AAPN-Tiny: A Compact Edge-Deployable Adaptive Attention Pyramid Architecture for Multi-Class Fault Diagnosis in Solar Photovoltaic Modules

Rayappa David Amar Raj, Rama Muni Reddy Yanamala, Archana Pallakonda, and Kanasottu Anil Naik

***Abstract*—Solar PV arrays are susceptible to various faults, such as hotspots, cracks, and Potential Induced Degradation, which can impair efficiency and longevity. Traditional fault detection methods are time-intensive and limited in accuracy, especially for large-scale installations. This paper proposes an Adaptive Attention Pyramid Network (AAPN) for accurate and efficient fault detection in PV modules. AAPN integrates depthwise separable convolutions, squeeze-and-excitation blocks, and adaptive attention mechanisms to achieve high accuracy in identifying fault types across different classification complexities. Extensive experimentation on a comprehensive dataset of infrared PV images, organized into 12 fault classes, demonstrated AAPN's high classification accuracy of up to 96% in binary and 92% in 12-class classification scenarios. The proposed model is tested using an infrared solar module dataset for 2-class, 8-class, 11-class, and 12-class fault categories. Its effectiveness is compared with numerous existing deep-learning models for various fault classes. An ablation study was conducted to evaluate the impact of different architectural components, such as depthwise separable convolutions and squeeze-and-excitation blocks, on the model's performance, showing an optimal trade-off between accuracy and computational efficiency. The proposed architecture model is very lightweight, utilizing only 0.8 million parameters. Its effective balance between high accuracies and low parameter utilization makes it highly suitable for deployment on drone-based edge devices, facilitating on-site real-time PV fault monitoring, maintenance, and detection. Additionally, the model has been successfully implemented on the Google Coral Edge TPU, achieving 40.2 ms inference time per image, confirming its efficiency and suitability for real-time applications in resource-constrained environments.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/9716

***Index Terms*—Adaptive attention pyramid network, batch normalization, CNN, detection, Fault classification, infrared.**

I. Introduction

SOLAR panels are prone to various defects, such as hot spots, microcracks, delamination, and Potential Induced Degradation (PID), which can reduce performance and lifespan [1], [2]. Traditional inspection methods, like visual checks and I-V curve analysis, are labor-intensive and often miss internal faults. Advanced techniques like electroluminescence and thermal imaging help but require expensive equipment [3]. AI-based strategies, particularly those using Convolutional Neural Networks (CNNs), offer scalable, accurate, and cost-effective solutions for real-time defect detection [4], [5]. CNNs can identify a wide range of faults under diverse conditions, integrate with drones for remote monitoring, and enhance PV system reliability by reducing maintenance costs and boosting efficiency. The overview of classical, advanced PV array fault detection and analysis techniques is shown in Fig. 1.

Table I reviews notable CNN-based fault classification techniques using electroluminescence (EL) and thermographic imaging. Models like CNN, ResNet, EfficientNet, and YOLO demonstrate strong image analysis capabilities for detecting faults such as hot spots and micro-cracks. Techniques like contrast enhancement and GAN-based augmentation enhance accuracy and effectively address imbalanced datasets. For instance, EfficientNet with NCA feature selection achieves high fault categorization accuracy, and the Complementary Attention Network improves focus on faulty regions. Optimized models, such as dual-flow CNN on FPGA, cater to low-power, embedded systems, while edge-cloud models enable real-time assessment for large solar farms. However, challenges persist, including high computational demands (e.g., ensemble ResNets, GAN-augmented UDenseNet, YOLOv5), reliance on high-resolution data (e.g., HRNet with SeFNet), and limited generalizability (e.g., domain adaptation CNNs, VGG-16). These issues, along with model complexity and resource dependencies, hinder the practicality of existing approaches for real-time or resource-constrained applications. To overcome these challenges, a lightweight adaptive attention pyramid network model is proposed, balancing efficiency and performance. The major contributions of proposed work are as follows:

- Proposed the Adaptive Attention Pyramid Network (AAPN), a neural network model for PV module fault detection, designed to improve accuracy and efficiency through depthwise separable convolutions and squeeze-and-excitation (SE) blocks, reducing computational load

The associate editor coordinating the review of this manuscript and approving it for publication was Bruno Henrique Groenner Barbosa (*Corresponding author: Rayappa David Amar Raj*).

Rayappa David Amar Raj is with the Amrita School of AI, Amrita Vishwa Vidyapeetham, Coimbatore, India (e-mail: rd_amarraj@cb.amrita.edu).

R. M. R. Yanamala is with the Indian Institute of Information Technology, Design and Manufacturing, Kancheepuram, Chennai, India (e-mail: yanamalamunireddy@iiitdm.ac.in).

A. Pallakonda, and K. A. Naik are with the National Institute of Technology, Warangal, Telangana, India (e-mails: ap23csr1r06@student.nitw.ac.in, and anilnaik205@nitw.ac.in).

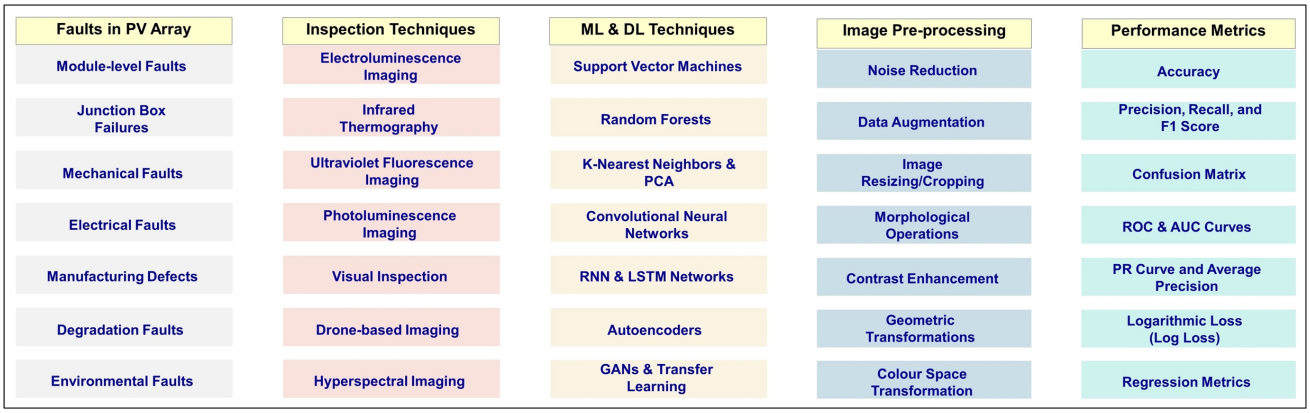

Fig. 1. Overview of classical, advanced PV array fault detection and analysis techniques.

TABLE I
DETAILED LITERATURE SURVEY ON RECENT DEEP LEARNING ARCHITECTURES FOR PV FAULT CLASSIFICATION

Ref	Model Name	Remarks	Advantages	Drawbacks
[6]	DNN with ResNet Ensemble	Residual network ensemble technique for categorizing anomalies for thermographic images	Good accuracy in binary and multi-class classification, effective detection in PV modules	Possible computational intensity resulting from the ensemble method
[4]	CNN-based Fault Classification	CNN + data augmentation on thermographic images for unbalanced datasets	Enhances accuracy for multiple classes using cross-validation & augmentation methods	Less accuracy in multi-class categorization due to high class unbalance
[7]	Edge-CNN on Jetson Nano	Real-time anomaly detection employing IR images on Jetson Nano for edge PV inspection applications	Effective edge deployment characterized by minimal latency and rapid image processing	Less accuracy on smaller edge device models compared to server models
[8]	EfficientNetB0 + NCA Feature Selection	EfficientNet with NCA for classification in infrared images	High accuracy and sensitivity in defect classification, appropriate for large datasets	Relies on exterior feature selection & classifier, rising pipeline complexity
[9]	Coupled UDenseNet+GAN	UDenseNet with GAN augmentation to enhance unbalanced data classification	Attains better accuracy in both binary and multi-class cases	Complex augmentation pipeline & high computations for GAN training
[10]	Improved MobileNet-V3	Improved MobileNet-V3 with transfer learning and data augmentation for UAV monitoring	Effective for real-time inspection with UAVs, accurate across multiple defects	Reliance on aerial imaging quality + UAV precision
[11]	HRNet+SeFNet	HRNet with self fusion network for detection in high-resolution EL images	Enhances multi-resolution feature extraction for precise defection	Complex model needing high-resolution inputs, huge computation cost
[12]	ResNet18+YOLO	Automatic identification with ResNet18 and YOLO on PV modules EL images	High F-1 score and efficient defect localization for large datasets	Huge computations for training YOLO and ResNet
[13]	Dual CNN+Telemetry Data	Region-based CNN for condition monitoring employing thermography & telemetry data	Strong fault detection in PV systems with varied environmental data	Telemetry needs may restrict real-time functionality on conventional devices
[14]	CNN-ResGRU Fusion Model	Fusion of CNN + Res-GRU for classification based on I-V curve & temperature data	Good classification accuracy with robust anti-interference capability	Complexity raised due to fusion architecture & multi-data input
[15]	Complementary Attention Network	Integrates channel + spatial attention for defect identification on PV cell EL images	Improved attention on defect regions, mitigates background noise	High resource requirements due to dual attention networks
[16]	GBH-YOLOv5	YOLOv5 with Ghost convolution + BottleneckCSP for defect detection	Enhanced mAP and inference speed for multi-scale fault defects	Complex model due to added layers and modules for multi-scale target detection
[17]	Deeplabv3 + ResNet-50 Backbone	Semantic segmentation for defect detection in silicon PV EL images	Attains high F-1 score and accuracy with efficient pixel-level classification	Moderate efficacy on less-probable faults; needs high-resolution data
[18]	Hybrid SDAE + CNN for micro-crack	Integrates short-term (SDAE) and long-term CNN features for micro-crack cell detection	Enhances detection accuracy, integrating short-term and long-term features	Less generalization without dataset adaptation
[19]	EfficientNet-B0 + CLAHE and GCAM	Lightweight defect detection with contrast improvement and attention for enhanced visualization	High accuracy with lightweight model; attention improves feature interpretability	Restricted to visible features, reliance on contrast quality
[20]	VarifocalNet + Custom Bottleneck	Modified VarifocalNet for rapid detection with a novel bottleneck module	Achieves speed and accuracy efficiently for real-time defect detection	Complex due to custom bottleneck layers

while enhancing feature extraction.

- Integrated enhanced pyramid blocks, adaptive attention, and SE layers in AAPN to enable multi-scale feature extraction, capturing both fine-grained and high-level spatial features for diverse and complex PV fault detection across varying image scales.
- The proposed model is tested using an infrared solar module dataset, which comprises 12 distinct fault classes (covering complex defect types such as multicell-hotspots, vegetation-induced defects, shading) under performance metrics such as accuracy, loss, precision, recall, F1-score, confusion matrix, ROC-AUC Curves for 2-

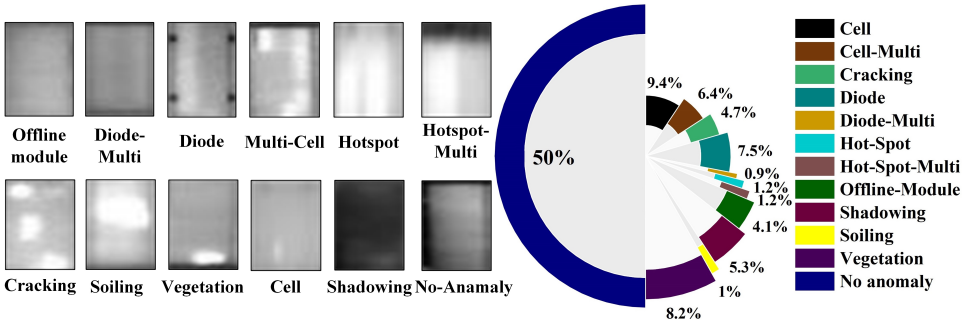

Fig. 2. Overview and the detailed composition of various defect categories in infrared solar module dataset.

class, 8-class, 11-class, and 12-class fault categories.

- The proposed model's effectiveness is compared with the existing deep learning models in terms of accuracy and parameter utilization for various fault classes.
- An ablation study was perfomed to validate the impact of different architectural components, such as depth-wise separable convolutions and squeeze-and-excitation blocks, on the model's performance, showing an optimal trade-off between accuracy and computational efficiency.
- The model has been successfully implemented on the Google Coral Edge TPU, achieving 40.2 ms inference time per image, confirming its suitability for real-time applications in resource-constrained environments.

II. Data Pre-processing and Augmentation

The distinct datasets available for evaluation and data augmentation approaches are described as follows:

A. Standard Datasets for Evaluation

Several benchmark datasets are available to evaluate CNN models for fault detection in solar modules. The ELPV dataset [21] contains 2,624 grayscale images of faulty and non-faulty modules with various defects like cracks and PID. The PVEL-AD dataset [22] includes 36,543 near-IR images with 12 fault types and 40,358 bounding boxes for industrial fault categorization. The Photovoltaic System Thermography Dataset [23] has 277 drone-captured thermographic images for detecting hot spots and snail trails. The PV Multi-Defect dataset [24] provides 1,108 grayscale images with five fault types for model evaluation. Additionally, an infrared solar module dataset [4] used in this paper offers 20,000 IR images (24×40 pixels) classified into 12 categories (Fig. 2), including healthy modules and 11 fault types such as hot spots, cracks, shading, and soiling. These datasets provide a comprehensive resource for training and evaluating models, enabling efficient predictive maintenance.

B. Data Augmentation

One of the main limitations in PV anomaly classification is the lack of large, balanced datasets, requiring augmentation [25] to boost sample size and improve CNN model performance.

Image augmentation with flipping: For each image in the dataset, a random horizontal flip and a random vertical flip are applied, controlled by F_H and F_V. Thus, the transformation for each image I_i in the dataset can be described as Equation 1

$$\hat{I}_i(x,y) = \begin{cases} I(W-x-1,y) & \text{if } F_H = 1 \text{ and } F_V = 0 \\ I(x,H-y-1) & \text{if } F_H = 0 \text{ and } F_V = 1 \\ I(W-x-1,H-y-1) & \text{if } F_H = 1 \text{ and } F_V = 1 \\ I(x,y) & \text{if } F_H = 0 \text{ and } F_V = 0 \end{cases} \tag{1}$$

Image augmentation with translational operation: The translation is implemented via a translation matrix (Equation 2). This matrix is used to shift each pixel in the image by dx and dy, applying the translation operation given by Equation 3.

$$T = \begin{bmatrix} 1 & 0 & dx \\ 0 & 1 & dy \end{bmatrix} \tag{2}$$

$$I_T(x,y) = I\left(\begin{bmatrix} 1 & 0 & dx \\ 0 & 1 & dy \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}\right) \tag{3}$$

Augmentation with clipping: After applying the brightness adjustment, pixel values may exceed the valid range (between 0 and 1 for normalized images). To ensure the pixel values remain valid, we clip the values to this range given by Equation 4.

$$I_B(x,y) = \text{clip}(I(x,y) + \Delta, 0, 1) \tag{4}$$

where $\text{clip}(a,b,c)$ ensures that the value a is constrained between b and c, Δ is random value generated within the range $[-\text{max\_delta}, \text{max\_delta}]$. $\mathcal{U}$ is uniform distribution.

III. Proposed Model Architecture

The proposed model (Fig. 3) is built around three principal components: an Adaptive Attention Pyramid, Enhanced Pyramid Blocks, and Squeeze-and-Excitation (SE) blocks. The figure includes the actual tensor dimensions and intermediate shapes at individual stages of the network. The input and output tensor sizes (H, W, C_in, C_out) for all the blocks are detailed, along with detailed convolution operations (such as number of filters, kernel sizes, and stride/padding) for all the layers offering the data flow, depth, and complexity of the architecture. The architecture was conceived with three main design objectives: (1) to achieve high classification accuracy for multi-class PV fault detection, (2) to minimize computational complexity and memory usage for suitability

on edge devices, and (3) to ensure scalability across different image resolutions and varying fault complexities, enabling deployment on both resource-rich and resource-constrained platforms.

Unlike conventional CNNs, which often trade efficiency for accuracy, the Adaptive Attention Pyramid Network (AAPN) integrates multi-scale feature learning with lightweight operations, allowing for robust performance in scenarios with overlapping fault patterns and low-resolution IR inputs.

Algorithm 1 AAPN with SE and Enhanced Pyramid Blocks

Require: Input tensor $X \in \mathbb{R}^{H\times W\times C}$, filters F, reduction ratio r, classes C

Squeeze-and-Excitation (SE) Block:
Input: Tensor X
$S \leftarrow \text{GlobalAveragePooling2D}(X)$
$Z \leftarrow \sigma(\text{W}_2 \cdot \text{ReLU}(\text{W}_1 S))$, $\text{W}_1 \in \mathbb{R}^{C\times \frac{C}{r}}$, $\text{W}_2 \in \mathbb{R}^{\frac{C}{r}\times C}$
$\tilde{X} \leftarrow X \odot Z$ // Element-wise multiplication
Output: $\tilde{X}$
Enhanced Pyramid Block (EPB):
Input: Tensor X, filters F
$X_1 \leftarrow \text{DepthwiseConv2D}(X) \rightarrow \text{Conv}_{1\times1}(X_1) \rightarrow \text{BatchNorm}(X_1)$
$X_2 \leftarrow \text{Conv}_{3\times3}(X_1) \rightarrow \text{SE\_Block}(X_2) \rightarrow \text{BatchNorm}(X_2)$
$X_{\text{out}} \leftarrow \text{MaxPooling2D}(X_2)$
Output: X_{out}
Adaptive Attention Pyramid Network (AAPN):
Input: $X_{\text{input}} \in \mathbb{R}^{40\times24\times1}$, classes C
$X_0 \leftarrow \text{Conv}_{3\times3}(X_{\text{input}}, 32) \rightarrow \text{BatchNorm}(X_0)$
$X_1 \leftarrow \text{EPB}(X_0, 64)$
$X_2 \leftarrow \text{EPB}(X_1, 128)$
$X_3 \leftarrow \text{EPB}(X_2, 256)$
$G \leftarrow \text{GlobalAveragePooling2D}(X_3)$
$D \leftarrow \text{ReLU}(\text{Dense}(G, 128)) \rightarrow \text{Dropout}(D, 0.4)$
$Y_{\text{pred}} \leftarrow \text{Softmax}(\text{Dense}(D, C))$

A. Adaptive Attention Pyramid Network Architecture

The AAPN model (Algorithm 1), shown in Fig. 3, begins with a 3×3 convolutional layer (32 filters) followed by batch normalization. This initial stage is critical for capturing low-level spatial gradients and edges, which are foundational in differentiating PV module defects. Batch normalization reduces internal covariate shift, accelerating convergence and providing regularization against overfitting in early stages.

Each Adaptive Pyramid Block uses depthwise separable convolutions—a factorization of standard convolution into depthwise and pointwise stages—substantially reducing computational load and parameter count. This is essential for edge deployment, where memory bandwidth and processing capability are limited. For example, with $C_{in} = 64$, $C_{out} = 128$, $k = 3$, $H = W = 24$, a standard convolution requires approximately 42.5M MACs, whereas the equivalent depthwise separable convolution needs only 9.2M MACs—a $\sim$78% reduction without significant loss in representational power.

The computational formulations for conventional and depthwise convolutions are given in Equations (5)–(8):

$$Output[i,j,c_{out}] = \sum_{x=0}^{k-1}\sum_{y=0}^{k-1}\sum_{z=0}^{c_{in}-1} \times Input[i+x, j+y, z] \times Kernel[x,y,z,c_{out}] \tag{5}$$

$$\text{Mul}_{\text{conv}} = C_{out} \times (H-k+1) \times (W-k+1) \times C_{in} \times k^2 \tag{6}$$

Here, i and j are spatial coordinates of the output feature map, c_{out} is output channel, k is the kernel size, $Input$ is the input feature map, and $Kernel$ is the convolution kernel.

$$Output[i,j,c_{out}] = \sum_{x=0}^{k-1}\sum_{y=0}^{k-1} \times Input[i+x, j+y, c_{out}] \times Kernel[x,y,c_{out}] \tag{7}$$

$$\text{Mul}_{\text{depthwise}} = C_{in} \times (H-k+1) \times (W-k+1) \times k^2 \tag{8}$$

Here, i and j represent the spatial coordinates of the output feature map, c_{out} denotes the output channel, k is the kernel size, $Input$ is the input feature map, and $Kernel$ is the depthwise convolution kernel.

B. Enhanced Pyramid Block

The Enhanced Pyramid Block (EPB) extends traditional pyramid structures by integrating depthwise convolutions, batch normalization, and SE recalibration within a single modular unit (Fig. 4). The 3×3 depthwise convolution reduces parameter count while retaining fine spatial resolution, which is critical for small fault patterns in PV cells. Batch normalization [Equations (9), (10)] ensures stable activation distributions, allowing deeper networks to be trained without gradient instability. The embedded SE block adaptively recalibrates channel-wise features, amplifying channels with high defect-relevant responses and attenuating less informative ones. This is particularly beneficial for classes with visually similar features (e.g., hotspot-multi vs. cell-multi), where subtle discriminative cues matter.

$$\hat{x}^{(k)} = \frac{x^{(k)} - \mu^{(k)}}{\sqrt{\sigma^{(k)2} + \epsilon}} \tag{9}$$

$$y^{(k)} = \gamma^{(k)}\hat{x}^{(k)} + \beta^{(k)} \tag{10}$$

In this context, $x^{(k)}$ represents the k-th feature in a batch, while $\mu^{(k)}$ and $\sigma^{(k)2}$ denote the mean and variance of the k-th feature across the batch, respectively. The term ϵ is a small constant included for numerical stability. The normalized value is represented as $\hat{x}^{(k)}$, and $\gamma^{(k)}$ and $\beta^{(k)}$ are learnable parameters (scale and shift) that enable the model to potentially recover the original distribution if needed.

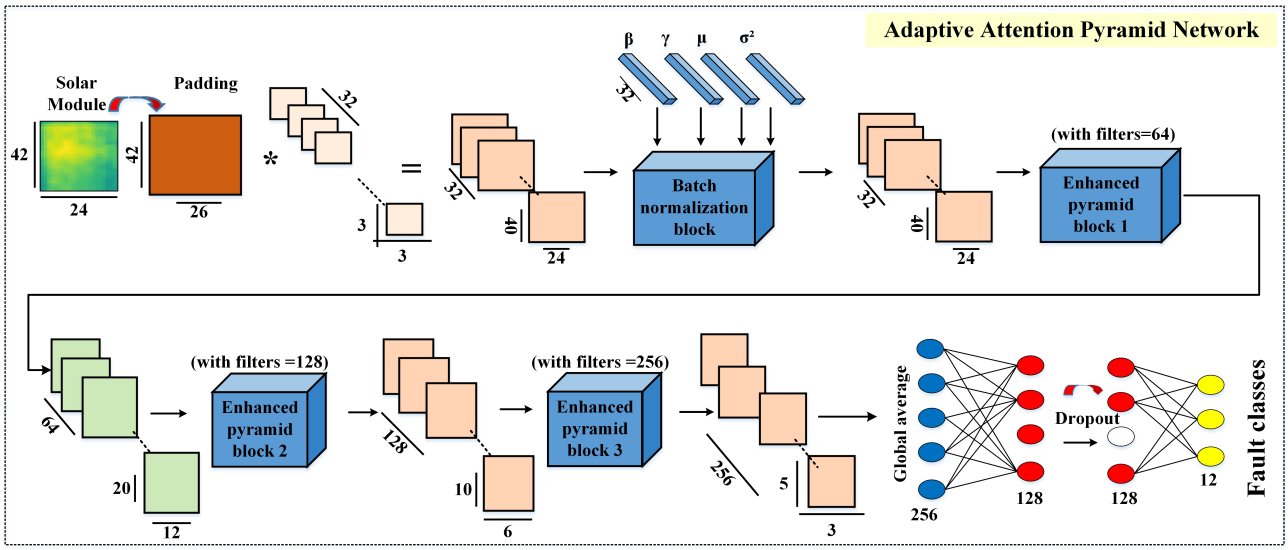

Fig. 3. Internal architecture of the proposed Adaptive Attention Pyramid Network (AAPN) for photovoltaic fault diagnosis.

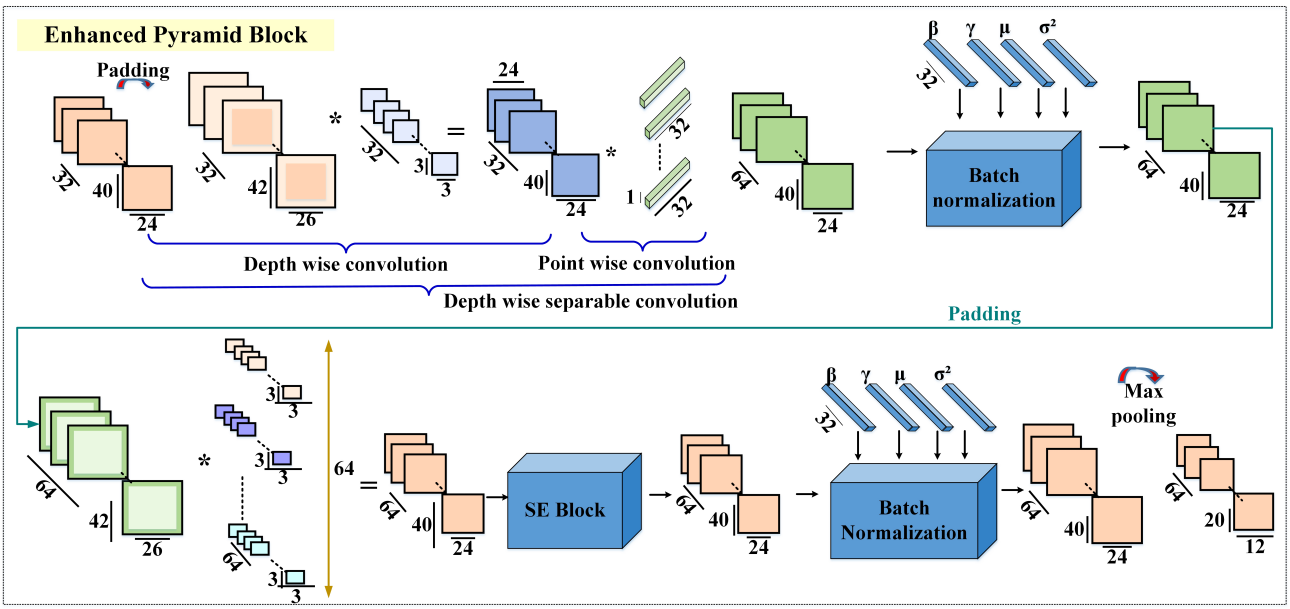

Fig. 4. Internal architecture of the Enhanced Pyramid Block used in the Adaptive Attention Pyramid Network.

C. Squeeze-and-Excitation Block

The SE block applies global average pooling to capture holistic contextual information from each channel, producing a $(1, 1, C)$ descriptor. Two dense layers (Equation 11) then compute channel-wise weights, which rescale the input feature maps. This reweighting mechanism improves the network's attention to defect-related patterns while suppressing irrelevant background noise (e.g., thermal gradients not caused by faults).

$$z = \sigma(W * \text{ReLU}(W_1 * x)) \tag{11}$$

where W_1 is the first dense layer and W is the final dense layer. The block output is then multiplied with the input tensor as shown in Fig. 5 to obtain a recalibrated feature map, prioritizing the most significant features across channels.

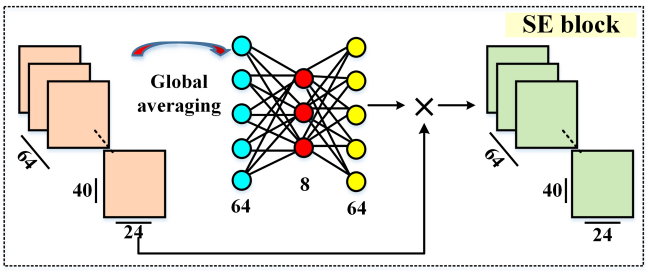

Fig. 5. Internal architecture of squeeze-and-excitation block.

D. Ablation Study on Depthwise Separable Convolutions for Computational Efficiency and Accuracy

To substantiate the selection to employ depthwise separable convolutions (DSC) for lessening computational complexity,

we implemented ablation research (Table II) evaluating various configurations of DSC with traditional convolutions (Conv). The objective was aimed to determine the optimum balance between model accuracy and parameter count. We tested multiple combinations with varying Depthwise Seperable Convolutions (DSC) and Convolutions (Conv), noting both the size of the model (in millions of parameters) and its corresponding accuracy for each combination. The configurations examined comprised 0DSC + 1Conv, providing 91% accuracy with 0.445M parameters; the proposed 1DSC + 1Conv, demonstrating 92% accuracy with 0.878M parameters; 2DSC + 1Conv, yielding 90% accuracy with 0.969M parameters; 3DSC + 1Conv, delivering 89% accuracy with 1.061M parameters; 4DSC + 1Conv, yielding 86% accuracy with 1.151M parameters; 0DSC + 2Conv, yielding 92% accuracy with 1.221M parameters; and 1DSC + 2Conv, providing 91% accuracy with 1.653M parameters.

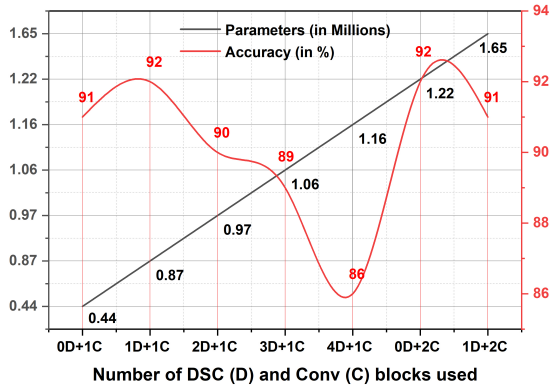

Fig. 6. Performance of the proposed model with varying number of depthwise separable convolutions and conventional convolutions blocks.

The results obtained indicated that the proposed 1DSC + 1Conv configuration exhibited the optimal balance, with the highest accuracy (92%) while maintaining the lowest parameter count (0.878M). Increasing the quantity of DSC layers (e.g., 2DSC + 1Conv, 3DSC + 1Conv) caused increasing parameter counts but lowered accuracy, implying diminishing accuracy from extra DSC layers (Fig. 6). The 0DSC + 2Conv configuration, utilising solely conventional convolutions, gained comparable accuracy (92%) but with a higher parameter count (1.221M), indicating the effectiveness of DSC in decreasing parameters while ensuring accuracy. The findings substantially endorse the application of DSC to decrease computing complexity while ensuring high accuracy, providing the recommended approach optimal to be implemented in environments with limited resources, such as UAV-based monitoring systems. The results from this ablation examination offer a good basis for the architectural choices in the proposed methodology.

E. Trade-off Analysis of Model Configurations

In the proposed model configurations and final selection (Table III), we propose a clear trade-off analysis contrasting accuracy, inference time, and parameter count for the considered configurations. The configurations were tested based on their performance and suitability for deployment on resource-constrained devices, with the results shown in Table 6. Configuration 4 found as the optimal configuration, exhibiting the trade-off between accuracy, inference time, and parameter efficiency. The analysis shows how Configuration 4 outperforms others, specifically in terms of attaining enhanced accuracy of 92% while maintaining a relatively low inference time of 40.2 ms and an optimal parameter count of 0.88M, making it suitable for edge-based monitoring applications. The trade-off analysis exhibits that the Configuration 4 provides the highest accuracy while ensuring efficient computational performance, surpassing configurations with lower accuracy or higher parameter counts. Despite configurations 1 and 2 offering faster inference times (Table III), they compromise accuracy, with Configuration 4 yielding the best optimum. The hyperparameters, including the number of enhanced blocks and dense layers in the SE block, were chosen based on testing to guarantee that Configuration 4 would provide optimal performance. By choosing three enhanced blocks and two dense layers in the SE block, a good balance between model complexity and computational efficiency is attained. Hence, the Configuration 4 was selected for its capability to yield high accuracy and efficiency without compromising real-time deployment in resource-constrained environments. This inclusive analysis establishes the rationale for choosing the Configuration 4 and elucidates the decision-making procedure for subsequent deployments.

TABLE II
COMPARISON OF DIFFERENT CONFIGURATIONS IN TERMS OF PARAMETERS AND ACCURACY

Configuration	Parameters (Million)	Accuracy (%)
0 DSC + 1 Conv	0.445	91
Proposed (1 DSC + 1 Conv)	0.878	92
2 DSC + 1 Conv	0.969	90
3 DSC + 1 Conv	1.061	89
4 DSC + 1 Conv	1.151	86
0 DSC + 2 Conv	1.221	92
1 DSC + 2 Conv	1.653	91

TABLE III
TRADE-OFF ANALYSIS-ACCURACY, INFERENCE TIME, PARAMETERS

Config.	Enh. Pyr. Block	# Dense Layers in SE Block	Acc. (%)	Inf. Time	Param.
1	1	2	83.9	18.3	0.051
2	2	2	90.13	30.3	0.22
3	3	1	90.87	40.3	0.94
4	3	2	92	40.2	0.87
5	3	3	91.11	40.2	0.90

F. Computational Efficiency

The proposed AAPN model achieves comparable or better accuracy while reducing FLOPs by up to two orders of magnitude relative to large CNNs like VGG16 and ResNet50. When quantized to 8-bit, the model size drops to ≈0.8 MB, enabling deployment on devices with stringent memory and power budgets highlighting extreme computational efficiency

with only 0.8M parameters and 0.09G FLOPs, AAPN requires 99.4% fewer parameters and 99.4% fewer operations than VGG16, while being an order of magnitude smaller than EfficientNetB0 and MobileNetV2. The corresponding model size of 3.1 MB (in FP32) is small enough to fit comfortably within the memory of microcontrollers and low-cost edge AI accelerators. These metrics confirm that AAPN is well-suited for energy-efficient, low-latency PV fault detection in embedded and UAV-based monitoring systems.

IV. Experimental Results and Analysis

The proposed AAPN attained substantial performance across differing categorization complexities, exhibiting its adaptability, efficiency, and robustness in addressing diverse defect types of PV modules. The accuracy, loss, ROC plots, confusion matrices (Figs. 7, 8, 9, 10) obtained from 2-class, 8-class, 11-class, and 12-class categorization tasks underline the model's strengths and limitations, specifically regarding precision, recall, and F1-score (Fig. 11) across multifarious defect types. This section delivers a clear class-wise investigation and explores the architectural options contributing to the architecture's performance.

Class-wise Fault Classification Analysis: In 2-Class Classification, the AAPN model achieved an impressive 96% accuracy, demonstrating robust binary classification capabilities. The model's hierarchical pyramid architecture, enhanced with Depthwise Separable Convolutions and Squeeze-and-Excitation (SE) blocks, allowed effective focus on global and local image patterns, ensuring precise differentiation between faulty and non-faulty cases. The model attained 91% accuracy in 8-Class Classification, efficiently distinguishing among 8 fault types. Notably, Diode (Class-2) and DiodeMulti (Class-3) achieved high F1-scores of 0.99 and 0.95, respectively, due to AAPN's multi-scale feature extraction and attention mechanisms. Some overlap in features led to lower F1-scores for Cell-Multi (Class-1) at 0.86 and HotSpot-Multi (Class-5) at 0.84. Shadowing (Class-7) and offline module (Class-6) maintained robust performance with 0.92 F1-scores.

AAPN achieved 87% accuracy, effectively managing complex multi-class (11-class) classification scenarios. Diode (Class-3) and Diode-Multi (Class-4) stood out with 0.98 and 0.97 F1-scores, while Cell-Multi (Class-1) and HotSpot-Multi (Class-6) had moderately lower F1-scores of 0.79 and 0.81 due to overlapping features. Soiling (Class-9) faced challenges, achieving a recall of 0.58 and an F1-score of 0.72, highlighting a need for additional tuning. The model achieved an overall accuracy of 92%, excelling in highly complex 12-class classifications. Diode (Class-3) and No-fault (Class-7) achieved 0.98 and 0.97 F1-scores, respectively. Classes such as Soiling (Class-10) and HotSpot (Class-5) had lower recall (0.64) and F1-scores (0.84), indicating overlapping features. Shadowing (Class-9) and Cell-Multi (Class-1) achieved 0.88 and 0.77 F1-scores, suggesting the potential for improvement with additional data or tuning. AAPN consistently delivers high accuracy across various classification complexities, leveraging advanced feature extraction and attention mechanisms. However, fine-tuning and targeted data augmentation could enhance performance for challenging categories.

TABLE IV
Class-wise Accuracy and Parameter Utilization Comparison

Ref	Model name	Classification Accuracy (in %) 2c	8c	11c	12c	parameters (million)
[6]	VGG16	83	-	-	68	14.7 M
[6]	Mobile Net	73	-	-	57	42 M
[6]	ResNet50	84	-	-	69	23.6 M
[6]	SVM	88.7	-	-	73.8	-
[6]	Raw Data	90.6	-	-	76.7	0.1 M
[6]	DA	93.3	-	-	84.1	0.1 M
[6]	DA + Focal loss	92.3	-	-	82.3	0.1 M
[6]	DA + Oversample	93.5	-	-	82	0.1 M
[6]	Ensemble model	94.4	-	-	85.9	15.75 M
[4]	CNN	92.5	78.85	66.43	-	-
[7]	CNN	-	-	-	85.4	2.76 M
[7]	VGG16	-	-	-	70	14.7 M
[7]	Mobile Net	-	-	-	57	4.2 M
[7]	ResNet50	-	-	-	78	23.6 M
[7]	RF	-	-	-	56.1	-
[7]	SVM-PCA	-	-	-	69.4	-
[9]	Raw Data	93.12	-	68.4	-	-
[10]	MobileNet-V3	96.6	-	-	-	5.4 M
[10]	I-MobileNet-V3	97.8	-	70.82	-	3.4 M
[26]	ResNet50	93.8	88.96	-	88.69	25.6 M
[26]	VGG16	94.65	89.48	-	89.17	138 M
[27]	SqueezeNet	85.02	-	75.15	-	-
[27]	ShuffleNet	91.5	-	82.94	-	-
[27]	AlexNet	93.2	-	-	-	-
[28]	ShuffleNet V2	-	-	84.06	-	-
Proposed Model		**96**	**91**	**87**	**92**	**0.8 M**

TABLE V
Accuracy Values obtained for 8-class Classification

Ref.	Model	Accuracy (%)	Parameters
[29]	K-means & Inception & residual	89	0.9 M
[30]	Lightweight inception residual	89	-
[29]	VGG16	76.27	134.3 M
[29]	VGG19	76.25	139.6 M
[29]	GoogleNet	75.17	6.6 M
[29]	ResNet18	78.3	11.6 M
[29]	EfficientNet	81.75	4 M
[29]	LIRNet	89.1	0.9 M
Proposed Model		**91**	**0.8M**

Performance Evaluation of the Model on Different Data Splits: The data splitting approach utilised consists of three different schemes: 80% for training, 15% for testing, and 5% for validation; 80% for training, 10% for testing, and 10% for validation; and 70% for training, 20% for testing, and 10% for validation. Over these splits, the model continually exhibited higher classification accuracy, especially in the 2-class, where accuracy varied from (95–96)% across the splits. The 8-class had comparably below-par performance, with accuracy varying from (88–91)%. Conversely, the 11c and 12c classes exhibited lower accuracy values than the 2-class classification, although maintained good performance, with 11c accuracy varying between (85–88)% and 12c accuracy between (90–92)%. The results demonstrate that varying class complexities impact the model's classification efficacy, with fewer intricate classes obtaining higher accuracy rates. The model showed robust performance in terms of precision,

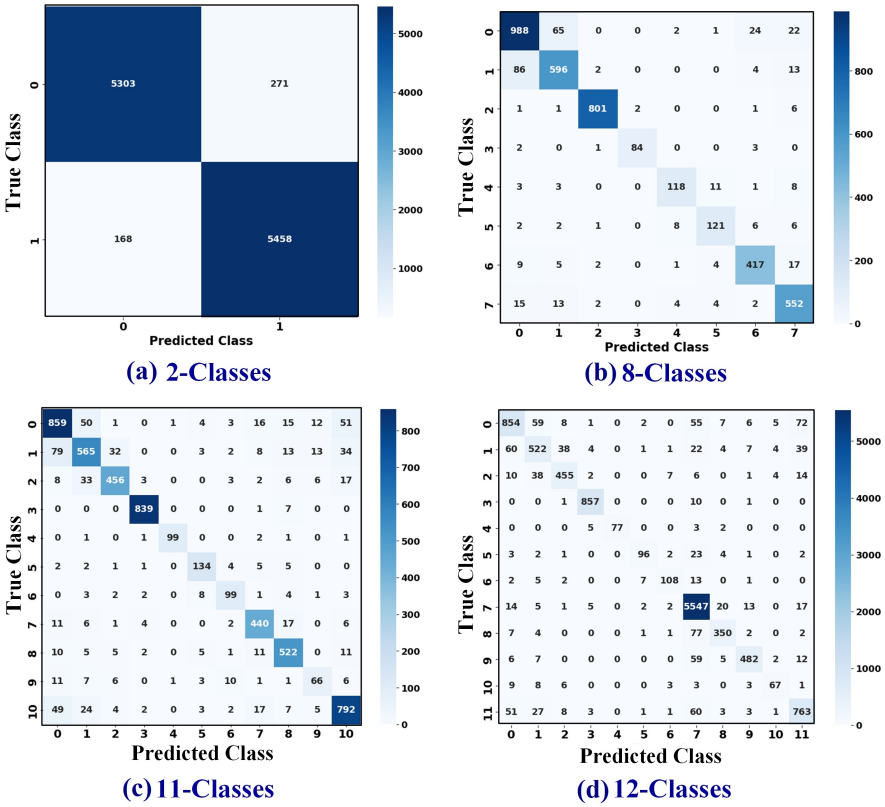

Fig. 7. Confusion matrices obtained for (a) 2-class, (b) 8-class, (c) 11-class and (d) 12-class fault classification tasks.

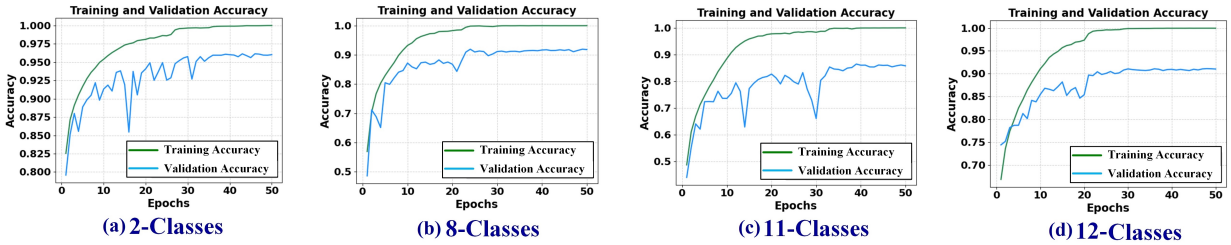

Fig. 8. Training and validation accuracy plots obtained for (a) 2-class, (b) 8-class, (c) 11-class and (d) 12-class fault classification tasks.

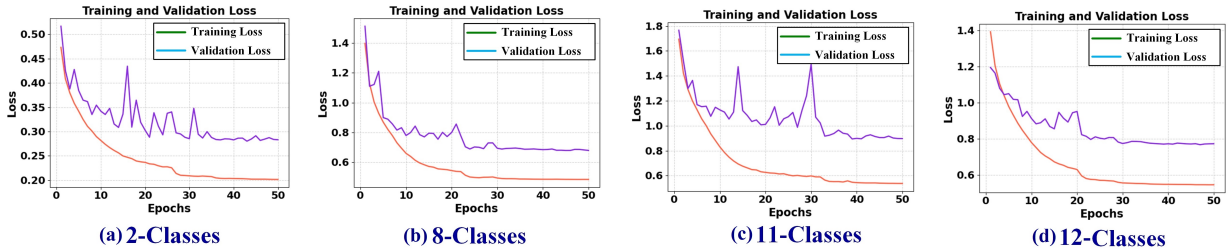

Fig. 9. Training and validation loss plots obtained for (a) 2-class, (b) 8-class, (c) 11-class and (d) 12-class fault classification tasks.

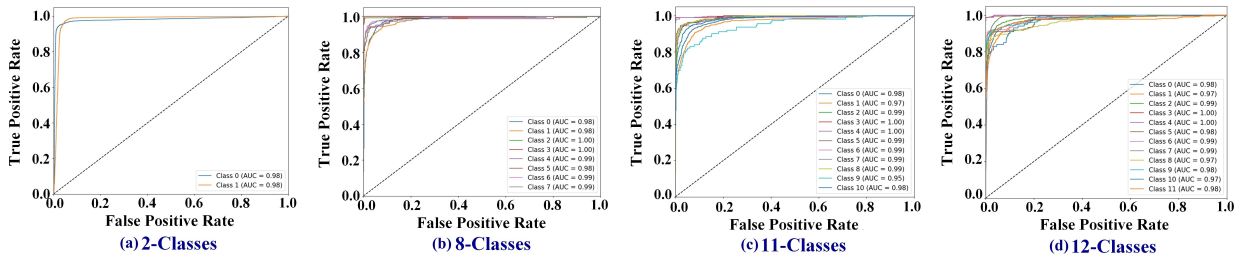

Fig. 10. ROC curves obtained for (a) 2-class, (b) 8-class, (c) 11-class and (d) 12-class fault classification tasks.

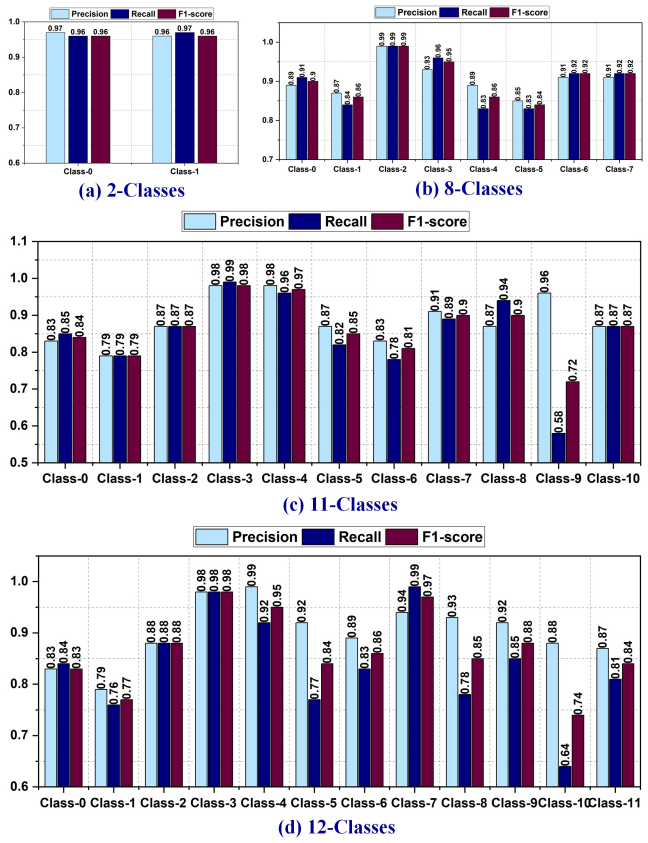

Fig. 11. Precision, recall, F1-score values obtained for (a) 2-class, (b) 8-class, (c) 11-class and (d) 12-class fault classification tasks.

recall, and F1-score across the majority of classes. The 2-class continually attained superior results for all metrics, with F1-scores varying between 96% and 96% across the splits, confirming the model's competence in balancing precision and recall successfully. In 8-class classification, the performance was marginally inferior, with F1-scores between 88% and 91%, implying that the model encountered challenges with identifying certain defects. The results reveal that the model exhibits strong performance across numerous splits; nonetheless, the intricacy of the challenge of classification influences performance, especially for more difficult fault classifications such as 8c, 11c, and 12c.

V. Critical Performance Comparison Analysis

The comprehensive comparison of accuracies and parameter count utilized by distinct models are given in Table.IV, Table V, and Table VII.

The comprehensive comparison of models highlights the trade-off between parameter count and accuracy across fault classification tasks. Traditional models like VGG16 and ResNet50 achieve decent accuracy in certain categories (e.g., 94.65% for VGG16 in 2-class) but exhibit inconsistencies and high computational costs, with parameter counts of 138M and 25.6M, respectively. Lightweight models like MobileNet (4.2M parameters) prioritize efficiency but struggle with accuracy, achieving only 57% in 12-class classification. Similarly, raw data and simple augmentation methods use minimal parameters (0.1M) but deliver inadequate accuracy (76.7% in 12-class), making them unsuitable for precision-critical applications. Improved MobileNet-V3 offers better performance (e.g., 97.8% in 2-class) but lacks generalization across tasks. In contrast, the proposed AAPN achieves exceptional accuracy—96% (2-class), 91% (8-class), 87% (11-class), and 92% (12-class)—with the lowest parameter count (0.8M). Its consistent, high performance and computational efficiency make it ideal for real-time, scalable PV monitoring systems, outperforming existing models by balancing accuracy and efficiency effectively.

VI. Implementation of the Proposed AAPN Model on Google Coral Edge TPU

The proposed AAPN was successfully executed on the Google Coral Edge TPU board to test its real-time performance and practicality (Fig. 13) for deployment in resource-limited environments.

The Edge TPU is a specialized AI-accelerator devised for low-power, high-speed inference, making it an exemplary choice for operating lightweight models like AAPN. After quantizing the model into an 8-bit TensorFlow Lite format, an impressive inference time of 40.2 milliseconds per image has been attained during the fault class classification as noted in Table.VIII. This rapid inference speed is essential for applications like real-time fault detection in solar PV systems, where swift and efficient processing is needed for on-site diagnostics. Besides the low inference time, the proposed model established energy efficiency, consuming only 2 watts during inference on the Edge TPU, making it highly suited for autonomous drone-based monitoring applications or other edge-based solutions that need both real-time processing and low power consumption. The compact model size, with only 0.87 million parameters, additionally contributes to the model's efficiency, qualifying it to be deployed in environments with less computational resources. This implementation emphasizes the practicality of AAPN for real-time, energy-efficient anomaly detection in embedded and industrial environments, delivering a potent solution for predictive maintenance in PV systems and other edge-based applications.

The accuracy and parameter count trade-off comparison of models is shown in Fig. 12. The proposed AAPN model is highly suitable for FPGA-based drone deployment [32] in solar fault monitoring due to its remarkably low parameter count (0.8M) and exceptional classification accuracy of 96%. Unlike conventional deep learning models such as VGG16 (138M parameters) and ResNet50 (25.6M parameters), AAPN achieves outstanding accuracy while significantly lowering computational overhead, making it ideal for FPGA-based edge devices with limited resources. By employing depthwise separable convolutions and squeeze-and-excitation blocks, the model guarantees efficient feature extraction while maintaining low latency, allowing drones to process infrared images in real-time without relying on cloud-based inference. The parallel execution on FPGA hardware enables onboard AI-driven fault detection, facilitating fast decision-making and autonomous adjustments based on detected panel defects. The

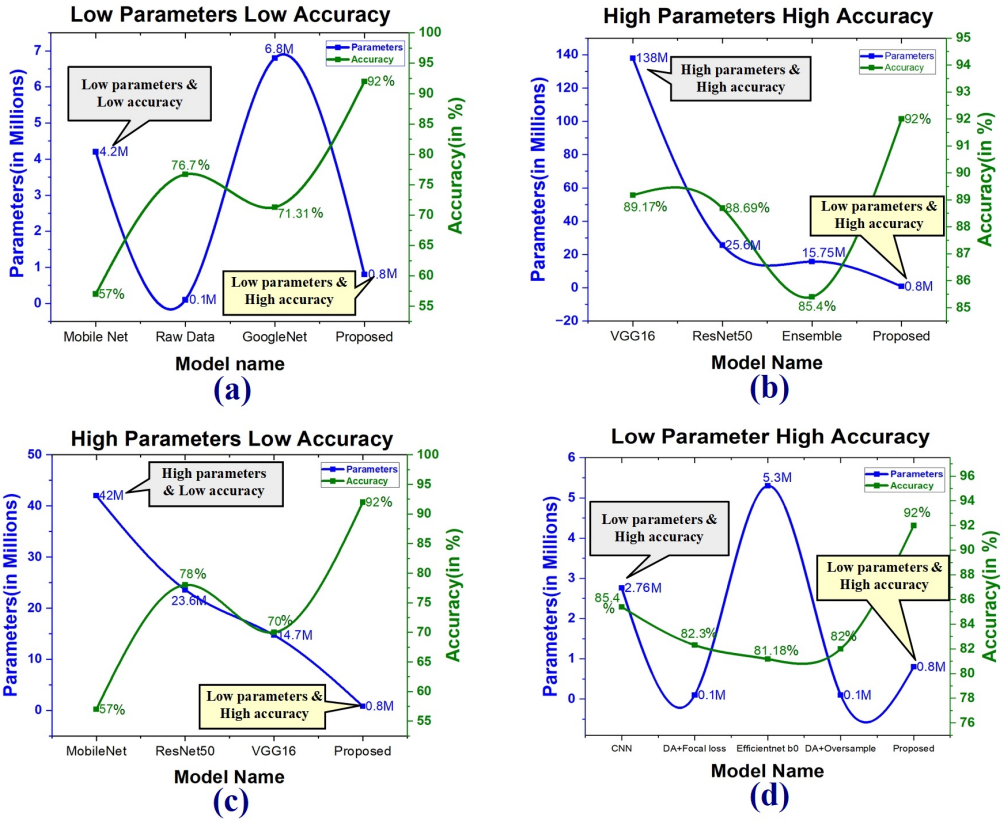

Fig. 12. Comparison of model performance in terms of parameter count and classification accuracy for different architectures: (a) Low parameters with low accuracy, (b) High parameters with high accuracy, (c) High parameters with low accuracy, and (d) Low parameters with high accuracy.

TABLE VI
PERFORMANCE EVALUATION OF THE MODEL ON DIFFERENT DATA SPLITS

Class	X_train	X_test	X_val	Total Images	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
X_train=80%, X_test=15%, X_val=5%								
2-Class	64,000	12,000	4,000	80,000	95	96	96	96
8-Class	23,094	4,330	1,443	28,868	88	90	88	89
11-Class	32,000	6,000	2,000	40,000	85	87	82	84
12-Class	64,000	12,000	4,000	80,000	90	88	82	85
X_train=80%, X_test=10%, X_val=10%								
2-Class	64,000	8,000	8,000	80,000	96	96	96	96
8-Class	23,094	2,886	2,886	28,868	91	91	90	91
11-Class	32,000	4,000	4,000	40,000	88	89	86	87
12-Class	64,000	8,000	8,000	80,000	92	89	84	86
X_train=70%, X_test=20%, X_val=10%								
2-Class	56,000	16,000	8,000	80,000	96	96	96	96
8-Class	20,207	5,773	2,886	28,868	90	90	88	89
11-Class	28,000	8,000	4,000	40,000	87	87	84	85
12-Class	56,000	16,000	8,000	80,000	90	87	81	84

lightweight architecture of AAPN makes it especially efficient for battery-powered UAVs, reducing energy consumption and extending flight time, which is crucial for large-scale solar farm inspections. In contrast to other CNN architectures that demand high-end GPUs, AAPN is optimized for FPGA deployment, balancing low memory usage, computational efficiency, and high accuracy. This provides scalability for fleets of autonomous drones capable of efficiently surveying and monitoring extensive solar farms. With onboard AI inference, lower power consumption, and industry-leading classification accuracy, the proposed AAPN model proves to be the most efficient deep learning architecture for real-time, FPGA-driven UAV-based PV monitoring applications.

VII. LIMITATIONS AND FUTURE SCOPE

While the dataset utilized in this analysis is comprehensive and encompasses a combination of fault classes, to address

TABLE VII
ACCURACY VALUES (%) OBTAINED FOR 12-CLASS CLASSIFICATION

Ref	Model	Accuracy	Ref	Model	Accuracy
[8]	Resnet18	74.83	[8]	VGG16	73.94
[8]	Resnet50	79.15	[8]	VGG19	73.91
[8]	Resnet101	79.12	[8]	Squeezenet	74.68
[8]	Darknet19	76.99	[31]	CE	86.9
[8]	Mobilenetv2	77.47	[31]	LDAM Loss	86.8
[8]	Darknet53	76.2	[31]	Balancedsoftmax	86.3
[8]	Xception	77.52	[31]	CE+DRW	85.9
[8]	Efficientnetb0	81.18	[31]	Focal loss	87.3
[8]	Shufflenet	78.1	[31]	LDAW+DRW	86
[8]	Nasnetmobile	75.7	[31]	CE+CMO	86.2
[8]	Nasnetlarge	77.09	[31]	Manifoldmixup	87
[8]	Densent201	77.86	[31]	PatchUp	86.8
[8]	Inceptionv3	76.63	[31]	Cutmix	86
[8]	Inceptionresnetv2	79.42	[31]	Mixup	85.5
[8]	Googlenet	71.31	[31]	SaliencyMix	86.2
[8]	Alexent	77.3	[31]	Class balance	87.6
	Proposed Model				**92**

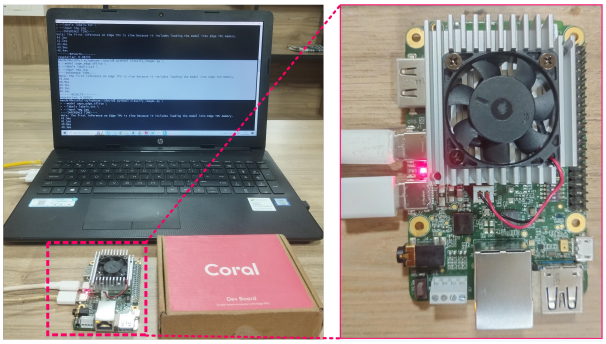

Fig. 13. Implementation of proposed AAPN model on Edge TPU.

TABLE VIII
CORAL DEV BOARD SPECIFICATIONS AND PROPOSED AAPN INFERENCE PERFORMANCE

Parameter	Value / Description
Model Format	TensorFlow Lite (.tflite, quantized and compiled for Edge TPU)
Inference Latency	40.2 ms per image (Edge TPU with Proposed AAPN.tflite model)
Host Processor	NXP i.MX 8M Quad-core Cortex-A53
Edge Accelerator	Google Edge TPU (4 TOPS at 0.5 W per TOP)
Power Consumption	2 Watts during inference
RAM	512 MB LPDDR4
Storage	8 GB eMMC + microSD card support
Operating System	Mendel Linux (Debian-based)

the generalization concerns, as the present experiments were performed on a single infrared dataset, future work will concentrate on cross-dataset validation to evaluate the model's capability to generalize across heterogeneous real-world conditions, including divergences in sensor types, environmental factors, and image resolutions. We will furthermore examine domain adaptation and transfer learning methods to improve the model's robustness and performance on diverse datasets demonstrating the model's ability for real-world deployment in various PV monitoring environments, providing its scalability and effectiveness across a wide spectrum of operational conditions. The Google Coral Dev Board was selected for its real-time inference efficiency and practicality for lightweight models, such as AAPN, which has 0.8 million parameters and an inference time of 40.2 ms/image, showcasing its potential for FPGA-based processors. Nevertheless, FPGA-specific metrics, such as LUT utilization and power consumption, are important for further validation. The evolution of a hardware-accelerated FPGA architecture will be examined in future work, building on the results from the Edge TPU board.

VIII. CONCLUSIONS

The proposed Adaptive Attention Pyramid Network (AAPN) successfully addressed the challenges of existing fault detection models, achieving high classification accuracy across complex multi-class scenarios. The integration of depthwise separable convolutions and SE blocks in the pyramid architecture enables efficient multi-scale feature extraction, effectively handling the diversity and complexity of PV module faults. Experimental results highlight the model's exceptional accuracies, with up to 96% in binary classification, 91% in 8-class, 87% in 11-class, and 92% in 12-class fault classifications. Further, the proposed model is tested using performance evaluation metrics such as accuracy, loss, precision, recall, F1-score, confusion matrix, ROC-AUC Curves for fault categories. Besides, its effectiveness is compared with existing deep-learning models for various fault classes. An ablation study was perfomed to validate the impact of different architectural components, such as depthwise separable convolutions and squeeze-and-excitation blocks, on the model's performance, showing an optimal trade-off between accuracy and computational efficiency. Additionally, the model has been successfully implemented on the Google Coral Edge TPU, achieving 40.2 ms inference time per image, confirming its efficiency and suitability for real-time applications in resource-constrained environments. The proposed architecture model is very lightweight, utilizing only 0.8M parameters. Its effective balance between high accuracies and low parameter utilization makes it highly suitable for deployment on drone-based edge devices, facilitating on-site real-time PV fault monitoring, maintenance, and detection.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] B. Sezen, C. C. Cerasi *et al.*, "Solar cell busbars surface defect detection based on deep convolutional neural network," *IEEE Latin America Transactions*, vol. 21, no. 2, pp. 242–250, 2023. [Online]. Available: doi:10.1109/TLA.2023.10015216

[2] N. S. Maddileti, R. Namburi, R. D. A. Raj, R. M. R. Yanamala, and A. Pallakonda, "Dcgan-driven minority class augmentation for lightweight yolo-based photovoltaic defect localization suitable for edge deployment," *IEEE Transactions on Device and Materials Reliability*, 2025. [Online]. Available: doi:10.1109/TDMR.2025.3592416

[3] A. E. Nieto, F. Ruiz, D. Patino, and O. Ramirez, "Classification of electric faults in photovoltaic systems based on voltage-power curves," *IEEE Latin america transactions*, vol. 19, no. 12, pp. 2071–2078, 2021. [Online]. Available: doi:10.1109/TLA.2021.9480149

[4] A. Guisández Hernández and S. P. Santos, "Modelling and experimental validation of aging factors of photovoltaic solar cells," *IEEE Latin America Transactions*, vol. 19, no. 8, pp. 1270–1277, 2021. [Online]. Available: doi:10.1109/TLA.2021.9475857

[5] L. R. Yeddula, A. Pallakonda, R. D. A. Raj, R. M. R. Yanamala, K. K. Prakasha, and M. S. Kumar, "Yolov8n-gbe: A hybrid yolov8n model with ghost convolutions and bifpn-eca attention for solar pv defect localization," *IEEE Access*, 2025. [Online]. Available: doi:10.1109/ACCESS.2025.3584249

[6] M. Le, D. K. Nguyen, V.-D. Dao, N. H. Vu, H. H. T. Vu *et al.*, "Remote anomaly detection and classification of solar photovoltaic modules based on deep neural network," *Sustainable Energy Technologies and Assessments*, vol. 48, p. 101545, 2021. [Online]. Available: https://doi.org/10.1016/j.seta.2021.101545

[7] M. Le, D. Le, and H. H. T. Vu, "Thermal inspection of photovoltaic modules with deep convolutional neural networks on edge devices in auv," *Measurement*, vol. 218, p. 113135, 2023. [Online]. Available: https://doi.org/10.1016/j.measurement.2023.113135

[8] Z. B. Duranay, "Fault detection in solar energy systems: A deep learning approach," *Electronics*, vol. 12, no. 21, p. 4397, 2023. [Online]. Available: https://doi.org/10.3390/electronics12214397

[9] 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, p. 4918, 2023. [Online]. Available: https://doi.org/10.3390/s23104918

[10] C. Tang, H. Ren, J. Xia, F. Wang, and J. Lu, "Automatic defect identification of pv panels with ir images through unmanned aircraft," *IET Renewable Power Generation*, vol. 17, no. 12, pp. 3108–3119, 2023. [Online]. Available: https://doi.org/10.1049/rpg2.12831

[11] X. Zhao, C. Song, H. Zhang, X. Sun, and J. Zhao, "Hrnet-based automatic identification of photovoltaic module defects using electroluminescence images," *Energy*, vol. 267, p. 126605, 2023. [Online]. Available: https://doi.org/10.1016/j.energy.2022.126605

[12] X. Chen, T. Karin, and A. Jain, "Automated defect identification in electroluminescence images of solar modules," *Solar Energy*, vol. 242, pp. 20–29, 2022. [Online]. Available: https://doi.org/10.1016/j.solener.2022.06.031

[13] Á. H. Herraiz, A. P. Marugán, and F. P. G. Márquez, "Photovoltaic plant condition monitoring using thermal images analysis by convolutional neural network-based structure," *Renewable Energy*, vol. 153, pp. 334–348, 2020. [Online]. Available: https://doi.org/10.1016/j.renene.2020.01.148

[14] W. Gao and R.-J. Wai, "A novel fault identification method for photovoltaic array via convolutional neural network and residual gated recurrent unit," *IEEE access*, vol. 8, pp. 159 493–159 510, 2020. [Online]. Available: DOI:10.1109/ACCESS.2020.3020296

[15] B. Su, H. Chen, P. Chen, G. Bian, K. Liu, and W. Liu, "Deep learning-based solar-cell manufacturing defect detection with complementary attention network," *IEEE Transactions on Industrial informatics*, vol. 17, no. 6, pp. 4084–4095, 2020. [Online]. Available: DOI:10.1109/TII.2020.3008021

[16] L. Li, Z. Wang, and T. Zhang, "Gbh-yolov5: Ghost convolution with bottleneckcsp and tiny target prediction head incorporating yolov5 for pv panel defect detection," *Electronics*, vol. 12, no. 3, p. 561, 2023. [Online]. Available: https://doi.org/10.3390/electronics12030561

[17] J. Fioresi, D. J. Colvin, R. Frota, R. Gupta, M. Li, H. P. Seigneur, S. Vyas, S. Oliveira, M. Shah, and K. O. Davis, "Automated defect detection and localization in photovoltaic cells using semantic segmentation of electroluminescence images," *IEEE Journal of Photovoltaics*, vol. 12, no. 1, pp. 53–61, 2021. [Online]. Available: DOI:10.1109/JPHOTOV.2021.3131059

[18] X. Qian, J. Li, J. Cao, Y. Wu, and W. Wang, "Micro-cracks detection of solar cells surface via combining short-term and long-term deep features," *Neural Networks*, vol. 127, pp. 132–140, 2020. [Online]. Available: https://doi.org/10.1016/j.neunet.2020.04.012

[19] Q. Liu, M. Liu, C. Wang, and Q. J. Wu, "An efficient cnn-based detector for photovoltaic module cells defect detection in electroluminescence images," *Solar Energy*, vol. 267, p. 112245, 2024. [Online]. Available: https://doi.org/10.1016/j.solener.2023.112245

[20] Y. Jia, G. Chen, and L. Zhao, "Defect detection of photovoltaic modules based on improved varifocalnet," *Scientific Reports*, vol. 14, no. 1, p. 15170, 2024. [Online]. Available: https://doi.org/10.1038/s41598-024-66234-3

[21] A. Korovin, A. Vasilev, F. Egorov, D. Saykin, E. Terukov, I. Shakhray, L. Zhukov, and S. Budennyy, "Anomaly detection in electroluminescence images of heterojunction solar cells," *Solar Energy*, vol. 259, pp. 130–136, 2023. [Online]. Available: https://doi.org/10.1016/j.solener.2023.04.059

[22] Y. Cao, D. Pang, Q. Zhao, Y. Yan, Y. Jiang, C. Tian, F. Wang, and J. Li, "Improved yolov8-gd deep learning model for defect detection in electroluminescence images of solar photovoltaic modules," *Engineering Applications of Artificial Intelligence*, vol. 131, p. 107866, 2024. [Online]. Available: https://doi.org/10.1016/j.engappai.2024.107866

[23] E. A. Mejia, H. L. Correa, É. Mejía, R. Girón, A. David, N. Rodriguez, and S. Esperanza, "Photovoltaic system thermal images," *Mendeley Data: Amsterdam, The Netherlands*, 2019. [Online]. Available: DOI:10.17632/82vzccxb6y.2

[24] L. Li, Z. Wang, and T. Zhang, "Gbh-yolov5: Ghost convolution with bottleneckcsp and tiny target prediction head incorporating yolov5 for pv panel defect detection," *Electronics*, vol. 12, no. 3, p. 561, 2023. [Online]. Available: https://doi.org/10.3390/electronics12030561

[25] W. Tang, Q. Yang, K. Xiong, and W. Yan, "Deep learning based automatic defect identification of photovoltaic module using electroluminescence images," *Solar Energy*, vol. 201, pp. 453–460, 2020. [Online]. Available: https://doi.org/10.1016/j.solener.2020.03.049

[26] E. 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," *Energy Conversion and Management*, vol. 314, p. 118718, 2024. [Online]. Available: https://doi.org/10.1016/j.enconman.2024.118718

[27] D. Korkmaz and H. Acikgoz, "An efficient fault classification method in solar photovoltaic modules using transfer learning and multi-scale convolutional neural network," *Engineering Applications of Artificial Intelligence*, vol. 113, p. 104959, 2022. [Online]. Available: https://doi.org/10.1016/j.engappai.2022.104959

[28] D. Sriraman and R. Ramaprabha, "Application of machine learning and convolutional neural networks for the fault detection and classification monitoring system in pv plants," in *2023 9th International Conference on Electrical Energy Systems (ICEES)*. IEEE, 2023, pp. 694–699. [Online]. Available: doi:10.1109/ICEES57979.2023.10110284

[29] S.-H. Lee, L.-C. Yan, and C.-S. Yang, "Lirnet: A lightweight inception residual convolutional network for solar panel defect classification," *Energies*, vol. 16, no. 5, p. 2112, 2023. [Online]. Available: https://doi.org/10.3390/en16052112

[30] H. Chen, A. Zhang, C. Gong, W. Liang, and Z. Wang, "Fault diagnosis method for photovoltaic panels based on improved shufflenet v2 and infrared images," in *2022 7th International Conference on Power and Renewable Energy (ICPRE)*. IEEE, 2022, pp. 447–451. [Online]. Available: doi:10.1109/ICPRE55555.2022.9960396

[31] S. Wu, Y. Kong, R. Xu, Y. Guo, Z. Chen, and X. Zheng, "A feature space class balancing strategy-based fault classification method in solar photovoltaic modules," *Engineering Applications of Artificial Intelligence*, vol. 136, p. 108991, 2024. [Online]. Available: https://doi.org/10.1016/j.engappai.2024.108991

[32] S. S. Priya, P. S. Sanjana, R. M. R. Yanamala, R. D. Amar Raj, A. Pallakonda, C. Napoli, and C. Randieri, "Flight-safe inference: Svd-compressed lstm acceleration for real-time uav engine monitoring using custom fpga hardware architecture," *Drones*, vol. 9, no. 7, p. 494, 2025. [Online]. Available: https://doi.org/10.3390/drones9070494

Rayappa David Amar Raj is currently working as an Assistant Professor in Amrita School of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Coimbatore, Chennai, India. He completed his Ph.D from NIT Warangal, India. He holds a Master's degree from NIT Agartala, India. His research interests include fault detection and classification in PV systems, power systems, smart grids, image processing, image encryption, neural networks, deep learning, dynamic reconfiguration strategies, PV array optimization, and cybersecurity.

Rama Muni Reddy Yanamala is working as an Assistant Professor in Department of Electronics and Communication Engineering, Indian Institute of Information Technology Design and Manufacturing (IIITD&M), Kancheepuram. He completed his Ph.D from National Institute of Technology, Warangal, India. He holds a Master's degree from the Indian Institute of Technology, Madras, India. His areas of research include Hardware Accelerators for deep learning applications, FPGA-based designs, IoT based edge device applications, Embedded projects, RISC-v debugger design using Bluespec language.

Archana Pallakonda is a research scholar in National Institute of Technology, Warangal, India. She holds a Master's degree from Kakatiya University, Warangal, India. Her areas of research include Biometric security, machine learning, cybersecurity, deep learning, cryptography, image processing, blockchain, transformation techniques, cloud computing, network security, bioinformatics.

Kanasottu Anil Naik is a faculty at National Institute of Technology, Warangal, India. He holds PhD and Master's degree from the IIT Roorkee and NIT Hamirpur respectively. His areas of research include renewable energy systems grid integration, intelligent controller applications in power system, micro-grid stability and control.