: https://doi.org/10.1177/14759217211053776.

[37] E. A. Ramadan, N. M. Moawad, B. A. Abouzalm, A. A. Sakr, W. F. Abouzaid, and G. M. El-banby, "An innovative transformer neural network for fault detection and classification for photovoltaic modules," vol. 314, no. June, 2024.

[38] J. Chen, Y. Lu, Q. Yu, X. Luo, and C. V Feb, "TransUNet : Transformers Make Strong Encoders for Medical Image Segmentation," pp. 1–13.

[39] Rohith G, D. S. Manish, R. Narasimhan A, A. U. Dhavale, and R. R. John, "SparkNet–A Solar Panel Fault Detection Deep Learning Model," IEEE Access, vol. 13, no. March, pp. 75599–75617, 2025, doi: https://doi.org/10.1109/ACCESS.2025.3564714.

[40] A. F. Amiri, S. Kichou, H. Oudira, A. Chouder, and S. Silvestre, "Fault Detection and Diagnosis of a Photovoltaic System Based on Deep Learning Using the Combination of a Convolutional Neural Network (CNN) and Bidirectional Gated Recurrent Unit (Bi-GRU)," Sustainability, vol. 16, no. 3, p. 1012, 2024, doi: https://doi.org/10.3390/su16031012.

[41] Y. Haruna, S. Qin, A. H. Adama Chukkol, A. A. Yusuf, I. Bello, and A. Lawan, "Exploring the synergies of hybrid convolutional neural network and Vision Transformer architectures for computer vision: A survey," Eng. Appl. Artif. Intell., vol. 144, no. December 2024, 2025, doi: https://doi.org/10.1016/j.engappai.2025.110057.

[42] R. A. M. Rudro et al., "SPF-Net: Solar panel fault detection using U-Net based deep learning image classification," Energy Reports, vol. 12, no. July, pp. 1580–1594, 2024, doi: https://doi.org/10.1016/j.egyr.2024.07.044.

[43] Z. B. Duranay, "Fault Detection in Solar Energy Systems: A Deep Learning Approach," Electron., vol. 12, no. 21, 2023, doi: https://doi.org/10.3390/electronics12214397.

[44] Naureen, Sadanand, P. Lohia, D. Dwivedi, and S. Ameen, "A Comparative Study of Quantum Dot Solar Cell with Two Different ETLs of WS2 and IGZO Using SCAPS-1D Simulator," Solar, vol. 2, no. 3, pp. 341–353, 2022, doi: https://doi.org/10.3390/solar2030020.

[45] B. V. Faassen, J. Serrano, and P. D. Rosero-Montalvo, "Micro-Fracture Detection in Photovoltaic Cells With Hardware-Constrained Devices and Computer Vision," IEEE Trans. Emerg. Top. Comput. Intell., vol. 14, no. 8, pp. 1–9, 2025, doi: https://doi.org/10.1109/TETCI.2025.3579274.

[46] C. V. Prasshanth, S. N. Venkatesh, V. Sugumaran, and M. Aghaei, "Enhancing photovoltaic module fault diagnosis: Leveraging unmanned aerial vehicles and autoencoders in machine learning," Sustain. Energy Technol. Assessments, vol. 64, no. October 2023, p. 103674, 2024, doi: https://doi.org/10.1016/j.seta.2024.103674.

[47] R. F. Pamungkas, I. B. K. Y. Utama, and Y. M. Jang, "A Novel Approach for Efficient Solar Panel Fault Classification Using Coupled UDenseNet," Sensors, vol. 23, no. 10, 2023, doi: https://doi.org/10.3390/s23104918.

[48] H. S. Mohd Mokhtar, A. S. Abdul Nasir, M. F. Naim Tajuddin, M. H. Abdul Nasir, and K. Ananda Rao, "Integrating Deep Transfer Learning and Image Enhancement for Enhancing Defective Photovoltaic Cells Classification in Electroluminescence Images," Iran. J. Electr. Electron. Eng., vol. 21, no. 2, pp. 1–15, 2025, doi: https://doi.org/10.22068/IJEEE.21.2.3571.

[49] D. Lang and Z. Lv, "A PV cell defect detector combined with transformer and attention mechanism," Sci. Rep., vol. 14, no. 1, pp. 1–14, 2024, doi: https://doi.org/10.1038/s41598-024-72019-5.

[50] S. Sudarshan, Mohana, G. Ambika, A. Kavitha, K. Nataraj, and B. S. Sudhangowda, "Object Detection using Vision Transformer and Deep Learning for Computer Vision Applications," Proc. 7th Int. Conf. Intell. Sustain. Syst. ICISS 2025, pp. 1524–1529, 2025, doi: https://doi.org/10.1109/ICISS63372.2025.11076364.

[51] S. Naveen Venkatesh, B. Rebecca Jeyavadhanam, A. M. Moradi Sizkouhi, S. M. Esmailifar, M. Aghaei, and V. Sugumaran, “Automatic detection of visual faults on photovoltaic modules using deep ensemble learning network,” Energy Reports, vol. 8, pp. 14382–14395, 2022, doi: https://doi.org/10.1016/j.egyr.2022.10.427.

[52] S. Hassan and M. Dhimish, “A Survey of CNN-Based Approaches for Crack Detection in Solar PV Modules: Current Trends and Future Directions,” Solar, vol. 3, no. 4, pp. 663–683, 2023, doi: https://doi.org/10.3390/solar3040036.

[53] M. R. Rahman, S. Tabassum, E. Haque, M. M. Nishat, F. Faisal, and E. Hossain, “CNN-based Deep Learning Approach for Micro-crack Detection of Solar Panels,” 2021 3rd Int. Conf. Sustain. Technol. Ind. 4.0, STI 2021, no. July 2022, 2021, doi: https://doi.org/10.1109/STI53101.2021.9732592.

[54] B. Karakas and S. Kulluk, “A Hybrid CNN-SVM Algorithm for Detecting Manufacturing Defects,” IEEE Access, vol. PP, no. Dl, p. 1, 2025, doi: https://doi.org/10.1109/ACCESS.2025.3630968.

[55] C. R. Sergiu Deitsch a b d, Vincent Christlein d, Stephan Berger c, Claudia Buerhop-Lutz c, Andreas Maier d, Florian Gallwitz a b, “Automatic classification of defective photovoltaic module cells in electroluminescence images,” Sol. Energy, vol. 185, 2019, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0038092X19302014

[56] S. Hassan and M. Dhimish, “Dual spin max pooling convolutional neural network for solar cell crack detection,” Sci. Rep., vol. 13, no. 1, pp. 1–16, 2023, doi: https://doi.org/10.1038/s41598-023-38177-8.

[57] M. Hussain, H. Al-Aqrabi, and R. Hill, “PV-CrackNet Architecture for Filter Induced Augmentation and Micro-Cracks Detection within a Photovoltaic Manufacturing Facility,” Energies, vol. 15, no. 22, 2022, doi: https://doi.org/10.3390/en15228667.

[58] A. Di Tommaso, A. Betti, G. Fontanelli, and B. Michelozzi, “A Multi-Stage model based on YOLOv3 for defect detection in PV panels based on IR and Visible Imaging by Unmanned Aerial Vehicle,” Nov. 2021, [Online]. Available: http://arxiv.org/abs/2111.11709

[59] T. Selvaraj et al., “Environmental Fault Diagnosis of Solar Panels Using Solar Thermal Images in Multiple Convolutional Neural Networks,” Int. Trans. Electr. Energy Syst., vol. 2022, 2022, doi: https://doi.org/10.1155/2022/2872925.

[60] P. Plant, F. Detection, U. Deep, and L. Techniques, “Learning Techniques,” pp. 1–16, 2022.

[61] S. Naveen Venkatesh et al., “Photovoltaic Module Fault Detection Based on Deep Learning Using Cloud Computing,” Sci. Program., vol. 2023, 2023, doi: https://doi.org/10.1155/2023/8805817.

[62] R. Al-Mashhadani et al., “Deep learning methods for solar fault detection and classification: A review,” Inf. Sci. Lett., vol. 10, no. 2, pp. 323–331, 2021, doi: https://doi.org/10.18576/isl/100213.

[63] N. V. S, D. Sripada, S. V, and M. Aghaei, “Detection of visual faults in photovoltaic modules using a stacking ensemble approach,” Heliyon, vol. 10, no. 6, p. e27894, 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e27894.

[64] N. V. Sridharan, S. Vaithiyanathan, and M. Aghaei, “Voting based ensemble for detecting visual faults in photovoltaic modules using AlexNet features,” Energy Reports, vol. 11, no. December 2023, pp. 3889–3901, 2024, doi: https://doi.org/10.1016/j.egyr.2024.03.044.

[65] N. Kaur, “Hybrid image splicing detection: Integrating CLAHE, improved CNN, and SVM for digital image forensics,” Expert Syst. Appl., vol. 273, p. 126756, 2025, doi: https://doi.org/10.1016/j.eswa.2025.126756.

[66] I. Bakurov, M. Buzzelli, R. Schettini, M. Castelli, and L. Vanneschi, “Structural similarity index (SSIM) revisited: A data-driven approach,” Expert Syst. Appl., vol. 189, p. 116087, 2022, doi: https://doi.org/10.1016/j.eswa.2021.116087.

[67] S. L. P, S. S, and M. S. Rayudu, “IoT based solar panel fault and maintenance detection using decision tree with light gradient boosting,” Meas. Sensors, vol. 27, p. 100726, 2023, doi: https://doi.org/10.1016/j.measen.2023.100726.

[68] K. Chahine, “Tree-Based Algorithms and Incremental Feature Optimization for Fault Detection and Diagnosis in Photovoltaic Systems,” Eng, vol. 6, no. 1, 2025, doi: https://doi.org/10.3390/eng6010020.

[69] I. Charabi, M. B. Abidine, B. Fergani, and M. Oussalah, “DeepF-SVM: A new hybrid deep learning model for enhanced sensor-based human activity recognition,” Cluster Comput., vol. 28, no. 14, 2025, doi: https://doi.org/10.1007/s10586-025-05636-y.