An International Journal of Optimization and Control: Theories & Applications
ISSN: 2146-0957 eISSN: 2146-5703
Vol.16, No.6, pp.2218-2248 (2026)
https://doi.org/10.36922/IJOCTA026320204

RESEARCH ARTICLE

Electroluminescence image-based photovoltaic defect severity classification using convolutional neural network architectures and ensemble learning

Fouad Sleibi[1]*, Emma Qumsiyeh[1], and Nahla E. Zakzouk[2]

[1]Department of Information Technology, Faculty of Engineering and Information Technology, Palestine Ahliya University, Bethlehem, Palestine

[2]Department of Electrical and Control Engineering, College of Engineering and Technology, Arab Academy for Science and Technology and Maritime Transport, Alexandria, Egypt

fuadsleibi@gmail.com; e.qumsiyeh@paluniv.edu.ps; nahlaezzeldin@aast.edu

ARTICLE INFO

Article history:
Received: August 7, 2026
Revised: October 6, 2026
Accepted: October 8, 2026
Published online: October 19, 2026

Keywords:
Photovoltaic defect diagnosis
Electroluminescence imaging
Convolutional neural networks
Ensemble learning
Imbalanced datasets
Macro-F1 score
Classification accuracy

ABSTRACT

Accurate fault detection and defect severity classification in photovoltaic (PV) modules are essential for predictive maintenance and reliable energy production. Electroluminescence (EL) imaging provides a fast, non-contact, and effective approach for identifying internal defects in PV cells, including cracks, inactive regions, and structural abnormalities. However, automatic severity classification remains challenging due to class imbalance and subtle visual differences between intermediate defect levels. This study investigated transfer learning using several convolutional neural network (CNN) architectures, including ResNet18, ResNet34, VGG16, VGG19, and EfficientNet-B0, for multi-class PV defect severity classification using publicly available EL images within a unified experimental framework. Among the individual models, VGG16 achieved the best test performance with an accuracy of 74.54% and a macro-F1 score of 0.56, indicating that single CNN models remain insufficiently robust for reliable multi-class severity classification, particularly under class imbalance and overlapping severity levels. To address these limitations, ensemble learning was employed to combine complementary CNN models, significantly improving classification performance to approximately 90% accuracy and 0.84 macro-F1. In addition, gradient-weighted class activation mapping analysis was used to provide qualitative insight into the regions of the EL images that influenced model predictions. The ensemble models produced visually consolidated and interpretable attention patterns across the representative samples compared with the individual models. Overall, the proposed ensemble-based framework provides a practical and computationally efficient approach for improving PV defect severity classification using EL imaging.

CC BY

1. Introduction

Photovoltaic (PV) technology has gained prominence as one of the most important renewable energy sources because of its environmental advantages, scalability, and growing contribution to sustainable energy generation.[1] With the ongoing trend of solar system implementation on a global scale, the long-term reliability, efficiency, and operational stability of PV modules have become significant technical and economic issues. Since PV modules are often affected by many factors such as environment, materials, and processes, any fault in solar cells can reduce power generation, accelerate module degradation, and increase maintenance costs if not detected early.[2] Therefore, accurate identification of defect severity is essential for enabling effective predictive maintenance and reducing operational risks.[3]

*Corresponding Author

Photovoltaic fault detection techniques (FDTs) include visual inspection, model-based techniques, and data-driven approaches.[2-5] Among different PV FDTs, data-driven approaches depend mainly on collecting and analyzing large datasets representing both normal and faulty operating conditions.[6] These methods learn the relationship between input and output signals directly from the data, making them more flexible and suitable for real-world PV fault diagnosis applications.[7]

Data-driven PV FDTs utilize different types of data, which can be mainly categorized as electrical measurements, imaging data, and environmental parameters.[3] Electrical signals, including current and voltage measurements, are commonly used to evaluate the operational condition of PV modules and detect abnormalities related to electrical or thermal faults. Environmental data, including temperature, humidity, and irradiance, are used for fault detection since changing weather conditions can significantly influence PV performance. As per image-based data, they offer a non-invasive, fast, and efficient monitoring approach by capturing visible or thermal abnormalities such as cracks, hot spots, discoloration, and poor cell connections.[8]

Hence, among data-driven approaches, image-based fault detection has gained significant importance, especially in large-scale or complex PV installations where manual inspection is difficult due to its fast inspection capability, non-contact operation, and ability to automatically identify defects using advanced image processing and deep learning (DL) techniques.[9] Top of FormBottom of Form Among image-based PV FDTs, electroluminescence (EL) imaging has become one of the most accurate techniques for identifying internal defects and microcracks in PV cells.[10] EL images may reveal cracks, inactivity, and finger breakages, along with other structural abnormalities that might not be evident during normal observation. Manual interpretation of EL images is, however, time-consuming, subjective, and difficult to scale in industry, where large numbers of modules have to be inspected regularly. As a result, machine learning (ML) and DL have been increasingly used to automatically analyze defects over the past several years.[11-13]

Deep convolutional neural networks (CNNs) have proven quite effective for image classification tasks because of their capacity to learn hierarchical visual features directly from raw image data.[14] Hence, CNN methods provide a viable solution to the PV inspection problem, where defect warning patterns and severity across EL images are to be detected. However, multi-class severity classification remains challenging in the case of small, visually challenging, and imbalanced PV datasets where intermediate faults exist, as the visual patterns may reflect a mix of normal and severely malfunctioning cells.[15,16] Most existing studies focus on binary classification or large-scale datasets, with limited attention to imbalanced multi-class severity classification. Hence, in the case of an imbalanced dataset, these classification models tend to be biased toward majority classes and struggle to capture subtle differences between adjacent severity levels.

In this paper, these issues are addressed by investigating the classification performance of various pretrained CNN backbones[17], to classify multi-class PV defect severity (functional, slight, moderate, severe) based on the EL images dataset that features some imbalanced intermediate-severity labels. The electroluminescence photovoltaic (ELPV) dataset is publicly available and is used in this study as the experimental benchmark.[18] Furthermore, two adaptation strategies are addressed: full fine-tuning and two-stage fine-tuning.[19] Every single model is trained with a common preprocessing, augmentation, optimization, and evaluation setup to enable objective comparison. Since the task is an imbalanced multi-class classification problem, the analysis focused on macro-F1 and per-class recall, rather than overall accuracy.[20] Additionally, gradient-weighted class activation mapping (Grad-CAM) is incorporated to provide qualitative visual interpretations of model attention and to analyze model behavior across severity levels.[21] Finally, ensemble learning strategies are also investigated to overcome the limitations of single models by combining the most effective architectures. Ensemble learning has been widely shown to improve classification performance in complex and imbalanced tasks[22], thus improving detection of minority classes and achieving more balanced performance across severity levels by leveraging complementary model predictions.

Since the combined investigation of CNN comparison, transfer learning, ensemble learning, and explainability on imbalanced multi-class ELPV datasets remains insufficiently explored, the main contributions of this work are summarized as follows:

(i) A comprehensive and fair comparison of several pretrained CNN architectures, including ResNet18, ResNet34, VGG16,

VGG19, and EfficientNet-B0, under a unified experimental framework for multi-class severity classification using the ELPV dataset. This study aims to highlight the stability, generalization capability, and effectiveness of pretrained CNN models in distinguishing different defect severity levels. The comparison also highlights the continuing difficulty of identifying intermediate-severity labels.

(ii) An investigation of different transfer learning approaches, associated with training data enhancement using a weighted random sampler (WRS), as well as full fine-tuning and staged fine-tuning, for handling imbalanced severity classification problems in PV defect analysis.

(iii) The incorporation of Grad-CAM explainability techniques to visualize model attention regions, interpret classification decisions, and analyze the difficulties associated with intermediate severity categories.

(iv) A detailed analysis of the limitations of conventional CNN models when dealing with imbalanced multi-class datasets, especially in the presence of subtle visual differences and overlapping severity levels.

(v) The development and evaluation of ensemble learning strategies to improve classification accuracy, enhance robustness, and achieve better class balance across different defect severity categories, thus addressing the limitations of individual CNN architectures.

2. Literature review

Recent studies have addressed various methodologies for classifying PV cells into four defect-probability categories using the ELPV dataset introduced in Ref.[18] These studies span multiple research directions, including preprocessing and segmentation, regression, classification frameworks, and advanced DL-based approaches. **Table 1** presents a comparison of these studies.

In Ref.[18], a pretrained VGG19 deep regression network was introduced. Although acceptable classification accuracy was achieved, severe class imbalance was an issue. In Ref.[23], a hybrid deep and ML approach was proposed in which deep features were extracted using pre-trained CNN architectures such as ResNet50, VGG16, VGG19, and DarkNet19, followed by support vector machine (SVM) classification. Although the results were good, the minority classes remained imbalanced due to small sample sizes of defective items. Whereas in Ref.[24], an enhanced hybrid DL approach is proposed where ResNet152 and Xception architectures were fused with coordinate attention, mixed pooling, and feature fusion methods to enhance feature extraction and defect localization. Despite the relatively higher achieved accuracy and enhanced class balance, this approach suffers from very high implementation complexity and computation burden, which affect its practicality. Again, high implementation complexity and computation cost were noted in Ref.[25], which proposed an approach comprising the architectures of HazELNet, SqueezeNet, and MobileNet-v2. In Ref.[26], although ensemble DL methods were introduced, they did not always outperform single models, and in all cases, low accuracies were achieved. Finally, single CNN models were introduced associated with generative adversarial network (GAN)-based augmentation in Ref.[27] Although this augmentation improved class balance, they increase computational complexity and training requirements.

Overall, studies that dealt with the considered ELPV dataset generally followed three main directions: single pretrained CNN models, highly customized hybrid architectures, and advanced augmentation-based frameworks. Although these approaches achieved promising results, several limitations remain unresolved. Single-model approaches suffered from limited robustness and poor handling of imbalanced multi-class severity levels, especially for intermediate defect categories. Hybrid architectures, combining CNNs with SVMs or advanced feature fusion techniques, improved performance but introduced higher architectural complexity, computational costs, and challenging optimization procedures. Furthermore, lightweight multi-scale CNN frameworks and GAN-based augmentation methods required extensive preprocessing, sophisticated augmentation pipelines, or high graphics processing unit (GPU) resources to achieve competitive results. Although ensemble learning has been explored previously, existing ensemble studies reported relatively low performance on highly imbalanced ELPV datasets and concluded that ensembles do not always outperform strong standalone CNNs on small datasets. In addition, limited attention was given to systematically analyzing how complementary behaviors among pretrained CNNs could improve robustness, interpretability, and minority-class

Table 1. Previous studies involving the ELPV dataset introduced in Ref.[18]

Study	Model/method	Features number	Imbalance handling	Implementation complexity	Training time	Accuracy	Limitation/gap
Buerhop-Lutz *et al.*[18]	Pretrained VGG19 deep regression network	Deep CNN features extracted using VGG19	Class down-weighting due to severe imbalance	Low	Not reported	88.42%	Severe class imbalance
Demirci *et al.*[23]	Hybrid DFB-SVM + L-CNN	1,024–5,120 features, best: selected 2,000 features using mRMR	Online & offline data augmentation + SVM robustness to imbalance	Moderate	~135 min for L-CNN	90.57%	The minority classes remained imbalanced due to small sample sizes of defective items
Wang *et al.*[24]	Hybrid ResNet152 + Xception + coordinate attention + mixed pooling + feature fusion	2,048 features from Xception + deep residual features from ResNet152	Data augmentation + class weighting	Very High	~56–80 epochs convergence	92.13%	Very high implementation complexity and computation burden
Al-Otum[25]	HazELNet + SqueezeNet + MobileNet-v2 using transfer learning and multi-scale CNN	Multi-scale features using 3×3, 5×5 kernels + deep features	Extensive data augmentation + balanced modified ELPV dataset	High	HazELNet: 50 epochs	93.3%	High implementation complexity and computation burden
Tella *et al.*[26]	Deep learning ensemble methods (voting & bagging) using AlexNet, SENet, GoogleNet, Xception, ViT, DarkNet53, ResNet18, and SqueezeNet	Deep CNN feature representations extracted from 8 architectures	Binary conversion + ensemble diversity + batch sampling	Moderate	40–50 epochs	68.36% (voting), 73.02% (binary ResNet18)	Ensemble methods did not always outperform single models
Ebied *et al.*[27]	ResNet152, DenseNet201, DenseNet169, ResNet101, GAN, transfer learning	EL image features extracted automatically by CNNs	GAN-based oversampling (AUG300)	High	High due to deep models and GAN training	90.13% using ResNet152 with GAN augmentation	Although GAN-based augmentation improves class balance, they increase computational complexity and training requirements

Abbreviations: CNN: Convolutional neural network; DFB-SVM: Deep feature-based support vector machine; EL: Electroluminescence; ELPV: Electroluminescence photovoltaic; GAN: Generative adversarial network; L-CNN: Lightweight convolutional neural network; mRMR: Minimum redundancy maximum relevance; ResNet: Residual network; SENet: Squeeze-and-excitation network; SVM: Support vector machine; VGG19: Visual Geometry Group 19-layer network; ViT: Vision transformer.

sensitivity in severity classification tasks.

Therefore, the research gap lies in the lack of a practical and computationally efficient framework that can balance classification performance, robustness, interpretability, and implementation simplicity while effectively handling imbalanced multi-class EL datasets. This motivates the proposal of an ensemble strategy that combines the complementary strengths of multiple pretrained CNN models without requiring highly complex feature-fusion architectures or expensive augmentation frameworks.

3. Methodology

This study adopted a unified experimental framework to ensure a fair, controlled comparison of different CNN architectures and transfer learning strategies, followed by the application of ensemble DL methods. All CNN models were trained and evaluated under consistent preprocessing, augmentation, optimization, and evaluation conditions, allowing performance differences to be attributed to model design rather than experimental variation.

Figure 1 shows the complete experimental architecture developed in this work to classify PV defects by severity. The workflow starts with the ELPV dataset, in which defect probabilities are categorized into four severity levels (functional, slight, moderate, and severe), followed by stratification of the data into training, validation, and test sets. Class imbalance is compensated for by weighted sampling applied only to the training set. Consistent image preprocessing, including resizing, normalization, and contrast-limited adaptive histogram equalization, is then performed, with data augmentation applied during training to enhance model robustness. Several pretrained CNNs are initialized and trained under identical settings using different fine-tuning strategies (full fine-tuning and two-stage fine-tuning), with multiple runs conducted to assess reproducibility. Results are averaged, and standard deviations are calculated. The top-performing models are selected based on

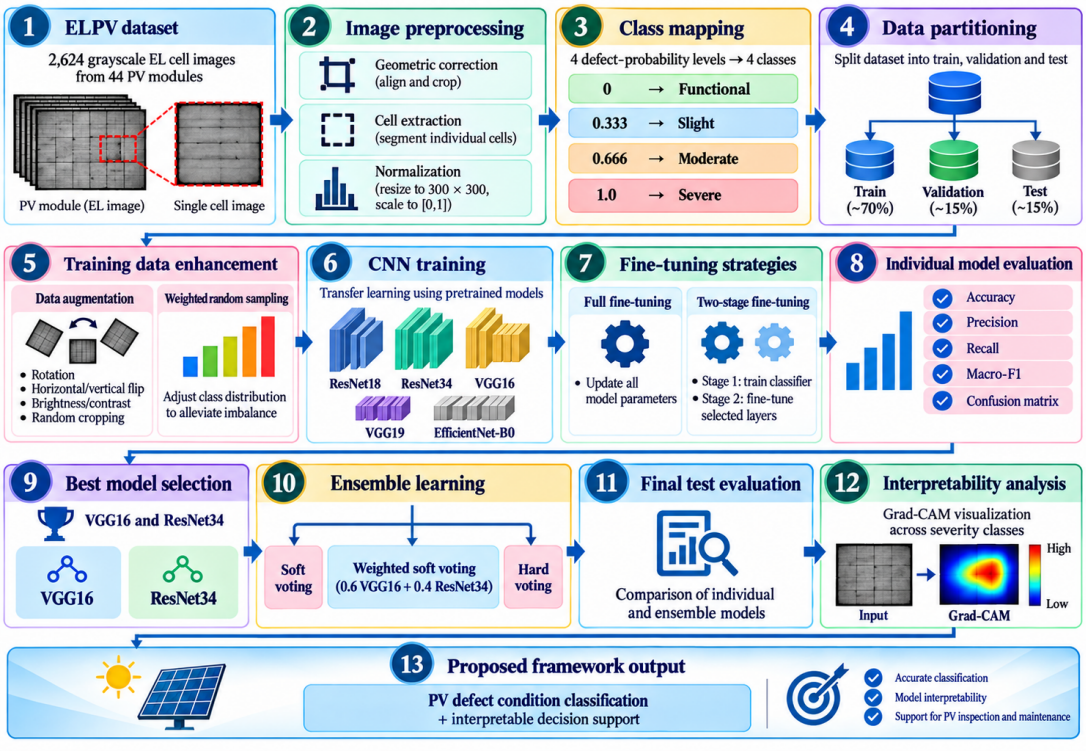


Figure 1. Overall workflow of the proposed methodology
Abbreviations: CNN: Convolutional neural network; EL: Electroluminescence; ELPV: Electroluminescence photovoltaic; Grad-CAM: Gradient-weighted class activation mapping; PV: Photovoltaic; ResNet: Residual network; VGG: Visual Geometry Group.

validation macro-F1 scores and then evaluated on the test set. To enhance classification performance, ensemble models are constructed using soft, weighted, and hard voting strategies. Finally, Grad-CAM visualization is employed for both individual and ensemble models to explore model interpretability and behavior across varying defect severity levels.

3.1. Dataset

The ELPV dataset[18] was considered in this study. The dataset consists of 2,624 grayscale images, each 300 × 300 pixels with 8-bit intensity depth, taken from 44 PV modules. Each image corresponds to a single solar cell extracted from a complete module EL image. The images were preprocessed with geometric processing before extraction to eliminate lens distortions introduced during acquisition and to ensure the samples shared a common spatial representation.

In addition, the data underwent geometric and pixel-intensity normalization. Geometric normalization ensured that all extracted solar cells were aligned and resized to a standard 300 × 300-pixel resolution, preserving the same perspective and scale. Pixel-intensity normalization was also used to match image brightness, with pixel values scaled to the range [0,1] during preprocessing. This normalization operation reduces diversity due to imaging conditions and enhances the stability and convergence of ML models during training.[28]

There are both working and faulty solar cells within the dataset, and the annotations indicate varying degrees of the likelihood of fault. Each picture is annotated with a defect-probability annotation. **Table 2** shows the distribution of the ELPV data by defect probability level and module type. The ELPV dataset contains four different levels of expert annotation for defect probability: 0, 0.333, 0.666, and 1.0. In this study, these original levels of annotation were operationally defined as functional, slight, moderate, and severe condition categories, respectively. The inherent ordering of the original annotations is retained, while simultaneously the task is formulated as a supervised four-class classification problem. Importantly, these categories denote increasing likelihood of a defect, but are not measures of the physical degradation, loss of electrical power, or independently determined engineering severity thresholds. It is worth noting that the considered four-class severity representation is not an arbitrary discretization introduced by the authors. It follows the original annotation structure of the ELPV benchmark, in which each EL image is assigned a defect-probability value. These values were obtained through expert assessment of the EL images and represent increasing degrees of defect occurrence/degradation. Thus, the four classes preserve the original ELPV annotation levels rather than representing newly defined clinical or physical severity thresholds.

However, the classification problem is also highly imbalanced, with most of the samples from the functional category (1,508 images) and the fewest from the moderate category (106 images), making the intermediate categories very difficult to classify. In terms of module type, the dataset includes 1,550 polycrystalline PV module images and 1,074 monocrystalline PV module images, enabling analysis of both PV technologies.

In summary, the considered ELPV dataset contains 2,624 cell-level EL images from only 44 PV modules, with 300 × 300 grayscale resolution. The four severity categories are also highly imbalanced: 1,508 functional, 295 slight, 106 moderate, and 715 severe images. Thus, this dataset is relatively limited in size and exhibits significant class imbalance, making direct training of deep CNN architectures prone to overfitting and biased class predictions. Therefore, the proposed framework employs transfer learning with data augmentation and training-data enhancement, weighted random sampling, and fine-tuning strategies to exploit the available samples while improving generalization across the four defect-severity classes. The augmentation increases the diversity of training inputs and helps mitigate overfitting, while transfer learning reduces the amount of task-specific data required.

3.2. Data partitioning

To construct an equitable and hierarchical experimental design, the dataset is partitioned into training, validation, and testing sets with strict non-overlap constraints, and bias in the training set is mitigated through weighted sampling.[22] To ensure stable model development and objective performance evaluation, the dataset was divided into three subsets: training (70%), validation (15%), and testing (15%), as shown in **Table 3**, yielding 1,836, 394, and 394 samples, respectively. The procedure was performed at the image level with strict non-overlap restrictions between subsets. Cells derived from the same PV module were stored in the same dataset partition to avoid potential data leakage. The model optimization was performed on the training set, and hyperparameter tuning and checkpoint

selection were performed on the validation set. The test set was kept hidden throughout training and used for the final performance assessment.

Table 2. Severity distribution by defect probability and module type

Severity level	Probability	Poly	Mono	Total
Functional	0%	920	588	1,508
Slight	33%	178	117	295
Moderate	67%	50	56	106
Severe	100%	402	313	715
Total	–	1,550	1,074	2,624

Abbreviations: Mono: Monocrystalline silicon; Poly: Polycrystalline silicon.

Table 3. Electroluminescence photovoltaic dataset partitioning

Subset	Number of samples
Training	1,836
Validation	394
Testing	394
Total	2,624

3.3. Class imbalance handling

The training subset analysis showed substantial variation in sample counts across the four severity categories. This type of distributional skew may bias the optimization process, since deep neural networks tend to favor majority classes during training. Thus, a WRS was used in training to resolve this effect.[28] Rather than modifying the stored dataset or duplicating samples offline, this approach assigned a sampling weight to each training instance, proportional to the inverse of its class frequency. As a result, minority classes are sampled more often during mini-batch construction, leading to equal exposure to all classes during training. This strategy allows the model to learn representative features from underrepresented classes without modifying the original dataset distribution. Notably, this balancing mechanism operates at the data-loading level; i.e., it was used only in the training subset, not in the validation or test sets, to avoid bias in evaluation.

Table 4 shows the initial severity distribution in the training subset before implementing any class-balancing strategy. The functional class contained the most samples (1,055 samples), whereas the moderate class contained the fewest (74). The slight and severe categories have 207 and 500 samples, respectively. This disparity in severity levels encourages the use of a sampling-based balancing scheme during training to balance class exposure. **Table 5** summarizes the PV type of the module in the training subset. The largest parts are monocrystalline modules (1,098 samples) and polycrystalline modules (738 samples). This distribution is useful for showing how the two module technologies are represented in the model's training data.

After using the WRS[28], the effective sampled class exposure became evenly distributed across severity categories. The sampling structure was used to ensure an equal number of instances in each class during training, as demonstrated in **Table 6**. Thus, the training distribution was balanced across all four severity groups during sampling. The sampling approach adopted, which includes imposing equal representation across classes, reduces the effect of optimization bias, resulting in the model's greater ability to learn severity patterns in minority classes. The sampling strategy ensures balanced class representation during training without increasing the actual dataset size.

Table 4. Original severity distribution in training subset by severity level

Severity level	Label ID	Number of samples
Functional	0	1,055
Slight	1	207
Moderate	2	74
Severe	3	500
Total samples	–	1,836

Table 5. Original severity distribution in training subset by module type

Module type	Number of samples
Monocrystalline	738
Polycrystalline	1,098
Total samples	1,836

3.4. Image transformation pipeline

To achieve standardized input representations for model training, a structured image transformation pipeline is applied to each input sample before it is fed into the neural network. It is a sequence of deterministic and random preprocessing operations on each image, which is then fed into the neural network. Since the original data consists of 300 × 300 grayscale images, all samples were resized to match the input dimensions of the selected CNN architectures. Because the tested models (ResNet, VGG, and EfficientNet) were trained on ImageNet and assume three-channel images, grayscale images were also converted to three-channel formats to ensure compatibility with pretrained weights.[17]

Table 6. Effective sampled severity exposure per epoch after applying weighted random sampler

Severity level	Label ID	Number of samples
Functional	0	1,055
Slight	1	1,055
Moderate	2	1,055
Severe	3	1,055
Total samples	–	4,220

3.4.1. Training transformations

Stochastic transformations were used during training to introduce controlled variability in the training data, enabling the model to generalize better to unseen samples, thus improving model robustness and reducing overfitting.[29]

The training pipeline involved:

- random resized cropping to 224 × 224 pixels
- random horizontal flipping
- random rotation within a limited angular range
- grayscale conversion to three channels
- conversion to tensor format
- intensity normalization

3.4.2. Validation and testing transformations

For validation and testing, only deterministic preprocessing operations were applied to ensure a stable, reproducible evaluation.[11] These included:

- resizing to a fixed intermediate resolution
- center cropping to 224 × 224 pixels
- grayscale conversion to three channels
- conversion to tensor format
- intensity normalization

Validation and testing were performed without stochastic augmentations to maintain evaluation consistency. The use of a systematic transformation pipeline makes the input dimensionality of an experiment uniform and enables data-driven regularization of the training data without compromising the integrity of evaluation.

3.5. Data augmentation strategy

Data augmentation was used to train the model, improving generalization and reducing overfitting. In contrast to expanding static datasets, the augmentation process was applied dynamically during training, i.e., stochastic transformations were performed on-the-fly as mini-batches were loaded. As a result, the model is subjected to slightly varied versions of the same image across different epochs without physically increasing the size of the stored data.[29,30] These changes are a realistic simulation of changes in orientation, positioning, and spatial representation, and maintain the structural properties of cell defects in PV cells.

In **Figure 2**, an original image is shown alongside several augmented versions. These changes introduce regulated variability into the training process while maintaining the structural properties of PV cell defects. The augmentation operations, incorporated into the training transformation pipeline and comprising controlled geometric perturbations, include random resized cropping, horizontal flipping, and small-angle rotations.

Figure 3 shows augmented training samples, each assigned to a defect-severity category. As the examples show, the idea of augmentation is applied uniformly across all classes without sacrificing defect patterns.

Notably, data augmentation was used only on the training set. Only deterministic preprocessing was applied to the validation and test subsets to ensure consistent, repeatable, and objective performance measurement. With on-the-fly augmentation, the training procedure can achieve greater input variability, be more resilient to small spatial distortions, and exhibit greater generalization, without modifying the ingrained evaluation protocol.

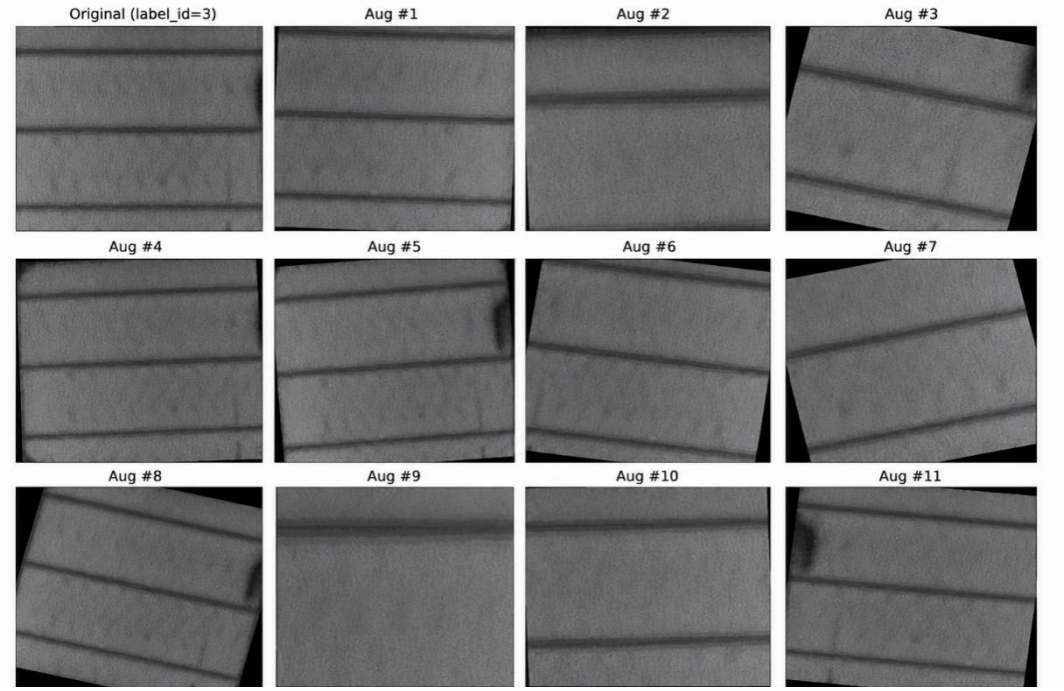


Figure 2. Examples of data augmentation operations applied to an electroluminescence solar cell image

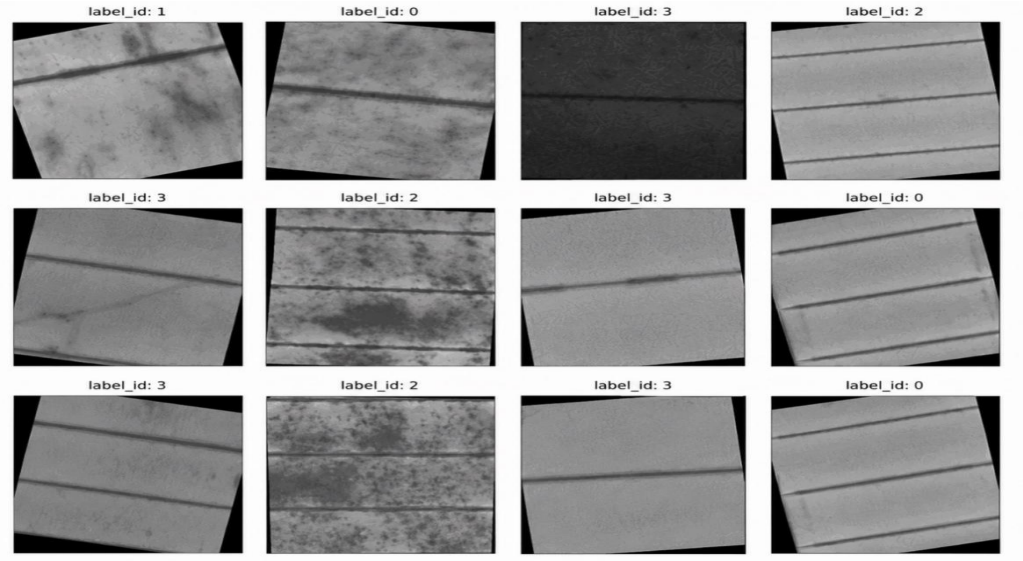


Figure 3. Augmented electroluminescence solar cell samples from different severity classes

3.6. Model architectures

The paper compares five popular CNN models for classifying the severity of PV defects using the ELPV dataset. The use of the following backbones, ResNet18, ResNet34, VGG16, VGG19, and EfficientNet-B0, is motivated by their diversity in architectural design, which provides a strong basis for examining interactions among feature-extraction capacity, model depth, and transfer-learning behavior.[17] Each model was pretrained on ImageNet weights and adapted to the four-class severity classification task.

ResNet18, as described in Ref.[31], is a residual CNN architecture that uses skip connections to enhance gradient flow and stabilize optimization in deep networks. Its relatively lightweight structure enables it to be computationally efficient while retaining a high level of representational power.

ResNet18 was selected in this research as a small baseline to assess the performance of residual learning in EL-based PV severity classification. On the other hand, ResNet34 is a deeper residual architecture.[31] Compared with ResNet18, it offers greater representational capacity via additional residual blocks while retaining the optimization benefits of skip connections. It is used to investigate whether increased network depth improves the discrimination of PV defect severity levels, particularly for visually ambiguous intermediate classes.

VGG16, proposed in Ref.[17], consists of a series of small 3x3 convolutional filters and then fully connected layers. Despite having more parameters than residual architectures, its simple, efficient hierarchical structure has made it a powerful baseline for transfer learning. VGG16 was considered in this study to evaluate the effectiveness of a classical deep CNN architecture for EL image classification. On the other hand, VGG19 is a deeper variant of the VGG family introduced in Ref.[17] It uses the same sequential convolutional architecture as VGG16, with more convolutional layers, allowing more comprehensive feature extraction. It is used to examine whether increased depth in classical CNN architectures improves performance in multi-class severity classification.

Finally, EfficientNet-B0, introduced in Ref.[32], is based on a compound scaling strategy that jointly optimizes network depth, width, and input resolution. This design aims to achieve a strong balance between predictive accuracy and computational efficiency. EfficientNet-B0 is included to assess whether modern, scaled architectures can deliver robust performance on EL-based severity classification tasks.

Because the original EL images are single-channel grayscale, they were converted to three-channel images so that all inputs to ImageNet-pretrained models could be used without distorting the underlying luminance data. In all considered architectures, the original final classification head was replaced with a task-oriented output head that indicates the four severity categories: functional, slight, moderate, and severe.

3.7. Training setup and implementation information

All models were trained within a unified experimental framework to ensure methodological consistency and enable fair comparison between the evaluated architectures. Hence, variations in performance are due to architectural considerations and fine-tuning strategies rather than differences in the training environment. The training process was implemented using PyTorch[33], with GPU[19] acceleration enabled when available to improve computational efficiency. Before optimization began, all models were transferred to either a Compute Unified Device Architecture (CUDA)-enabled GPU or the CPU, depending on hardware availability. To accelerate data transfer from CPU memory to GPU memory during training, the pin_memory option was enabled in the PyTorch data loader when CUDA was available.

The pretrained backbones were initially trained to classify images on ImageNet; thus, all EL images were downsampled to 224 × 224 pixels to match the networks' expected input size. All experiments used a batch size of 32 to balance computational efficiency and gradient stability during training. The training was conducted for 15 epochs with a learning rate of 1×10^{-4} and a weight decay coefficient of 1×10^{-4}, using the Adam optimizer.[34,35] The cross-entropy loss was developed as an optimization objective and is used for multi-class classification problems [20]

To facilitate transparency and reproducibility of the proposed methodology, an interactive implementation of the developed framework has been made publicly available through Hugging Face. The Space allows users to upload EL solar-cell images and obtain predictions from VGG16, ResNet34, Soft Voting, and Weighted Soft Voting, together with Grad-CAM visualizations. The implementation is available at Ref.[36]

3.8. Fine-tuning strategies

The experiment aimed to explore the impact of transfer-learning adaptation on PV severity classification, using two fine-tuning strategies across all analyzed CNN backbones: full fine-tuning and two-stage fine-tuning. The preprocessing pipeline, training setup, and evaluation metrics were applied to both strategies to enable a fair comparison.[31] In the full-fine-tuning setup, the entire pretrained backbone was free to change throughout training. ImageNet-pretrained weights were used to initialize the networks, and all layers of the model (convolutional layers and the classification head) were optimized together on the training data. The strategy enables the model to fully realign its acquired representations to the EL space and further refine feature learning for PV defect patterns. Nevertheless,

since all parameters are updated during training, this method can lead to overfitting on small datasets. Conversely, two-stage fine-tuning offers a more controlled adaptation by initially retaining generic features learned from ImageNet and subsequently fine-tuning the entire network, which often enhances convergence and improves generalization. However, since in full fine-tuning, all layers are updated from the beginning, allowing the network to fully adapt the ImageNet features to the EL image domain, full fine-tuning can outperform the two-stage fine-tuning strategy regarding accuracy.

3.9. Evaluation metrics

Model performance was evaluated using several metrics suitable for multi-class classification. Accuracy was used to measure the overall proportion of correctly classified samples.[20] Per-class F1-score, recall, and precision were also examined to provide a more detailed evaluation of model performance across individual severity categories. Since the dataset exhibits class imbalance, macro-F1 was adopted as the primary evaluation metric. Macro-F1 computes the F1-score independently for each class and averages the results, giving equal importance to all classes regardless of their distribution.[28] Additionally, macro-recall was analyzed to assess the model's sensitivity across different defect severity levels.[20] This is also called balanced accuracy since it accounts for class imbalance by averaging recall across all classes, providing a more reliable estimate of performance compared to standard accuracy. Furthermore, confusion matrices were analyzed to visualize misclassification patterns, particularly to understand errors between adjacent severity levels such as slight, moderate, and severe.

For multi-class classification:

$$Accuracy = \frac{\frac{1}{c}\sum_{i=1}^{c} TP(i)}{Total\ samples\ number} \tag{1}$$

$$Balanced\ Accuracy\left(Macro-Recall\right) = \frac{1}{c}\sum_{i=1}^{c} Recall\left(i\right) \tag{2}$$

$$Macro-F1 = \frac{1}{c}\sum_{i=1}^{c} F1\left(i\right) \tag{3}$$

$$F1\left(i\right) = \frac{2\times Precision\left(i\right)\times Recall\left(i\right)}{Precision\left(i\right)+Recall\left(i\right)} \tag{4}$$

$$Precision\left(i\right) = \frac{TP\left(i\right)}{TP\left(i\right)+FP\left(i\right)} \tag{5}$$

$$Recall\left(i\right) = \frac{TP\left(i\right)}{TP\left(i\right)+FN\left(i\right)} \tag{6}$$

where

c: Class number

$TP(i)$: True positives per class

$FP(i)$: False Positives per class

$FN(i)$: False Negatives per class

All models were tested in 10 separate runs, and the results are presented as the mean and standard deviation to demonstrate robustness and reproducibility.

4. Results

4.1. The results of evaluation metrics

Table 7 indicates that the fully fine-tuned versions of ResNet18, ResNet34, and VGG16 generally perform better than other versions. The maximum validation macro-F1 score for ResNet18 (full fine-tuning) reached 0.5850 ± 0.0150, indicating good overall classification performance and balanced results across classes. Additionally, ResNet18 had the best macro-precision (0.5900 ± 0.0140), indicating more credible, confident positive estimates. Conversely, ResNet34 (full fine-tuning) achieved a marginally lower macro-F1 score (0.5485 ± 0.0175), indicating moderate performance with less differentiation between class predictions. In the meantime, VGG16 (full fine-tuning) exhibited stable, consistent behavior, with a macro-F1 score of 0.5671 ± 0.0232 and a balanced between precision and recall, indicating consistency across runs.

These findings suggest that ResNet18 has the best overall performance in macro-F1 and precision, whereas VGG16 has a more balanced trade-off between precision and recall. ResNet34, in turn, demonstrates relatively worse performance yet is a competitive baseline. Altogether, the three architectures can be regarded as efficient and stable methods of PV defect severity classification, with ResNet18 being the most powerful in this research.

The per-class validation performance comparison in **Table 8** shows that each of the three architectures has its own strengths and

Table 7. Validation (Vad) performance comparison between the considered convolutional neural networks

Model	Fine-tuning (FT)	Val accuracy	Val precision (macro)	Sensitivity - recall (macro)	Val recall (macro)	Val F1 (macro)
ResNet18	Full FT	0.7520 ± 0.0120	0.5900 ± 0.0140	0.5900 ± 0.0180	0.5900 ± 0.0180	0.5850 ± 0.0150
ResNet18	Two-stage FT	0.7107 ± 0.0130	0.5595 ± 0.0150	0.5993 ± 0.0200	0.5993 ± 0.0200	0.5726 ± 0.0160
ResNet34	Full FT	0.7175 ± 0.0252	0.5563 ± 0.0257	0.5419 ± 0.0262	0.5419 ± 0.0262	0.5485 ± 0.0175
ResNet34	Two-stage FT	0.6980 ± 0.0140	0.5390 ± 0.0160	0.5597 ± 0.0190	0.5597 ± 0.0190	0.5471 ± 0.0150
EfficientNet-B0	Full FT	0.7259 ± 0.0120	0.3878 ± 0.0200	0.4322 ± 0.0220	0.4322 ± 0.0220	0.4005 ± 0.0201
EfficientNet-B0	Two-stage FT	0.6218 ± 0.0160	0.5103 ± 0.0180	0.5463 ± 0.0200	0.5463 ± 0.0200	0.4980 ± 0.0170
VGG16	Full FT	0.7437 ± 0.0126	0.5573 ± 0.0113	0.5601 ± 0.0168	0.5601 ± 0.0168	0.5671 ± 0.0232
VGG16	Two-stage FT	0.6726 ± 0.0140	0.5352 ± 0.0160	0.6001 ± 0.0210	0.6001 ± 0.0210	0.5509 ± 0.0170
VGG19	Full FT	0.6371 ± 0.0150	0.5263 ± 0.0170	0.5804 ± 0.0200	0.5804 ± 0.0200	0.5262 ± 0.0160
VGG19	Two-stage FT	0.6904 ± 0.0130	0.5565 ± 0.0160	0.6043 ± 0.0210	0.6043 ± 0.0210	0.5546 ± 0.0170

Table 8. Per-class validation performance for selected convolutional neural network models (full fine-tuning [FT])

Model	Class	Precision	Recall (sensitivity)	F1-score	Support
ResNet18 (Full FT)	Functional (0.0)	0.82 ± 0.02	0.88 ± 0.03	0.85 ± 0.02	226
	Slight (0.333)	0.44 ± 0.03	0.60 ± 0.04	0.51 ± 0.03	44
	Moderate (0.666)	0.30 ± 0.03	0.38 ± 0.04	0.33 ± 0.04	16
	Severe (1.0)	0.80 ± 0.02	0.74 ± 0.01	0.77 ± 0.02	108
ResNet34 (Full FT)	Functional (0.0)	0.83 ± 0.02	0.68 ± 0.03	0.75 ± 0.02	226
	Slight (0.333)	0.32 ± 0.01	0.56 ± 0.03	0.41 ± 0.03	44
	Moderate (0.666)	0.16 ± 0.04	0.31 ± 0.02	0.21 ± 0.04	16
	Severe (1.0)	0.67 ± 0.03	0.67 ± 0.02	0.67 ± 0.03	108
VGG16 (Full FT)	Functional (0.0)	0.82 ± 0.02	0.76 ± 0.02	0.79 ± 0.02	226
	Slight (0.333)	0.35 ± 0.03	0.61 ± 0.03	0.45 ± 0.03	44
	Moderate (0.666)	0.42 ± 0.04	0.44 ± 0.04	0.43 ± 0.04	16
	Severe (1.0)	0.83 ± 0.02	0.78 ± 0.02	0.80 ± 0.02	108

weaknesses. ResNet18 achieves the highest recall for the functional class (0.88 ± 0.03), indicating it is highly effective at identifying non-defective cells, which is one reason it is very accurate overall. Meanwhile, ResNet34 has a comparatively lower recall for the functional class (0.68 ± 0.03), though the other classes are relatively moderate and consistent, especially the slight and severe ones. In the meantime, VGG16 performs better on severe defects, with the highest precision (0.83 ± 0.02) and high recall (0.78 ± 0.02) in the severe category, reflecting its ability to identify obvious defect patterns. Nevertheless, the three models share a weakness in the moderate class (0.666), as the precision and F1-scores are far lower than in other classes. For example, in all models, the F1-score of this class is constrained (0.33 ± 0.04 in ResNet18, 0.21 ± 0.04 in ResNet34, and 0.43 ± 0.04 in VGG16). It implies that moderate defects are harder to differentiate in principle, probably because of their less pronounced visual characteristics and comparatively few training samples. Consequently, this weakness appears to be data-driven rather than model-driven, suggesting that improved data representation methods, such as class balancing or data augmentation, should be adopted.

The results of the tests in **Table 9** show differences in performance among the evaluated models, attesting to the reliability and stability of the tested models. VGG16 (full fine-tuning) outperformed the other architectures, achieving the highest test accuracy (0.7454 ± 0.0126) and macro-F1 score (0.5555 ± 0.0099), indicating strong generalization to unseen data. This high performance is due to its architectural design, which relies on deep sequential convolutional layers that effectively capture fine-grained local features. These features render VGG16 especially appropriate for EL images, in which defect patterns are often less pronounced and require more detailed extraction of spatial features.

ResNet18 (full fine-tuning) also showed competitive performance, with slightly lower accuracy and macro-F1, indicating its capability to provide consistent and dependable predictions across multiple runs. Among the fully fine-tuned models, VGG16 achieved the highest test macro-recall (0.5601 ± 0.0168). Across all evaluated fine-

Table 9. Test performance comparison (mean ± standard deviation over 10 runs)

Model	Fine-tuning (FT)	Best epoch	Test accuracy	Test precision (macro)	Test recall (macro)	Test F1 (macro)
ResNet18	Full FT	8	0.7358 ± 0.0226	0.5607 ± 0.0298	0.5529 ± 0.0296	0.5522 ± 0.0264
ResNet18	Two-stage FT	15	0.7183 ± 0.0200	0.5530 ± 0.0180	0.5599 ± 0.0200	0.5489 ± 0.0190
ResNet34	Full FT	8	0.7299 ± 0.0262	0.5563 ± 0.0271	0.5419 ± 0.0276	0.5421 ± 0.0224
ResNet34	Two-stage FT	11	0.7208 ± 0.0180	0.5290 ± 0.0170	0.5327 ± 0.0180	0.5302 ± 0.0160
EfficientNet-B0	Full FT	10	0.7280 ± 0.0180	0.5400 ± 0.0200	0.5160 ± 0.0210	0.5260 ± 0.0180
EfficientNet-B0	Two-stage FT	10	0.6142 ± 0.0200	0.4831 ± 0.0180	0.4894 ± 0.0190	0.4725 ± 0.0170
VGG16	Full FT	13	0.7454 ± 0.0126	0.5573 ± 0.0113	0.5601 ± 0.0168	0.5555 ± 0.0099
VGG16	Two-stage FT	14	0.6802 ± 0.0150	0.5239 ± 0.0140	0.5686 ± 0.0180	0.5331 ± 0.0150
VGG19	Full FT	15	0.6777 ± 0.0170	0.4967 ± 0.0180	0.5297 ± 0.0200	0.5007 ± 0.0160
VGG19	Two-stage FT	15	0.6751 ± 0.0160	0.5301 ± 0.0170	0.5349 ± 0.0190	0.5237 ± 0.0160

tuning configurations, VGG16 with two-stage fine-tuning achieved the highest test macro-recall (0.5686 ± 0.0180).

In general, the results confirm that ResNet18, ResNet34, and VGG16 are all effective architectures for classifying PV defect severity, and each model has its own advantages. While VGG16 excels in generalization and overall performance, ResNet18 offers high precision and stability, and ResNet34 performs better at capturing defective samples.

4.2. Confusion matrix analysis of convolutional neural network models across defect severity classes

As shown in **Figure 4**, both ResNet18 and ResNet34 perform well on the functional and severe classes, as indicated by high diagonal values. However, ResNet18 shows some confusion in the moderate class, which can be mistaken for surrounding classes such as functional and slight. This reveals the difficulty of separating intermediate classes, especially in unbalanced class scenarios. On the other hand, for ResNet34, there is some misclassification between the slight and moderate classes, especially the latter, which remains challenging to predict. This is due to the data's imbalance and the proximity of the severity levels.

As for VGG16, **Figure 5** shows that the model performs well in the functional and severe classes, with a large proportion of correct predictions along the diagonal. For the moderate class, VGG16 achieves better classification performance than previous models, though it still misclassifies neighboring classes. These results show improved feature discrimination, but class imbalance effects remain.

4.3. Comparative analysis between ResNet18, ResNet34, and VGG16 (full fine-tuning)

The confusion matrices show how the three models perform in classifying the four severity classes. Each model achieves high scores for correctly identifying the functional and severe classes, as shown by the high values on the diagonal. VGG16 is slightly more effective at identifying serious defects, and it incurs fewer misclassifications than other models. Conversely, there is significant confusion in the moderate category across all models, with misclassifications often directed towards the functional category. Such a tendency can be attributed to the intrinsic complexity of separating moderate flaws, given their non-obvious visual features and their scarcity in the sample. In general, the

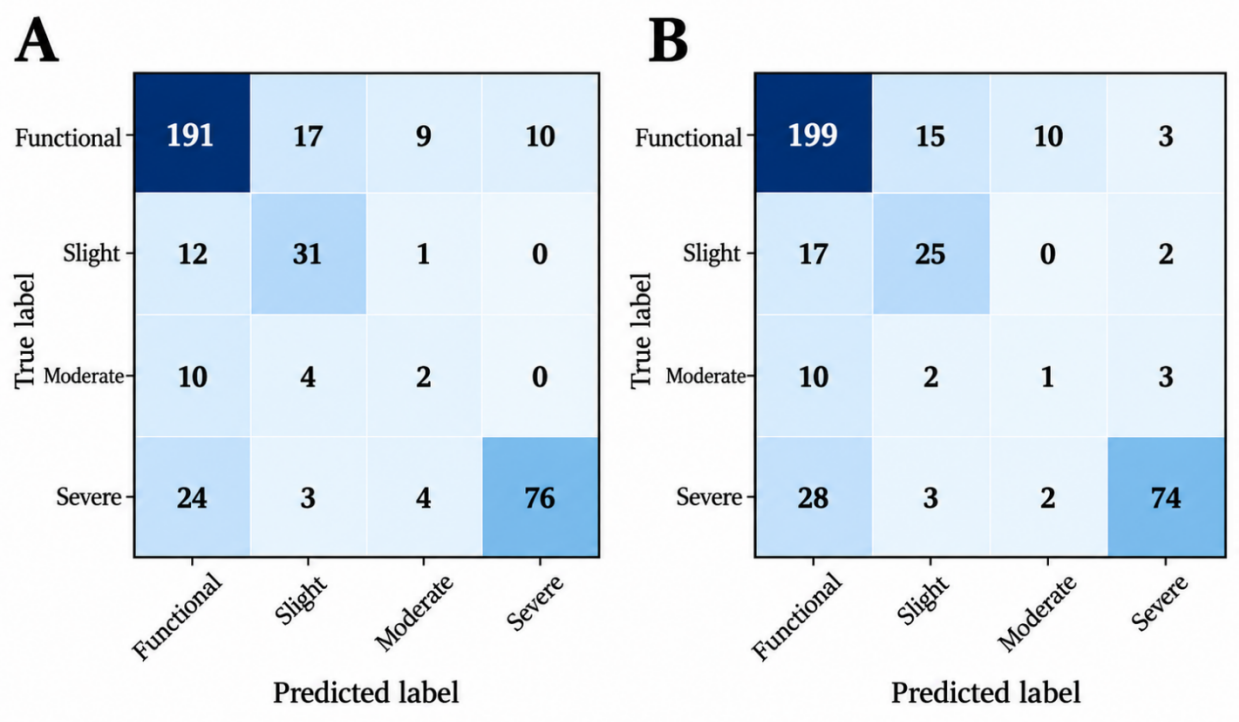


Figure 4. Confusion matrix with full fine-tuning with (A) ResNet18, (B) ResNet34 on the test set

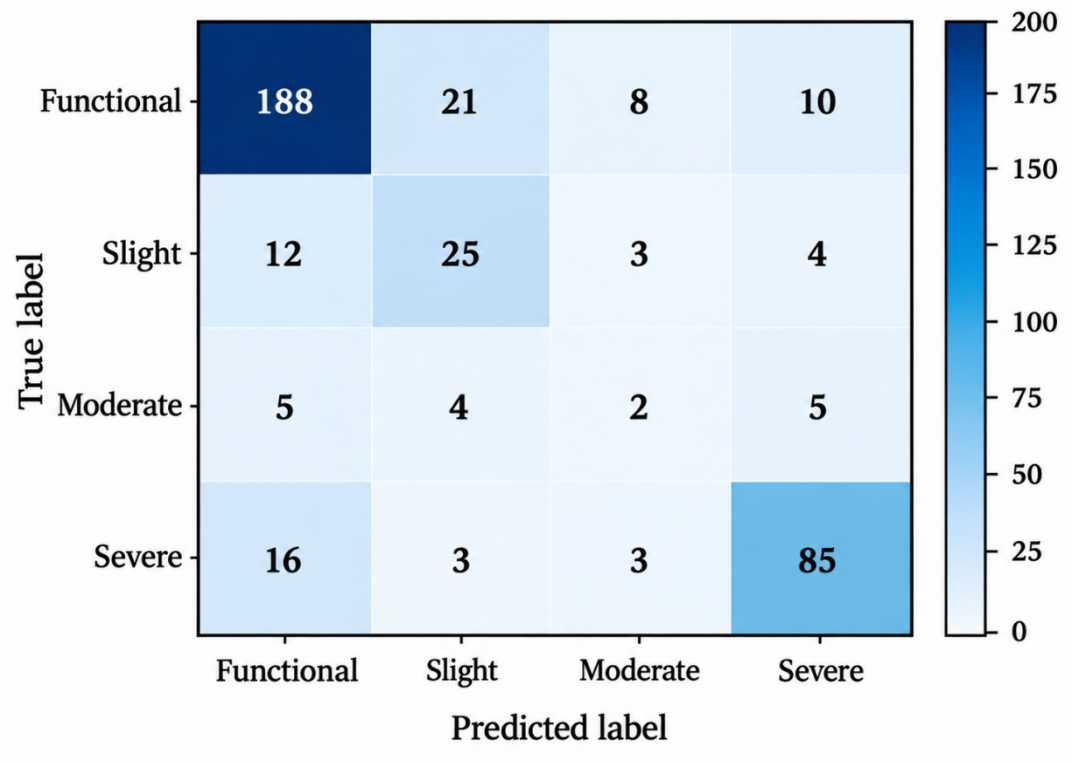


Figure 5. Confusion matrix of the VGG16 model (full fine-tuning) on the test set

findings validate the assumption that the three architectures are effective for well-defined classes. Still, there remains a problem with correctly classifying fewer classes at intermediate severity levels, especially the moderate class. To ensure the soundness, stability, and reliability of findings, the comparative results presented in **Tables 7–9** show detailed results for CNN architectures and fine-tuning parameters for classifying defect severity in the ELPV dataset.

VGG16 (full fine-tuning) was the most successful architecture in terms of overall performance on the test set, with the highest test accuracy (0.7454 ± 0.0126) and macro-F1 (0.5555 ± 0.0099). This means better generalization to unknown information. The high performance of VGG16 can be attributed to its sequential convolutional architecture, which is particularly effective at capturing fine-grained local features. These properties are critical for EL images, where defect patterns can be very delicate and require complex spatial feature detection.

The ResNet18 (full fine-tuning) was competitive, with the highest validation macro-F1 score of 0.5850 ± 0.0150. This outcome shows that ResNet18 achieved a good trade-off between precision and recall on the validation set. VGG16 had the best test macro-recall score in all the fine-tuning models, at 0.5601 ± 0.0168. ResNet34 had a lower test macro-recall of 0.5419 ± 0.0276 and different patterns of prediction and attention that enabled it to play a complementary role within the ensemble.

Table 8 in the per-class analysis, along with the confusion matrices, also reveals a clear pattern across all models. Although the functional and severe classes can be classified with high accuracy, all architectures have difficulty with the moderate class, with both recall and F1-score significantly lower. This challenge is evident in the confusion matrices, where moderate samples are often misclassified as either functional or slight. This data limitation is mostly data-based and not model-specific. The sample size of the moderate class is extremely small (16 instances in both validation and test sets), which implies that DL models cannot acquire representative features. Moreover, moderate defects often exhibit imperceptible visual features that straddle the boundaries between adjacent categories, making them harder to classify. The other factor is the domain gap between ImageNet and EL images. The fact that the models are pretrained on natural images means that their

learned feature representations may not fully align with the structural patterns evident in PV EL data. Although fine-tuning helps alleviate this problem, the dataset is relatively small, limiting the models' capacity to adapt.

Moreover, the results show that full fine-tuning is generally better than staged fine-tuning across most architectures. These findings indicate that multi-stage fine-tuning strategies do not always yield better performance, and in several instances, they result in marginally lower macro-F1 scores than full fine-tuning. This implies that freezing the initial layers during multi-stage training seems to restrict the model from learning low-level features specific to EL imagery, while in the full fine-tuning strategy, all network layers are enabled to adapt to the EL dataset, allowing the models to learn more domain-specific features associated with defect patterns and structural variations in solar cells.

Overall, the findings indicate that modern CNN models such as VGG16, ResNet18, and ResNet34 have a strong ability to classify PV defect severity. Still, their performance is average when evaluated using the class-balanced metric of macro-F1. This is largely due to class imbalance and data scarcity, which remain the main bottlenecks. To overcome those challenges, architecture-class-balancing methods, better data augmentation, or specialized methods are crucial for achieving improved performance in real-world PV inspection systems.

4.4. Ensemble-based performance enhancement

To enhance the classification performance of PV defect severity on the ELPV dataset, ensemble DL is employed to integrate information captured from several models, thereby complementing one another and improving classification performance, particularly in imbalanced settings.

4.4.1. Motivation for ensemble learning

Although fine-tuning improved the performance of individual DL models, the overall results remained limited when evaluated using class-balanced metrics such as macro-F1. While architectures such as ResNet18 and VGG16 achieved strong performance on dominant classes, their ability to generalize across all severity levels was restricted, particularly for the underrepresented moderate class (0.666). This is limited mostly by the class imbalance that results in biased predictions toward majority classes and reduced sensitivity to minority classes. Moreover, the various models had complementary strengths, and each architecture captured distinct features.

Ensemble learning is one proposed solution to these challenges, aimed at improving overall generalization and minimizing model-specific errors. Ensemble methods integrate information from several models, thereby complementing one another and improving classification performance, particularly in imbalanced settings. In this analysis, ensemble learning will be used to produce more balanced and reliable predictions across all severity classes.

For the weighted soft voting ensemble, a higher weight was assigned to VGG16 because it demonstrated stronger individual classification performance than ResNet34 during model evaluation. Accordingly, the ensemble weights were predefined as 0.6 for VGG16 and 0.4 for ResNet34. This weighting strategy was intended to give greater influence to the better-performing model while preserving the complementary contribution of ResNet34. The weights were fixed before the final test-set evaluation and were not optimized using the test results.

4.4.2. Results of ensemble methods

Three decision-level ensemble strategies are used for quantitative assessment of the performance of the ensemble: soft voting, weighted soft voting, and hard voting. The ensembles were made up by the fusion of the results of two models: VGG16 and ResNet34. The weights for both models were set to be equal when normal soft voting is used, while weighted soft voting gave a weight of 0.6 to VGG16 and 0.4 to ResNet34. VGG16 was assigned a higher weight due to its better performance per se, and ResNet34 was retained due to its complementary prediction and feature-attention behavior. Hard voting fused the last class predictions of both models using discrete decision aggregation. The classification reports of the three ensemble methods are shown in **Tables 10–12**, and these show, for each class, precision, recall, F1-score, and support, as well as overall accuracy and the macro-averaged precision, recall, and F1-score. The ensemble strategies showed a significant improvement in performance over the individual CNN models, with the macro-F1 score rising from around 0.55 to over 0.83.

The weighted soft voting ensemble achieved the best overall performance, with an accuracy of 0.8985 and a macro-F1 score of 0.8458, by assigning a weight of 0.6 to VGG16 and 0.4 to ResNet34. This result indicates that giving greater influence

Table 10. Soft voting ensemble (VGG16 + ResNet34): Test classification report

Class	Precision	Recall (sensitivity)	F1-score	Support
Functional (0.0)	0.9565	0.8761	0.9145	227
Slight (0.333)	0.7049	0.9773	0.8190	44
Moderate (0.666)	0.7059	0.7500	0.7273	16
Severe (1.0)	0.8899	0.8981	0.8940	107
Accuracy	-	-	0.8883	394
Macro average	0.8143	0.8754	0.8387	394
Weighted average	0.9000	0.8883	0.8906	394

Table 11. Weighted soft voting ensemble (0.6 VGG16 + 0.4 ResNet34): Test classification report

Class	Precision	Recall (sensitivity)	F1-score	Support
Functional (0.0)	0.9619	0.8938	0.9266	227
Slight (0.333)	0.7167	0.9773	0.8269	44
Moderate (0.666)	0.7059	0.7500	0.7273	16
Severe (1.0)	0.9065	0.8981	0.9023	107
Accuracy	-	-	0.8985	394
Macro average	0.8227	0.8798	0.8458	394
Weighted average	0.9089	0.8985	0.9007	394

Table 12. Hard voting ensemble (VGG16 + ResNet34): Test classification report

Class	Precision	Recall (sensitivity)	F1-score	Support
Functional (0.0)	0.9614	0.8805	0.9192	227
Slight (0.333)	0.7049	0.9773	0.8190	44
Moderate (0.666)	0.6667	0.7500	0.7059	16
Severe (1.0)	0.8981	0.8981	0.8981	107
Accuracy	-	-	0.8909	394
Macro average	0.8078	0.8765	0.8356	394
Weighted average	0.9034	0.8909	0.8936	394

to VGG16, while preserving the complementary contribution of ResNet34, produced more effective and balanced predictions. Standard soft voting also achieved strong performance, with an accuracy of 0.8883 and a macro-F1 score of 0.8387. Hard voting achieved an accuracy of 0.8909 and a macro-F1 score of 0.8356. Although hard voting slightly outperformed standard soft voting in terms of overall accuracy, both probability-based approaches, standard soft voting and weighted soft voting, achieved higher macro-F1 scores than hard voting. Among all evaluated ensemble

strategies, weighted soft voting provided the best balance between accuracy and class-balanced performance.

The moderate category remained the most challenging class because of its limited representation in the dataset. For the weighted soft voting ensemble, moderate-class recall increased from 0.44 for VGG16 to 0.75, while the corresponding F1-score increased from 0.43 to 0.7273. However, these results should be interpreted cautiously because the test set contained only 16 moderate samples. With such a small support size, class-specific performance estimates are more sensitive to a small number of prediction errors and are therefore subject to greater statistical uncertainty than those of the better-represented classes. Accordingly, the observed improvement in moderate-class performance should be considered encouraging evidence within the current benchmark rather than conclusive evidence of generalizable performance for this severity category. Validation on larger datasets containing substantially more intermediate-severity samples is required to establish the stability and reproducibility of this result.

In general, these findings confirm that ensemble learning is highly effective at improving classification robustness and generalization. The ensemble approach addresses individual model weaknesses, provides more balanced performance than any single model by integrating complementary models, and is more successful across all severity classes.

4.4.3. Comparative performance of ensemble techniques

The comparison of the three ensemble methods reveals that they have significantly enhanced classification performance over individual models. Nevertheless, one can see certain distinctions between the ensemble strategies. The best overall performance was achieved by the Weighted Soft-Voting ensemble, which achieved the highest accuracy and macro-F1 score. This implies that models with varying weights are assigned different weights, resulting in more effective and balanced predictions. The performance of soft voting was also high, slightly lower than that of the weighted approach, demonstrating the benefit of probability-based aggregation. Hard voting, on the other hand, exhibited comparatively lower performance, implying that the use of discrete class decisions is not as effective as the use of confidence scores. Overall, the findings suggest that probability-based ensemble algorithms, particularly Weighted Soft Voting, are the most reliable and robust solution for the considered ELPV dataset.

4.4.4. Quantitative analysis of ensemble performance improvements

Figure 6 shows that ensemble methods significantly improve performance compared to individual models. Although VGG16 and ResNet34 have average performance (macro-F1 $\approx$ 0.55), all ensemble methods dramatically improve performance, with macro-F1 exceeding 0.83. The weighted method yields optimal outcomes (accuracy = 0.8985, macro-F1 = 0.8458), indicating that appropriate weighting can improve accuracy and class distribution. Overall, the results confirm that ensemble learning provides more robust and reliable predictions, particularly in imbalanced multi-class scenarios. On the other hand, **Figure 7** shows that ensemble methods consistently improve recall across all classes compared to individual models. The most notable improvement is observed in the moderate category, where recall increases from 0.44 (VGG16) and 0.31 (ResNet34) to 0.75 across all ensemble strategies. This highlights the effectiveness of ensemble learning in capturing challenging and underrepresented patterns. In addition, ensemble models achieve very high recall in the slight category (0.9773) and maintain strong performance in the severe category (approximately 0.8981), indicating that these improvements are achieved without degrading performance on well-represented classes. Improvements are also observed in the functional class, although to a lesser extent.

4.4.5. Statistical assessment of ensemble improvement

To further evaluate the improvement achieved by the proposed ensemble approach, the weighted soft voting ensemble was compared with the best-performing individual CNN model, VGG16, using the held-out test set. VGG16 achieved a test accuracy of 0.7454 and a macro-F1 score of 0.5555, whereas the weighted soft voting ensemble achieved an accuracy of 0.8985 and a macro-F1 score of 0.8458. This corresponds to absolute improvements of 0.1531 in accuracy and 0.2903 in macro-F1. In addition, the moderate-class recall increased from 0.44 for VGG16 to 0.75 for the weighted soft voting ensemble. To quantify the statistical uncertainty associated with these improvements, a paired bootstrap procedure with 10,000 resamples was used to estimate

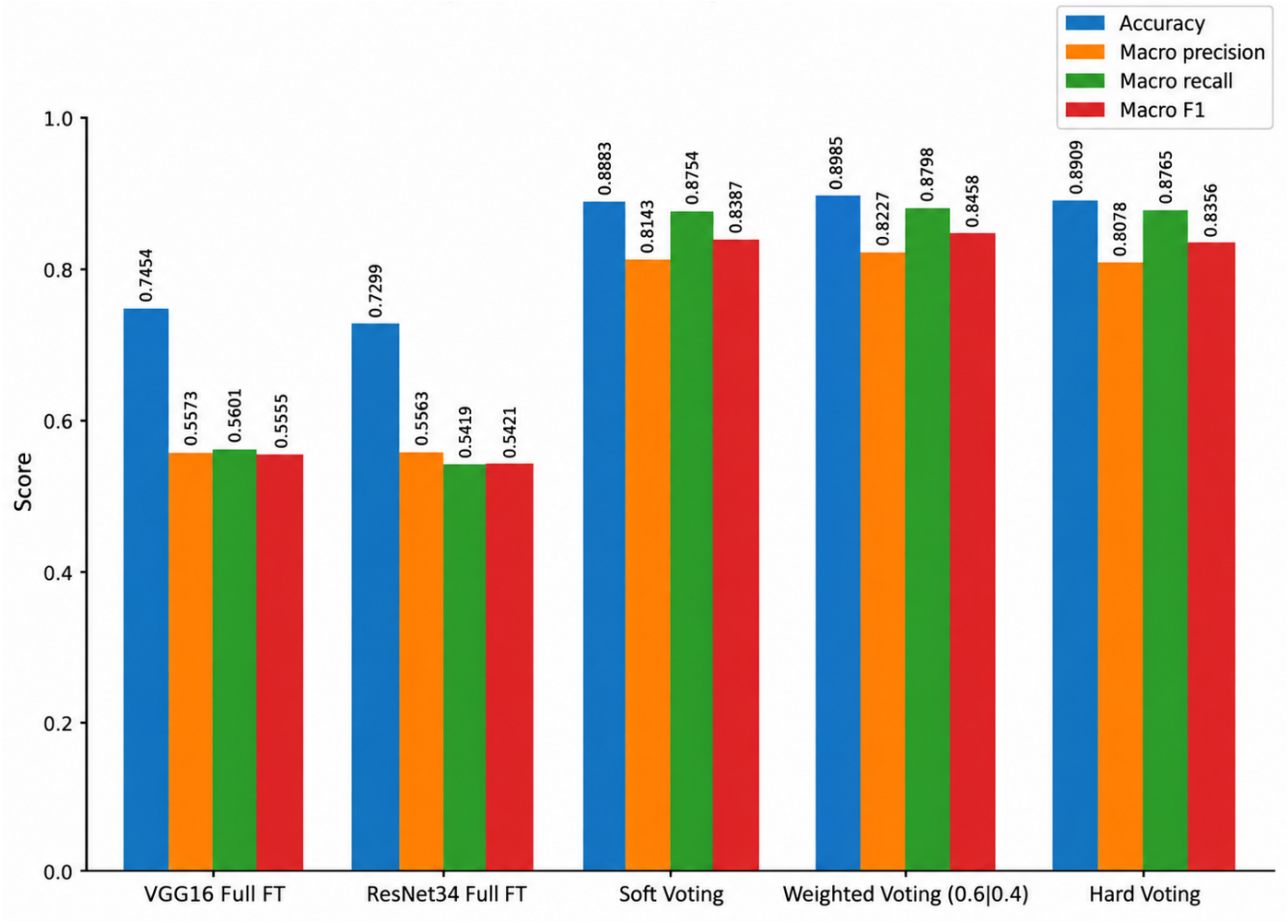


Figure 6. Overall performance comparison of individual models and ensemble methods on the test set

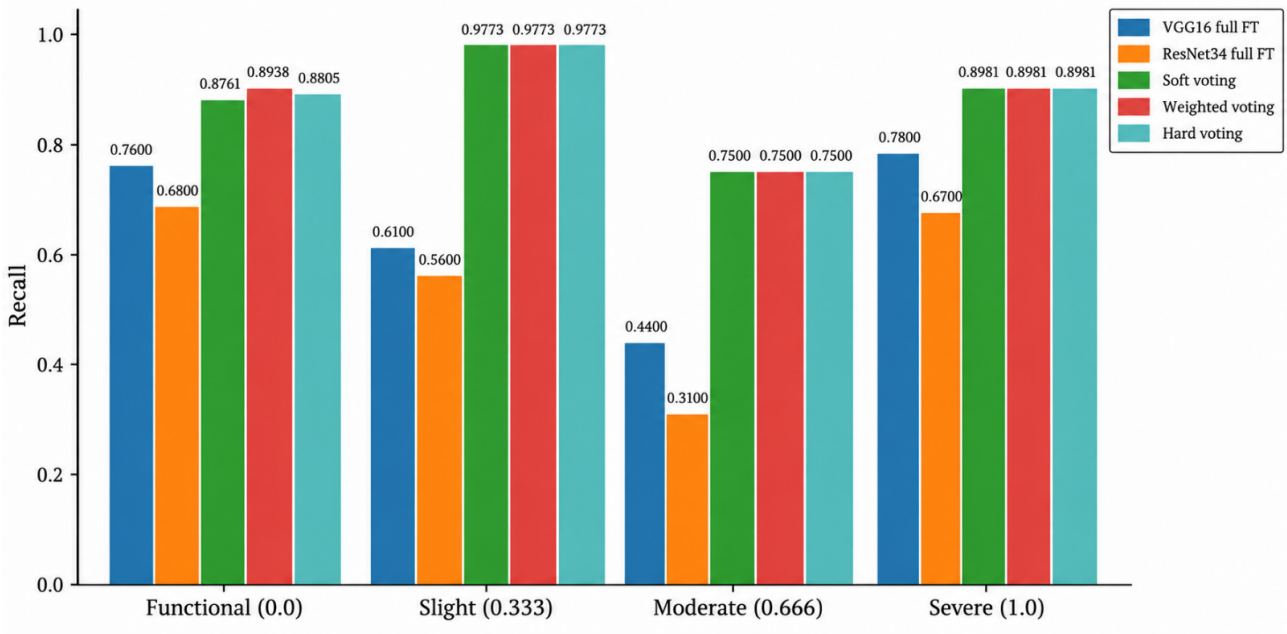


Figure 7. Per-class recall comparison of individual models and ensemble methods on the test set

95% confidence intervals for the differences in accuracy and macro-F1, while McNemar's test was applied to evaluate the paired classification errors between VGG16 and the Weighted Soft Voting ensemble. The statistical comparison is summarized in **Table 13**.

As shown in **Table 13**, the weighted soft voting ensemble achieved a marked improvement over the best-performing individual model across the principal evaluation metrics. Test accuracy increased from 0.7454 to 0.8985, corresponding to an absolute improvement of 0.1531, or 15.31 percentage points. Similarly, macro-F1 increased from 0.5555 to 0.8458, representing an absolute gain of 0.2903. The 95% confidence intervals excluded zero for both accuracy and macro-F1, providing statistical support for the observed ensemble improvement. McNemar's test further indicated a statistically significant difference in paired classification errors between VGG16 and the weighted soft voting ensemble ($p < 0.001$). The moderate-class recall also increased from 0.44 to 0.75; however, because this class contains only 16 test samples, this class-specific improvement should be interpreted cautiously and considered less statistically stable than the overall performance gains.

4.5. Grad-CAM visualization across different severity classes

This section presents a qualitative analysis of model interpretability using Grad-CAM visualizations for four representative PV samples with different defect severity levels (functional, slight, moderate, and severe). The objective is to evaluate how individual models (ResNet34 and VGG16) and ensemble approaches (soft voting and weighted averaging) interpret defect-related features and to assess the consistency between visual attention and prediction confidence. Grad-CAM provides insight into the regions that contribute most to the model's decision, enabling a deeper understanding of model behavior beyond quantitative metrics. This analysis is particularly important in PV defect detection, where interpretability and reliability are critical.

Table 14 presents Grad-CAM visualizations for representative EL samples from the four severity classes: functional, slight, moderate, and severe, as per the attention patterns produced by ResNet34, VGG16, soft voting, and weighted voting. This includes relative activation maps and the corresponding prediction probabilities.

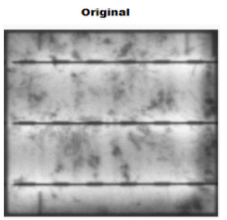
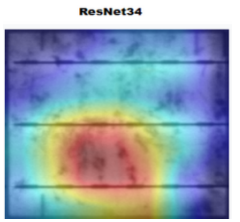
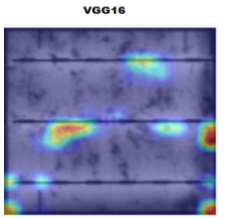
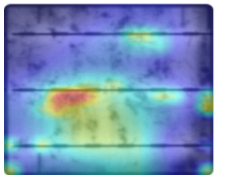
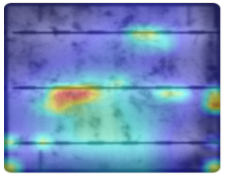
For the functional sample, the defect-free status was predicted with high confidence by all evaluated models. The corresponding Grad-CAM maps exhibited minimal activation, indicating that the models did not strongly focus on specific image regions in this example. For the slight-defect sample, VGG16 showed more focused activation around the subtle defect region, whereas ResNet34 exhibited broader activation over a wider surrounding area. The ensemble methods combined these complementary activation patterns and produced more consolidated attention maps. For the moderate-defect sample, which represents a challenging intermediate condition, VGG16 highlighted several localized regions, while ResNet34 exhibited a more diffuse activation pattern. The ensemble methods integrated these responses and emphasized regions that were consistently attended to by the individual models. For the severe-defect sample, all models produced strong activation around the dominant abnormal region, while the ensemble maps retained clear and visually interpretable attention patterns.

Overall, the Grad-CAM visualizations provide qualitative insight into the regions of the EL images attended to by the evaluated architectures. Across the representative samples, ResNet34 generally exhibited broader activation patterns, whereas VGG16 showed more focused and localized responses. The ensemble methods

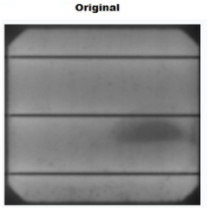
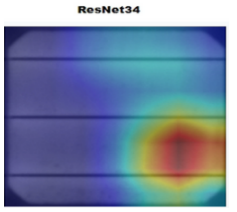
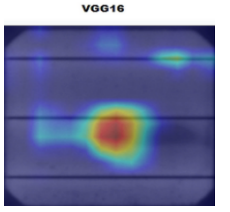
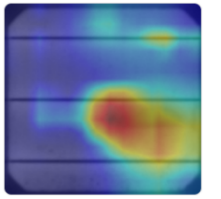
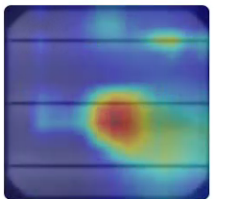
Table 13. Statistical comparison between VGG16 and the weighted soft voting ensemble on the test set

Metric	VGG16	Weighted soft voting	Absolute improvement	95% confidence interval for improvement	Statistical evidence
Accuracy	0.7454	0.8985	+0.1531	0.110, 0.196	McNemar $p < 0.001$
Macro-F1	0.5555	0.8458	+0.2903	0.221, 0.357	Bootstrap confidence interval excludes 0
Moderate recall	0.4400	0.7500	+0.3100	0.080, 0.530	Interpret cautiously; $n = 16$

Table 14. Grad-CAM visualizations for representative electroluminescence (EL) samples

Severity level (sample ID)	Original EL image	Grad-CAM activation maps for ResNet34	Grad-CAM activation for maps VGG16	Ensemble soft voting Grad-CAM	Ensemble weighted voting Grad-CAM
Functional (no defect; Cell2004)					

Model	Class	Prediction	Confidence	Probabilities
VGG16 prediction	Severity	Functional (0.0)	100.00%	Severe (1.0): 0.00017% Slight (0.333): 0.00443% Moderate (0.666): 0.00001% Functional (0.0): 99.99538%
ResNet34 prediction	Severity	Functional (0.0)	92.23%	Severe (1.0): 2.41495% Slight (0.333): 4.58831% Moderate (0.666): 0.76613% Functional (0.0): 92.23060%
Soft voting	Severity	Functional (0.0)	96.11%	Severe (1.0): 1.20756% Slight (0.333): 2.29637% Moderate (0.666): 0.38307% Functional (0.0): 96.11299%
Weighted voting	Severity	Functional (0.0)	96.89%	Severe (1.0): 0.96608% Slight (0.333): 1.83798% Moderate (0.666): 0.30646% Functional (0.0): 96.88947%

Slight defect (Cell1476)					

(*Cont'd...*)

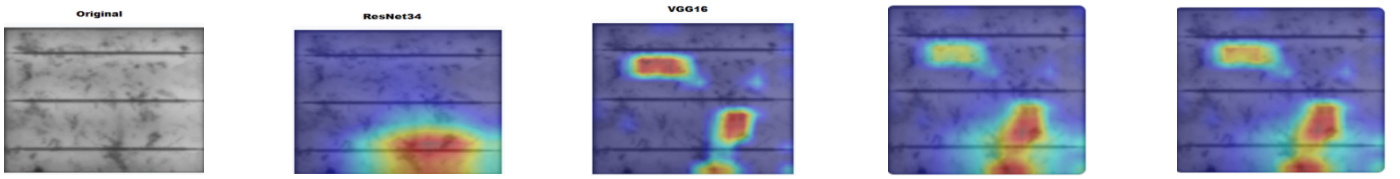
Table 14. (*Continued*)

Severity level (sample ID)	Original EL image	Grad-CAM activation maps for ResNet34	Grad-CAM activation for maps VGG16	Ensemble soft voting Grad-CAM	Ensemble weighted voting Grad-CAM

Corresponding classification results

Model	Class	Prediction	Confidence	Probabilities
VGG16 prediction	Severity	Slight (0.333)	99.89%	Severe (1.0): 0.07063% Slight (0.333): 99.88911% Moderate (0.666): 0.00010% Functional (0.0): 0.04016%
ResNet34 prediction	Severity	Slight (0.333)	99.54%	Severe (1.0): 0.16871% Slight (0.333): 99.53669% Moderate (0.666): 0.00396% Functional (0.0): 0.29064%
Soft voting	Severity	Slight (0.333)	99.71%	Severe (1.0): 0.11967% Slight (0.333): 99.71291% Moderate (0.666): 0.00203% Functional (0.0): 0.16540%
Weighted voting	Severity	Slight (0.333)	99.75%	Severe (1.0): 0.10986% Slight (0.333): 99.74815% Moderate (0.666): 0.00164% Functional (0.0): 0.14035%

Moderate defect (Cell0219)



(*Cont'd...*)

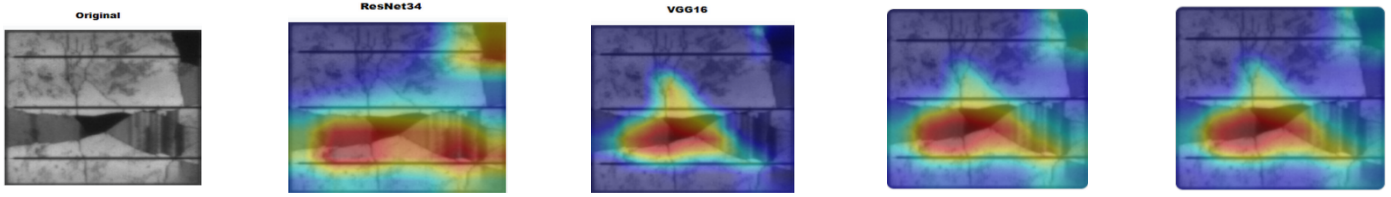
Table 14. (*Continued*)

Severity level (sample ID)	Original EL image	Grad-CAM activation maps for ResNet34	Grad-CAM activation for maps VGG16	Ensemble soft voting Grad-CAM	Ensemble weighted voting Grad-CAM

Model	Class	Prediction	Confidence	Probabilities
VGG16 prediction	Severity	Moderate (0.666)	99.98%	Severe (1.0): 0.01034% Slight (0.333): 0.00002% Moderate (0.666): 99.98369% Functional (0.0): 0.00594%
ResNet34 prediction	Severity	Moderate (0.666)	99.82%	Severe (1.0): 0.17883% Slight (0.333): 0.00260% Moderate (0.666): 99.81725% Functional (0.0): 0.00132%
Soft voting	Severity	Moderate (0.666)	99.90%	Severe (1.0): 0.09459% Slight (0.333): 0.00131% Moderate (0.666): 99.90047% Functional (0.0): 0.00363%
Weighted voting	Severity	Moderate (0.666)	99.92%	Severe (1.0): 0.07774% Slight (0.333): 0.00105% Moderate (0.666): 99.91711% Functional (0.0): 0.00409%

Corresponding classification results

Severe defect (Cell1629)



(*Cont'd...*)

Table 14. (*Continued*)

Severity level (sample ID)	Original EL image	Grad-CAM activation maps for ResNet34	Grad-CAM activation for maps VGG16	Ensemble soft voting Grad-CAM	Ensemble weighted voting Grad-CAM

	Model	Class	Prediction	Confidence	Probabilities
Corresponding classification results	VGG16 prediction	Severity	Severe (1.0)	100.00%	Severe (1.0): 100.00000% Slight (0.333): 0.00000% Moderate (0.666): 0.00000% Functional (0.0): 0.00000%
	ResNet34 prediction	Severity	Severe (1.0)	100.00%	Severe (1.0): 100.00000% Slight (0.333): 0.00000% Moderate (0.666): 0.00000% Functional (0.0): 0.00000%
	Soft voting	Severity	Severe (1.0)	100.00%	Severe (1.0): 100.00000% Slight (0.333): 0.00000% Moderate (0.666): 0.00000% Functional (0.0): 0.00000%
	Weighted voting	Severity	Severe (1.0)	100.00%	Severe (1.0): 100.00000% Slight (0.333): 0.00000% Moderate (0.666): 0.00000% Functional (0.0): 0.00000%

combined these complementary attention behaviors, producing visually consolidated activation maps for the selected examples. The corresponding prediction probabilities were also consistent with the classification outputs of the individual and ensemble models. However, this analysis was intentionally based on one representative example from each of the four severity labels and was therefore intended as a qualitative illustration of model attention behavior rather than as a quantitative validation of defect localization or interpretability. A more systematic evaluation using a larger number of samples and, where available, localization ground truth would be required to quantitatively assess the consistency and spatial correspondence of the activation maps.

5. Discussion

Hereby, a summary of experimental results as well as a comparative analysis between the developed approach and related work are highlighted. In addition, the main limitations of this work and future research directions are discussed.

5.1. Summarizing results

Among the investigated single CNN models (ResNet18, ResNet34, VGG16, VGG19, and EfficientNet-B0), ResNet18 (full fine-tuning) achieved the highest validation accuracy = 0.7520 ± 0.0120, validation macro-precision = 0.5900 ± 0.0140, validation macro-recall = 0.5900 ± 0.0180, and validation macro-F1 = 0.5850 ± 0.0150 while VGG16 (full fine-tuning) achieved the best overall test performance with test accuracy = 0.7454 ± 0.0126, macro-precision = 0.5573 ± 0.0113, macro-recall = 0.5601 ± 0.0168, and macro-F1 = 0.5555 ± 0.0099. Confusion matrices revealed that different individual CNN architectures performed reasonably, except for the moderate severity category classification, which remained the most difficult class because of subtle visual patterns, overlapping severity boundaries, and severe sample scarcity.

To overcome the limitations of individual CNN models, the framework introduced three ensemble-learning strategies: soft voting, weighted soft voting, and hard voting. Ensemble models combined the complementary feature representations of VGG16 and ResNet34, leveraging VGG16's localized defect sensitivity and ResNet34's global contextual understanding. The weighted soft voting ensemble achieved the highest overall classification performance with accuracy = 0.8985, macro-precision = 0.8227, macro-recall = 0.8798, macro-F1 = 0.8458, and weighted average F1-score = 0.9007. Functional-class F1-score reached 0.9266, slight-class F1-score reached 0.8269, moderate-class F1-score reached 0.7273, and severe-class F1-score reached 0.9023. Ensemble learning substantially improved moderate-class detection performance. Moderate-class recall increased from approximately 0.31–0.44 in individual CNN models to 0.75 across all ensemble strategies. Moderate-class F1-score improved from 0.21–0.43 in single models to approximately 0.71–0.73 in ensemble models.

The framework incorporated detailed Grad-CAM explainability analysis for ResNet34, VGG16, and weighted voting ensembles to investigate spatial attention behavior and PV defect localization mechanisms. Grad-CAM visualizations demonstrated complementary attention behaviors between CNN architectures. ResNet34 focused primarily on broad global contextual structures, whereas VGG16 concentrated on localized defect regions. Ensemble methods integrated these complementary representations to produce cleaner, more stable, and more interpretable activation maps aligned with actual PV defect regions.

5.2. Comparative analysis between related studies and the proposed framework

Table 15 provides a comparative analysis between the proposed weighted soft voting ensemble model and the relevant studies that deal with the same ELPV dataset and were presented in **Section 2**.

As per advanced frameworks, such as hybrid ResNet152 + Xception with attention and feature fusion, multi-scale networks combining HazELNet, SqueezeNet, and MobileNet-v2, or ResNet152 with GAN augmentation, they offer strong feature-extraction capabilities but suffer from high architectural complexity, computational cost, and difficult optimization. In contrast, the proposed weighted soft voting ensemble achieves an effective balance between classification performance, robustness, and implementation simplicity by combining the probabilistic outputs of two pretrained CNN models, VGG16 and ResNet34, without requiring complex feature-level fusion. The ensemble assigns a weight of 0.6 to VGG16 and 0.4 to ResNet34, allowing greater influence from the better-performing individual model while preserving the complementary contribution of ResNet34. This decision-level fusion improves predictive performance and

Table 15. Comparative analysis between the proposed weighted soft voting ensemble model and related work

Comparison criterion	**Buerhop-Lutz** *et al.* [18]	**Demirci** *et al.*[23]	**Proposed**	**Ebied** *et al.*[27]	**Al-Otum**[25]	**Wang** *et al.*[24]
Model/method	Pretrained VGG19 deep regression network	Hybrid DFB-SVM + L-CNN	Weighted soft voting ensemble	ResNet152 with GAN augmentation	HazELNet + SqueezeNet + MobileNet-v2 with transfer learning & multi-scale CNN	Hybrid ResNet152 + Xception + CA + mixed pooling + feature fusion
Architectural complexity	Low–moderate	Moderate	Moderate	High	High	Very high
Computational cost	Moderate	Moderate	Moderate	High	High	Very high
Training difficulty	Relatively easy	Moderate	Moderate	Difficult	Difficult	Very difficult
Feature extraction capability	Limited to single backbone features	Better due to CNN + SVM combination	Improved through multiple models	Strong deep feature extraction with synthetic data enhancement	Strong multi-scale feature extraction	Very powerful deep feature learning
Model integration method	Single-model framework	Hybrid ML + DL	Decision-level fusion	Single deep CNN + GAN-based augmentation	Multi-network feature fusion	Deep feature-level fusion with attention
Interpretability	Moderate	Moderate	Relatively high	Moderate–low	Lower	Lower
Generalization ability	Moderate	High	High–very high	Very good due to augmented diversity	Very high	Excellent (with sufficient data)
Robustness to imbalanced data	Limited	Moderate	Very high	Very Good because GAN generates minority samples	High–very high	Very high
Implementation practicality	High practicality	Moderate practicality	High practicality	Moderate	Moderate–low practicality	Low practicality

Abbreviations: CA: Coordinate attention; CNN: Convolutional neural network; DFB-SVM: Deep feature-based support vector machine; DL: Deep learning; GAN: Generative adversarial network; L-CNN: Lightweight convolutional neural network; ML: Machine learning; ResNet: Residual network; SVM: Support vector machine; VGG: Visual Geometry Group.

provides more balanced classification across the considered severity classes, particularly in the presence of class imbalance. Moreover, since the individual CNN models are trained independently, the proposed framework remains relatively easy to optimize, interpret, and maintain, making weighted soft voting a practical and efficient approach for automated PV inspection systems.

Compared with simpler single-model approaches such as the pretrained VGG19 network, the proposed weighted soft voting ensemble provides improved predictive performance by combining the complementary strengths of multiple CNN architectures. Although single models are easier to implement and computationally efficient, they often struggle with imbalanced multi-class datasets and subtle severity differences, leading to biased predictions toward majority classes. By aggregating class probabilities from independently trained models, the weighted soft voting ensemble achieves more balanced and reliable classification performance and improves sensitivity to minority classes while reducing dependence on the limitations of any single architecture. Among the three ensemble strategies evaluated in this study, weighted soft voting achieved the highest overall performance, with an Accuracy of 0.8985 and a macro-F1 score of 0.8458. Consequently, it provides improved predictive consistency and class-balanced performance without the excessive complexity associated with advanced hybrid feature-fusion frameworks.

Therefore, the weighted soft voting ensemble represents an effective middle-ground solution: it achieves the best overall performance among the evaluated ensemble strategies while maintaining a relatively simple decision-level fusion approach. It provides a favorable balance between predictive performance, class-balanced classification, implementation simplicity, and computational requirements, while avoiding the heavy architectural complexity of highly sophisticated hybrid DL frameworks (**Table 15**).

5.3. Innovation boundary

Previous studies on PV defect classification using EL images either relied on single CNN models with limited robustness against data imbalance or highly complex hybrid frameworks requiring expensive computation, advanced augmentation pipelines, and difficult optimization procedures. Although data augmentation techniques, including GAN-based augmentation, improved class balance and model generalization, they often increased computational complexity and training requirements. Moreover, existing ensemble methods still showed limited performance on small and imbalanced multi-class datasets. Therefore, there remains a need for a practical and computationally efficient approach that improves robustness, generalization, and minority-class sensitivity without introducing excessive architectural complexity. This research gap motivated the use of a weighted soft voting ensemble to combine the complementary strengths of multiple pretrained CNN models while maintaining implementation simplicity and reliable classification performance. Results verified the proposed ensemble model's capability in enhancing class balance compared to single CNN models and achieving accuracies comparable to results achieved by highly complex architectures, yet with relatively simpler implementation and less computational cost. Hence, this work focuses on how DL-based image classification can still be reliable when the available dataset is small and significantly class-imbalanced by developing and evaluating a practical DL framework for reliable multi-class image classification under class-imbalanced data conditions.

5.4. Limitations and future work

Although encouraging findings were obtained in this study, the proposed approach still faces limitations related to the small dataset size, class imbalance in intermediate severity levels, and dependence on EL image quality. However, the results reported should be viewed as performance in the specified benchmark and not as generalizable to the larger population outside the laboratory setting. Thus, further validation on larger and more diverse industrial data, from various PV technologies, manufacturers, imaging systems, and operating conditions is needed.

Future work may focus on integrating more advanced augmentation methods, adaptive class-balanced loss functions, attention-based lightweight architectures, and transformer-based models to further improve feature representation and minority-class recognition. In addition, future research can explore self-supervised learning approaches to reduce dependence on labeled data, as well as multimodal fusion of EL, thermal, and RGB images to enhance fault diagnosis reliability. Evaluating the proposed

framework on larger industrial-scale datasets and implementing real-time edge deployment for unmanned aerial vehicle- or Internet of Things-based PV monitoring systems also represent important future research directions.

6. Conclusion

This study presented a comparative investigation of DL models for PV defect severity classification using EL images, with the aim of improving automated PV inspection and predictive maintenance. Several pretrained CNN architectures, including ResNet18, ResNet34, EfficientNet-B0, VGG16, and VGG19, were evaluated using different transfer learning strategies, where full fine-tuning consistently achieved better performance than staged fine-tuning. Although individual models showed moderate results due to class imbalance and limited samples in intermediate severity categories, VGG16 and ResNet18 provided the best overall balance between accuracy and generalization, while ResNet34 demonstrated strong defect localization capability. To overcome the limitations of single models, ensemble learning techniques were introduced and significantly improved classification performance, increasing the macro-F1 score from approximately 0.55 to 0.84. Furthermore, Grad-CAM analysis confirmed that ensemble methods produced more reliable and interpretable attention maps by combining the complementary strengths of different models. Overall, the findings highlight the effectiveness of ensemble learning and explainable AI in enhancing the robustness, accuracy, and interpretability of PV defect severity classification systems, yet at relatively lower implementation complexity and computational burden compared to related work in the literature dealing with the same ELPV dataset.Top of Form

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: All authors
Data curation: Fouad Sleibi
Formal analysis: Fouad Sleibi, Emma Qumsiyeh
Investigation: Fouad Sleibi
Methodology: All authors
Software: Fouad Sleibi
Supervision: Emma Qumsiyeh, Nahla E. Zakzouk
Validation: All authors
Visualization: Fouad Sleibi
Writing – original draft: Fouad Sleibi
Writing – review & editing: Emma Qumsiyeh, Nahla E. Zakzouk

Availability of data

The data analyzed in this study are publicly available through the Electroluminescence Photovoltaic (ELPV) dataset repository. The dataset contains 2,624 electroluminescence images of photovoltaic cells extracted from 44 photovoltaic modules, together with defect-probability annotations and module-type information. The dataset is available at: https://github.com/zae-bayern/elpv-dataset. The official repository provides the dataset characteristics and licensing information.

AI tools statement

OpenAI ChatGPT was used during manuscript preparation to assist with language editing, grammatical refinement, and improving the text's clarity and organization. The authors critically reviewed and revised all AI-assisted content and take full responsibility for the accuracy, integrity, and final content of the manuscript.

References

1. Guo Q, Ingram A. Powering the Future: Advancements and Applications of Photovoltaic Systems with Political, Economic, and Environmental Considerations. *Eng Sci Technol.* 2024;5(1):109-120.

   https://doi.org/10.37256/est.5120243972

2. El-Banby GM, Moawad NM, Abouzalm BA, Abouzaid WF, Ramadan EA. Photovoltaic system fault detection techniques: a review. *Neural Comput Appl.* 2023;35(35):24829-24842.

   https://doi.org/10.1007/s00521-023-09041-7

3. Liu Y, Wu Y. Fault diagnosis of photovoltaic modules: A review. *Sol Energy.* 2025;293:113489.

   https://doi.org/10.1016/j.solener.2025.113489

4. Taghezouit B, Harrou F, Sun Y, Merrouche W. Model-based fault detection in photovoltaic systems: A comprehensive review and avenues for enhancement. *Result Eng.* 2024;21:101835.

   https://doi.org/10.1016/j.rineng.2024.101835

5. Mohamed KA, Zakzouk NE, Abdelgeliel M, Youssef KH. Faults and Faults' Diagnosis Techniques in Photovoltaics: An Overview. In: *2025 IEEE PES/IAS PowerAfrica*. IEEE; 2025:1-6.

   https://doi.org/10.1109/PowerAfrica65840.2025.11289065

6. Yao S, Kang Q, Zhou M, Abusorrah A, Al-Turki Y. Intelligent and Data-Driven Fault Detection of Photovoltaic Plants. *Processes.* 2021;9(10):1711.

   https://doi.org/10.3390/pr9101711

7. Mohamed KA, Zakzouk NE, Abdelgeliel M, Youssef KH. Online Fault Detection, Classification and Localization in PV Arrays Using Feedforward Neural Networks and Residual-Based Modeling. *Technologies.* 2026;14(2):130.

   https://doi.org/10.3390/technologies14020130

8. Ali MU, Zafar A, Kim KS. Imaging based fault diagnosis of photovoltaic modules: a systematic review. *Sol Energy.* 2026;314:114679.

   https://doi.org/10.1016/j.solener.2026.114679

9. Sadeghi R, Memme S, Morchio S, Fossa M, Parenti M. Infrared Thermography in Photovoltaic Systems: A Review for Maximizing Energy Yield and Long-Term Reliability. *Energies.* 2026;19(6):1570.

   https://doi.org/10.3390/en19061570

10. Del Prado Santamaría R, Dhimish M, Dos Reis Benatto GA, Kari T, Poulsen PB, Spataru SV. From Indoor to Daylight Electroluminescence Imaging for PV Module Diagnostics: A Comprehensive Review of Techniques, Challenges, and AI-Driven Advancements. *Micromachines.* 2025;16(4):437.

    https://doi.org/10.3390/mi16040437

11. Deitsch S, Christlein V, Berger S, *et al.* Automatic classification of defective photovoltaic module cells in electroluminescence images. *Sol Energy.* 2019;185:455-468.

    https://doi.org/10.1016/j.solener.2019.02.067

12. Hussain T, Hussain M, Al-Aqrabi H, Alsboui T, Hill R. A Review on Defect Detection of Electroluminescence-Based Photovoltaic Cell Surface Images Using Computer Vision. *Energies.* 2023;16(10):4012.

    https://doi.org/10.3390/en16104012

13. Almashhadani RAI, Hock GC, Bt Nordin FH, Abdulrazzak HN. Electroluminescence Images for Solar Cell Fault Detection Using Deep Learning for Binary and Multiclass Classification. *SSRG-IJEEE.* 2024;11(5):150-160.

    https://doi.org/10.14445/23488379/IJEEE-V11I5P114

14. Matusz-Kalász D, Bodnár I, Jobbágy M. An Overview of CNN-Based Image Analysis in Solar Cells, Photovoltaic Modules, and Power Plants. *Appl Sci.* 2025;15(10):5511.

    https://doi.org/10.3390/app15105511

15. Sonawane PR, Bhandari S, Patil RB, Al-Dahidi S. Comparative analysis of methods used to solve class imbalance problem in dataset for CNN-based fault diagnosis of solar photovoltaic panels. *Int J Thermofluids.* 2025;28:101303.

    https://doi.org/10.1016/j.ijft.2025.101303

16. Kumar R, Kim YW, Byun YC. Mitigating dataset imbalance using image-based stable diffusion and feature-level SMOTE for solar panel classification with CNNs. *Energy Rep.* 2026;15:109055.

    https://doi.org/10.1016/j.egyr.2026.109055

17. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv.* Preprint posted online 2014.

    https://doi.org/10.48550/ARXIV.1409.1556

18. Buerhop-Lutz C, Deitsch S, Maier A, *et al.* A benchmark for visual identification of defective solar cells in electroluminescence imagery. In: Proceedings of the 35th European Photovoltaic Solar Energy Conference and Exhibition (EU PVSEC); 2018:1287-1289.

    https://doi.org/10.4229/35THEUPVSEC20182018-5CV.3.15.

19. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM.* 2017;60(6):84-90.

    https://doi.org/10.1145/3065386

20. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag.* 2009;45(4):427-437.

    https://doi.org/10.1016/j.ipm.2009.03.002

21. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: *2017 IEEE International Conference on Computer Vision (ICCV).* IEEE; 2017:618-626.

    https://doi.org/10.1109/ICCV.2017.74

22. He H, Garcia EA. Learning from Imbalanced Data. *IEEE Trans Knowl Data Eng.* 2009;21(9):1263-1284.

    https://doi.org/10.1109/TKDE.2008.239

23. Demirci MY, Beşli N, Gümüşçü A. Efficient deep feature extraction and classification for identifying defective photovoltaic module cells in Electroluminescence images. *Expert Syst Appl.* 2021;175:114810.

https://doi.org/10.1016/j.eswa.2021.114810

24. Wang J, Bi L, Sun P, *et al.* Deep-Learning-Based Automatic Detection of Photovoltaic Cell Defects in Electroluminescence Images. *Sensors.* 2023;23(1):297.

https://doi.org/10.3390/s23010297

25. Al-Otum HA. Automatic Defect Detection and Classification in Electroluminescence Images of Pv Cells Using Convolutional Neural Networks. *SSRN.* Preprint posted online 2023.

https://doi.org/10.2139/ssrn.4422413

26. Tella H, Hussein A, Rehman S, Liu B, Balghonaim A, Mohandes M. Solar photovoltaic panel cells defects classification using deep learning ensemble methods. *Case Stud Therm Eng.* 2025;66:105749.

https://doi.org/10.1016/j.csite.2025.105749

27. Ebied MA, Munshi A, Alhuzali SA, El-sotouhy MM, Shehta AI, Elborlsy MS. Advanced deep learning modeling to enhance detection of defective photovoltaic cells in electroluminescence images. *Sci Rep.* 2025;15(1):31640.

https://doi.org/10.1038/s41598-025-14478-y

28. Japkowicz N, Stephen S. The class imbalance problem: A systematic study. *IDA.* 2002;6(5):429-449.

https://doi.org/10.3233/IDA-2002-6504

29. Deitsch S, Buerhop-Lutz C, Sovetkin E, *et al.* Segmentation of photovoltaic module cells in uncalibrated electroluminescence images. *Mach Vis Appl.* 2021;32(4):84.

https://doi.org/10.1007/s00138-021-01191-9

30. Liu Q, Liu M, Wang C, Wu QMJ. An efficient CNN-based detector for photovoltaic module cells defect detection in electroluminescence images. *Sol Energy.* 2024;267:112245.

https://doi.org/10.1016/j.solener.2023.112245

31. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE; 2016:770-778.

https://doi.org/10.1109/CVPR.2016.90

32. Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: *Proceedings of the 36th International Conference on Machine Learning.* Vol 97. Proceedings of Machine Learning Research; 2019:6105–6114. https://proceedings.mlr.press/v97/tan19a.html

33. Paszke A, Gross S, Massa F, *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv.* In: *Advances in Neural Information Processing Systems 32.* Curran Associates, Inc; 2019:8024-8035. https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

34. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. *arXiv.* Preprint posted online 2017.

https://doi.org/10.48550/ARXIV.1711.05101

35. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *arXiv.* Preprint posted online 2014.

https://doi.org/10.48550/ARXIV.1412.6980

36. Sleibi F. PV EL Defect Severity Ensemble [software]. Hugging Face Space. Accessed September 12, 2026. https://huggingface.co/spaces/fouadsl/PV-EL-Defect-Severity-Ensemble

Fouad Sleibi *received the B.Sc. degree in Communications and Information Technology from Al-Quds Open University, Palestine. He is currently pursuing the M.Sc. degree in Artificial Intelligence and Data Analytics at Palestine Ahliya University, Bethlehem, Palestine. Since 2010, he has been working as a Technology and Computer Teacher at Terra Santa School, where he also serves as a computer laboratory manager and contributes to networking, programming, and online learning activities. He is also an accredited trainer at the Artificial Intelligence and Innovation Center at Palestine Ahliya University, contributing to training activities and programs related to artificial intelligence and emerging technologies. His current academic and research interests include artificial intelligence, machine learning, deep learning, computer vision, data analytics, and intelligent systems.*

https://orcid.org/0009-0005-6797-439X

Emma Qumsiyeh *received the B.Sc. degree in Computer Engineering from the University of Aleppo, Syria, in 2011, and the Master of Arts in Teaching (MAT) in Science from Al-Quds Bard College, Palestine, in 2018. She received the Ph.D. degree in Information Technology Engineering from Al-Quds University, Palestine, in 2025. She is currently an Assistant Professor and the Coordinator of the Computer Engineering Program at the Faculty of Engineering and Information Technology, Palestine Ahliya University (PAU), Palestine. She also serves as the focal point for the IEEE Student Branch at PAU. She has published several research papers in international peer-reviewed journals and participated in*

international scientific conferences. Her research interests include artificial intelligence, machine learning, data science, bioinformatics, biomedical data analysis, and knowledge-based intelligent systems.

https://orcid.org/0000-0002-3797-5851

Nahla E. Zakzouk *received the B.Sc.and M.Sc. degrees from the Arab Academy for Science and Technology and Maritime Transport (AASTMT) , Egypt, in 2007 and 2009, respectively. In 2015, she received the Ph.D. degree in Electrical Engineering from the College of Engineering, Strathclyde University, U.K. Since 2009, she was working as an Assistant Teacher in AASTMT till 2015 when she was appointed as a Lecturer. She was promoted to Associate Professor in 2019, then Professor of Renewable Energy Systems in 2023. Since March 2025, she was appointed as the Head of the Electrical and Control Engineering Department in AASTMT. Her research interests include power quality, distributed generation, power converters and artificial intelligence in renewable energy systems.*

https://orcid.org/0000-0002-4421-0080

An International Journal of Optimization and Control: Theories & Applications (https://accscience.com/journal/ijocta)



This work is licensed under a Creative Commons Attribution 4.0 International License. The authors retain ownership of the copyright for their article, but they allow anyone to download, reuse, reprint, modify, distribute, and/or copy articles in IJOCTA, so long as the original authors and source are credited. To see the complete license contents, please visit http://creativecommons.org/licenses/by/4.0/.