

DRONE-BASED SOLAR PANEL FAULT CLASSIFICATION USING INFRARED IMAGES AND DEEP LEARNING: A COMPARATIVE STUDY OF A LIGHTWEIGHT CNN AND RESNET18

Farkhod Akhmedov¹, Toshtemir Abdikhafizovich Khujakulov², Halimjon
Ergashevich Khujamatov^{3*}

1. Department of Computer Engineering, University of Tashkent for Applied Sciences, Tashkent 100125, Uzbekistan; farhodsam92@gmail.com (F.A),
2. Department of Computer Engineering, University of Tashkent for Applied Sciences, Tashkent 100125, Uzbekistan; t.khujakulov@utas.uz
3. Department of Electronics and Instrumentation, Fergana State Technical University, Fergana 150100, Uzbekistan, kh.khujamatov@tuit.uz

Abstract

Efficient monitoring of photovoltaic (PV) modules is essential for sustaining energy yield and reducing manual inspection effort in solar power plants. Drone-based infrared (IR) imaging is a practical inspection modality because thermal anomalies often indicate electrical or physical faults. This paper studies automatic module-level fault screening on real IR module crops acquired by a small unmanned aerial system (sUAS). Because several fine-grained fault categories contain very few samples, the original multi-class labels were merged into a binary normal versus fault task. A compact custom CNN and a ResNet18 network were trained for 100 epochs under identical settings (96 × 96 grayscale input, batch size 32, class-weighted loss) on 1,523 images split into 1,065 training, 228 validation and 230 test images. On the held-out test set, ResNet18 reached 89.57% accuracy (95% CI 84.9–92.9%), 89.58% weighted F1, 95.99% ROC-AUC and 96.24% average precision, versus 76.09% accuracy, 75.65% weighted F1, 88.79% ROC-AUC and 91.96% average precision for the small CNN. The gain stems almost entirely from specificity: fault recall was similar (88.7% vs. 87.1%), but the false-alarm rate on healthy modules fell from 36.8% to 9.4% and the Matthews correlation coefficient rose from 0.522 to 0.791. Training-curve analysis shows that the small CNN is

capacity-limited and unstable on validation data, whereas ResNet18 converges smoothly with a moderate generalization gap. The results indicate that residual CNNs are a suitable baseline for IR-based PV module fault screening, with the caveat that evaluation was confined to a single-source dataset and a 230-image test set.

Keywords: photovoltaic fault detection; drone inspection; infrared thermography; computer vision; ResNet18; solar energy monitoring
Introduction Solar photovoltaic energy systems are expanding rapidly because they provide clean and scalable electricity generation. However, PV plants require continuous operation and maintenance to remain efficient. Faults such as hot cells, open-circuit modules, short-circuit patterns, shading and other abnormal thermal behaviors reduce output and can create safety risks if left undetected. Manual inspection is time-consuming in large installations, and electrical measurements, while useful for identifying under performing strings or modules, do not always localize fault patterns at the image level.

Drone-based inspection is an attractive alternative because unmanned aerial systems can survey a large plant quickly. Infrared thermography is particularly informative because many PV faults appear as temperature differences within a module. Combined with computer vision, IR imagery can support automatic screening of faulty modules and help maintenance teams prioritize physical inspection.

This paper addresses the computer-vision stage of such a workflow. Rather than a full 36 industrial inspection pipeline, it evaluates how well compact deep models classify cropped 37 IR module images into normal and faulty classes. The dataset is real, the training protocol 38 is reproducible, and the analysis uses standard classification metrics supplemented by 39 uncertainty estimates. The main contributions are:

A reproducible binary PV module fault classification pipeline built on real drone- 41 acquired IR module crops. A controlled comparison of a lightweight custom CNN and ResNet18 under an 43 identical 100-epoch training setup.

A multi-faceted evaluation: accuracy, balanced accuracy, precision, recall, F1, MCC, 45 Cohen's kappa, confusion matrices, ROC-AUC and average precision, with confidence 46 intervals and a significance test.

An analysis of training dynamics and error structure, and a discussion of how module- 48 level screening can support drone-based maintenance and energy-loss analysis. Related Work PV fault detection has been studied with electrical measurements, visible imaging, electroluminescence imaging and infrared thermography. Electrical methods suit systemlevel diagnosis, whereas image-based methods can localize visible or thermal patterns; IR thermography is especially relevant to drone inspection because abnormal heating often corresponds to operational faults.

UAV-based thermal imaging combined with deep learning has been shown to support automated plant monitoring. Bommers et al. [2] built a computer-vision tool that extracts PV modules from aerial IR videos and trained a ResNet-50 classifier for module anomalies, reporting more than 90% test accuracy, and stressed the value of semi-automatic module extraction for scalable inspection.

Public datasets have improved reproducibility. The Infrared Solar Modules dataset from Raptor Maps [3] contains 20,000 small IR module images in 12 classes (one no-anomaly class and eleven anomaly classes), and PVF-10 [4] was introduced for fine-grained UAV thermal-IR fault classification. The sUAS Infrared Image Dataset for PV Module Fault Classification [1] used here provides cropped grayscale IR module images organized by fault type, which suits a module-level classification experiment.

Deep learning is now the dominant approach for image-based PV module fault classification, with prior studies using custom CNNs, transfer learning, ResNet-style models, YOLO-based defect detection and segmentation-classification pipelines [5]–[7]. For small projects where only cropped module images are available, a full detection or segmentation system is unnecessary; a binary classifier answers the first practical question—is this module normal or faulty?—before fine-grained classification or full-scene localization is attempted.

Dataset and Preprocessing The experiments use the sUAS Infrared Image Dataset for Photovoltaic Module Fault Classification [1] (Zenodo, DOI 10.5281/zenodo.20860870), which contains cropped 8-bit grayscale IR images of annotated PV modules from an sUAS inspection workflow. The original classes are Cell, Multi-Cell, Open-Circuit Substring, Open-Circuit, Open-Circuit Multi-Substring, Short-Circuit Substring, Short-Circuit and Normal.

The labels were converted into a binary task: Normal (modules labeled normal) and Fault (all non-normal categories). This was chosen because some original classes contain very few examples, which would make a fair multi-class comparison impossible at this scale. After limiting the normal class to reduce imbalance, the working dataset contained 1,523 images (Table 1) divided into training, validation and test subsets (Table 2). All images 84 were resized to 96×96 pixels. Because the images are single-channel IR, classification relies 85 on thermal structure, shape and intensity patterns rather than color.

Table 1. Class distribution of the binary dataset.

Class	Images	Share
Fault	823	54.0%
Normal	700	46.0%
Total	1,523	100%

Table 2. Data split used in all experiments.

Subset	Images	Share	Class composition
Training	1,065	69.9%	not recorded
Validation	228	15.0%	not recorded
Test	230	15.1%	124 fault / 106 normal

Methodology Workflow The workflow has five stages: (i) import drone-acquired IR module crops; (ii) map the original fault types to a binary label; (iii) split into training, validation and test subsets; (iv) train two CNN classifiers for 100 epochs; and (v) evaluate the best validation checkpoint on the held-out test set. It can later be extended with full-image module detection, georeferencing and plant-level energy-loss estimation.

Models Small CNN. A lightweight convolutional network serves as a baseline to test whether the task can be learned without a deep architecture. It is computationally inexpensive and suitable for low-resource deployment, but may not capture complex thermal patterns as effectively as deeper networks.

ResNet18. ResNet18 [8] is a residual network whose identity shortcut connections improve gradient flow and ease optimization of deeper models. It is used here as a stronger CNN baseline for module-level fault classification.

Training setup and model selection Both models were trained on the same processed dataset with the configuration in

Table 3. A class-weighted loss compensates for the imbalance between fault and normal

Samples. Validation accuracy was monitored every epoch and the best checkpoint was saved; all reported test metrics use this saved checkpoint rather than the final-epoch weights. **Evaluation metrics** With fault as the positive class, the confusion matrix yields TP, FN, FP and TN. We report accuracy, balanced accuracy, precision, recall (sensitivity), specificity, F1 (macro and weighted), ROC-AUC and average precision (AP). Because accuracy alone can mask asymmetric errors, we add the Matthews correlation coefficient (MCC) [9] and Cohen's kappa [10], both of which account for chance agreement and class balance. Uncertainty is quantified with Wilson 95% intervals [11] for proportions and Hanley–McNeil approximate

Table 3. Training and evaluation configuration shared by both models.

Setting Value

Task Binary classification (fault = positive class) **Input** 96×96 single-channel IR image **Epochs** Batch size Loss Class-weighted loss **Model selection** Best validation accuracy (228 images) **Final evaluation** Held-out test set (230 images) intervals [12] for ROC-AUC. A two-proportion z-test on accuracy compares the models; 115 since both were evaluated on the same test images, a paired test (e.g., McNemar) would 116 be more powerful, so the unpaired test used here is conservative. ROC-AUC measures 117 ranking quality across all thresholds, while AP summarizes the

precision–recall trade-off, 118 which is more informative than ROC when the positive class matters most [13].

Experimental Results Overall comparison

Table 4. Test-set performance after 100 training epochs (n = 230). Acc. = accuracy; AP = average precision.

Model	Acc.	Bal. acc.	Macro F1	Weighted F1	ROC-AUC	AP
Small CNN	0.7609	0.7515	0.7530	0.7565	0.8879	0.9196
ResNet18	0.8957	0.8964	0.8953	0.8958	0.9599	0.9624
Δ (ResNet18 – Small CNN)	+0.1348	+0.1449	+0.1423	+0.1393	+0.0720	+0.0428

ResNet18 outperforms the small CNN on every metric (Table 4). Weighted F1 improves by 13.93 percentage points (0.7565 to 0.8958), and ROC-AUC rises from 0.8879 to 0.9599, indicating clearly better separation of normal and faulty modules. The smaller gain in AP (+4.28 points) reflects the fact that AP is already high for the weaker model because faults are the slightly larger class (54% of the test set), which raises the AP baseline.

Class-level performance

Table 5. Class-level results of ResNet18.

Class	Precision	Recall	F1-score	Support
Fault	0.9167	0.8871	0.9016	124
Normal	0.8727	0.9057	0.8889	106
Weighted average	0.8964	0.8957	0.8958	230

For ResNet18, the fault class reached 91.67% precision and 88.71% recall: predicted faults are usually correct, but 14 of 124 faulty test modules were missed. Normal-class recall was 90.57%, showing reliable recognition of healthy modules,

and the two classes have similar F1-scores (0.902 and 0.889), i.e., the model is not biased toward either class.

The small CNN shows a markedly asymmetric profile. Its fault recall (87.10%) is almost identical to that of ResNet18, but its normal-class recall is only 63.21% and its fault precision only 73.47%: it achieves sensitivity by over-flagging, assigning the fault label to many healthy modules.

Out of 230 test images, ResNet18 classified 206 correctly and misclassified 24 (14 false negatives, 10 false positives). The small CNN classified 175 correctly and misclassified

Table 6. Class-level results of the small CNN.

Class	Precision	Recall	F1-score	Support
Fault	0.7347	0.8710	0.7970	124
Normal	0.8072	0.6321	0.7090	106
Weighted average	0.7681	0.7609	0.7565	230

55 (16 false negatives, 39 false positives). The two models therefore miss almost the same 138 number of faults (14 vs. 16), but ResNet18 produces 74% fewer false alarms (10 vs. 39). 139 In maintenance applications false negatives are generally the costlier error, because an 140 undetected fault continues to degrade yield; the roughly 11% miss rate of both models is 141 therefore the main practical limitation. False alarms, however, are not free: the small CNN 142 would send technicians to inspect 39 healthy modules per 106, a workload that undermines 143 the point of automated screening.

Diagnostic metrics and statistical uncertainty

Table 7. Derived diagnostic metrics from the test confusion matrices (fault = positive class; n = 230).

Metric	Small CNN	ResNet18
Sensitivity / fault recall (TPR)	87.1%	88.7%
Specificity (TNR)	63.2%	90.6%

False-alarm rate (FPR)	36.8%	9.4%
Miss rate (FNR)	12.9%	11.3%
Positive predictive value (fault precision)	73.5%	91.7%
Negative predictive value	80.7%	87.3%
Matthews correlation coefficient	0.522	0.791
Cohen's kappa	0.511	0.791
Accuracy, 95% Wilson CI	76.1% [70.2, 81.1]	89.6% [84.9, 92.9]
ROC-AUC, 95% Hanley–McNeil CI	0.888 [0.845, 0.931]	0.960 [0.935, 0.985]

Table 7 makes the source of the improvement explicit. Sensitivity is statistically indistinguishable between the models (88.7% vs. 87.1%; a difference of two detections), whereas specificity rises from 63.2% to 90.6%. Chance-corrected agreement tells the same story: MCC increases from 0.522 to 0.791 and kappa from 0.511 to 0.791. The Wilson 95% intervals for accuracy (70.2–81.1% vs. 84.9–92.9%) do not overlap, and the unpaired two-proportion test gives $z = 3.90$ ($p < 0.001$) for the 13.5-point accuracy difference. The approximate AUC intervals are also disjoint (0.845–0.931 vs. 0.935–0.985), though only narrowly, so the AUC advantage should be considered suggestive rather than conclusive.

The intervals themselves are wide (about ± 4 points for ResNet18) because the test set has only 230 images. Training dynamics Figure 1 and Figure 2 shows the learning curves. ResNet18 improves steadily: training accuracy rises from 63.7% at epoch 1 to 93.9% at epoch 100 and training loss falls from 0.693 to 0.148. Validation accuracy follows the same trend with early fluctuations (standard deviation 4.1 points over epochs 1–20, 1.7 points over epochs 81–100). The best validation accuracy, 88.60% (202/228 images), occurs at epoch 99, which is the checkpoint used for testing. The lowest validation loss (0.310) occurs earlier, at epoch 90, and validation loss plateaus around 0.4 from roughly epoch 40 onward while training loss keeps decreasing.

At epoch 100 the train-validation accuracy gap is 6.2 points (about 6.5 points averaged over the last ten epochs), a moderate generalization gap indicating mild overfitting rather than

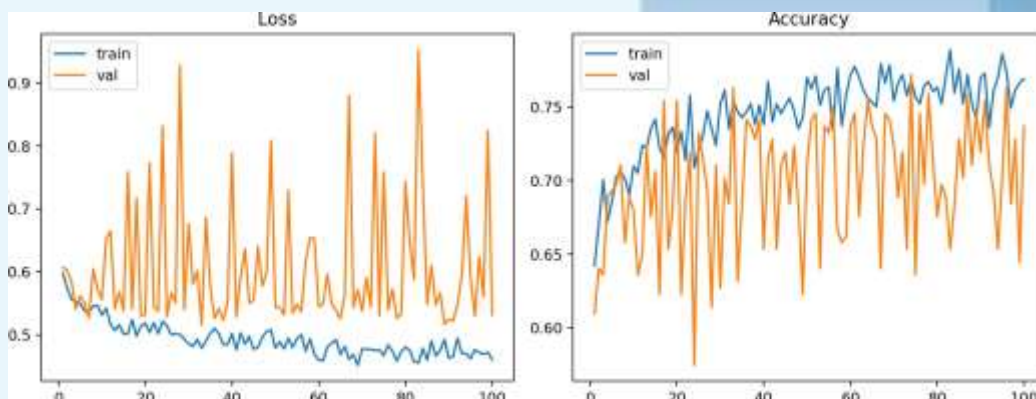


Figure 1. Training and validation loss (left) and accuracy (right) over 100 epochs for small CNN.

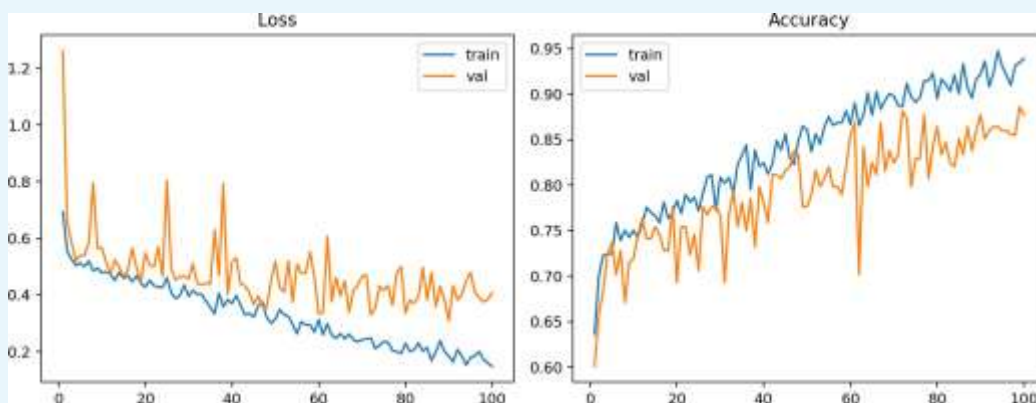


Figure 2. Training and validation loss (left) and accuracy (right) over 100 epochs for ResNet18 .

failure to generalize. Because one validation image corresponds to 0.44 percentage points, 166 selecting the checkpoint on validation accuracy is itself noisy; the mean validation accuracy 167 over the last 20 epochs (85.4%) is a more conservative indicator of expected behavior, and 168 the test accuracy of 89.6% is consistent with the upper range of that band.

The small CNN in Figure 1 behaves very differently. Training accuracy saturates at 170 roughly 0.75–0.79 and training loss at about 0.46–0.50 (values read from the figure; the numerical history for this run was not supplied), while validation loss swings erratically between about 0.5 and 0.95 and validation accuracy between about 0.57 and 0.78 even late in training. Because the training accuracy (≈ 0.77) is close to the final test accuracy (0.761), the small CNN appears to be capacity-

limited (underfitting) rather than overfit, and its unstable validation behavior implies that the selected checkpoint depends strongly on which epoch happened to score well on the validation split. The model was thus both less accurate and less reliable to train.

Figures 3 and 4 present the receiver operating characteristic (ROC) curves of the small CNN and ResNet18, respectively, for binary PV module fault classification, with fault treated as the positive class. The curves illustrate the relationship between the true-positive rate and false-positive rate across different classification thresholds. The dashed diagonal represents the performance of a random classifier.

The small CNN achieves an area under the ROC curve (AUC) of 0.888, indicating substantial discrimination between normal and faulty modules. Its true-positive rate reaches approximately 0.80 at a false-positive rate of 0.20. Beyond this region, further improvements in fault sensitivity require progressively higher false-positive rates, illustrating the trade-off between fault detection and unnecessary inspection of healthy modules.

ResNet18 achieves a higher AUC of 0.960, with its ROC curve positioned closer to the upper-left corner. At a false-positive rate of approximately 0.10, the model attains a truepositive rate of about 0.90. Compared with the small CNN, ResNet18 therefore provides higher fault sensitivity at comparable false-positive rates over most of the plotted range.

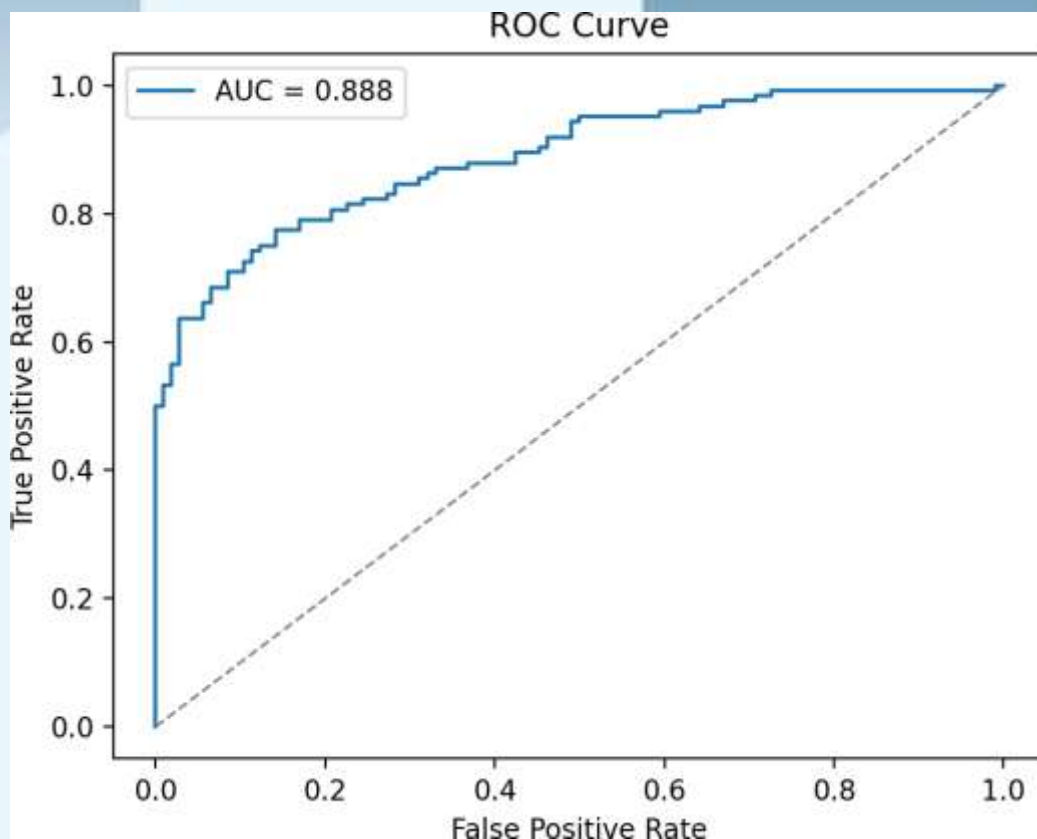


Figure 3. Receiver operating characteristic curve of the small CNN for PV module fault classification (AUC = 0.888). The dashed diagonal indicates random-classifier performance.

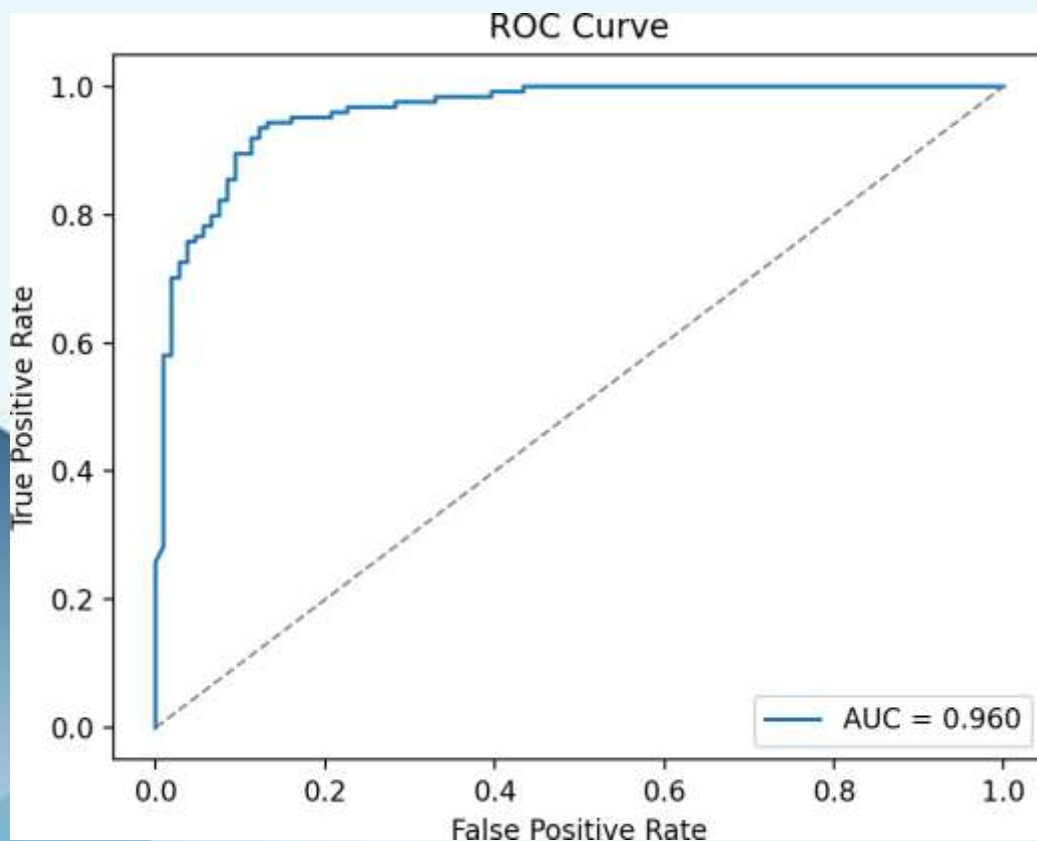


Figure 4. Receiver operating characteristic curve of ResNet18 for PV module fault classification (AUC

= 0.960). The dashed diagonal indicates random-classifier performance. The AUC increase of 0.072 supports improved discrimination on this test set, although statistical significance requires a paired comparison of the models' prediction scores.

Figure 5 presents the distribution of predicted fault probabilities produced by the small CNN for actual normal and faulty test modules. Faulty modules are predominantly concentrated at higher scores, with the largest frequency occurring near a predicted probability of 1.0. Normal modules generally receive lower scores, although their distribution extends into the higher-probability region. The overlap between the two distributions indicates that some healthy and faulty modules receive similar prediction scores, limiting their separation using a single decision threshold.

At a classification threshold of 0.5, normal modules with scores above the threshold generate false positives, whereas faulty modules with scores below it generate false negatives. The distribution therefore provides a visual explanation for the small CNN's false alarms and missed faults. Increasing the threshold could reduce false positives at the expense of fault recall; an appropriate operating threshold should be selected using validation data and the relative costs of these errors. The concentration of faulty samples near 1.0 indicates strong model confidence for many faults, but does not establish that the predicted probabilities are well calibrated.

Discussion The results demonstrate that the choice of model architecture substantially influences module-level photovoltaic fault classification under the evaluated conditions. ResNet18

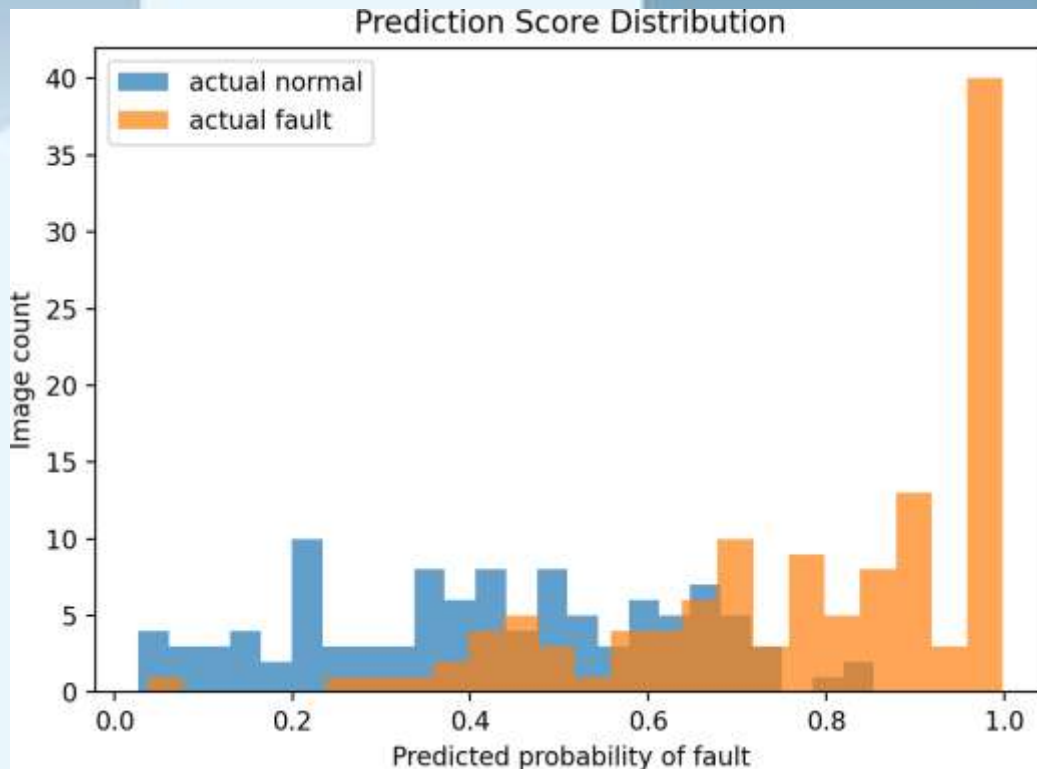


Figure 5. Distribution of predicted fault probabilities from the small CNN for actual normal and faulty test modules. Overlap between the class distributions illustrates the potential for false-positive and false-negative predictions at a fixed classification threshold.

achieved higher accuracy, balanced accuracy, F1-score and ROC-AUC than the compact 213 CNN. Its more balanced class-level performance suggests better separation of normal and 214 faulty thermal patterns. However, the comparison does not isolate the contribution of 215 residual connections from other architectural differences, including network depth and 216 capacity. Consequently, the improvement should be attributed to the evaluated ResNet18 217 configuration rather than to a single architectural component.

The principal improvement concerns the recognition of healthy modules. ResNet18 219 reduced false positives from 39 to 10, corresponding to approximately 74% fewer false alarms, while false negatives decreased only from 16 to 14. Specificity increased from 63.2% to 90.6%, whereas fault recall remained relatively similar at 87.1% and 88.7%. This distinction is relevant to maintenance planning: reducing false alarms can decrease unnecessary follow-up inspections, but the remaining missed faults limit the completeness of automated screening. These operational

benefits are potential implications of the classification results; inspection costs and maintenance outcomes were not measured directly.

The learning curves provide additional insight into the observed performance. ResNet18 exhibited improving training performance and a moderate train–validation gap, consistent with some overfitting. The compact CNN showed lower training accuracy and greater validation fluctuations, suggesting possible limitations in representation capacity or optimization. Nevertheless, learning curves from individual runs cannot establish the cause of these differences conclusively. Repeated training runs and controlled architectural experiments would help distinguish capacity limitations from sensitivity to initialization and training configuration.

The ROC analysis further indicates that ResNet18 provides stronger discrimination across classification thresholds, with an AUC of 0.960 compared with 0.888 for the small CNN. The small CNN’s score distribution shows overlap between healthy and faulty modules, explaining why a fixed threshold produces both false alarms and missed faults.

Threshold selection should therefore reflect the intended application. A lower threshold may be appropriate when missing a fault is particularly costly, whereas a higher threshold may reduce unnecessary inspections. Any threshold optimization or probability calibration should use validation data, with the test set reserved for final evaluation.

Within a drone-assisted monitoring workflow, the classifier could support maintenance prioritization by identifying module crops that warrant further examination. Its outputs could subsequently be associated with module identifiers, geographic coordinates, inverter records and electrical measurements. Such integration would allow suspected thermal anomalies to be cross-checked against operating conditions and production data. The present experiment evaluates image classification, however, and does not demonstrate closed-loop energy control or quantify recovered energy. In particular, ResNet18’s 11.3% fault miss rate supports its use as a screening aid alongside complementary inspection methods.

Limitations and future work Several limitations constrain the interpretation and generalizability of these findings. First, evaluation used a single-source dataset and a held-out test set of only 230 images. Although confidence intervals describe sampling uncertainty, they do not capture variation across sites, cameras, inspection conditions or training runs. External validation on independently acquired data is needed to assess transferability.

Second, the original fault categories were merged into a binary label. This formulation supports initial screening but does not identify fault type, severity or the appropriate corrective action. Aggregate fault recall may also conceal differences among the original categories. Future studies should report category-specific screening performance and investigate fine-grained diagnosis when sufficient examples are available.

Conclusion This study evaluated a compact CNN and ResNet18 for binary fault classification using drone-acquired infrared images of photovoltaic modules. Under the shared experimental protocol, ResNet18 achieved 89.57% test accuracy, 89.58% weighted F1, 0.9599 ROC-AUC and 0.9624 average precision, outperforming the compact CNN on all reported aggregate metrics.

The main benefit was a reduction in false alarms rather than a substantial increase in fault recall. ResNet18 correctly recognized 96 of 106 healthy modules, compared with 67 for the small CNN, while detecting 110 of 124 faulty modules, compared with 108. These results support ResNet18 as a promising baseline for prioritizing module inspections within a drone-assisted monitoring workflow. The remaining 14 missed faults nevertheless show that additional diagnostic evidence is needed for comprehensive assessment.

The findings are limited to binary classification of cropped images from one dataset and do not establish generalization to other installations, direct energy-loss estimation or automated energy control. Future work should prioritize independent validation, leakage-resistant data partitioning, repeated experiments, fault-specific analysis and integration with electrical monitoring. These steps would clarify the practical value of infrared computer vision for photovoltaic maintenance and provide a stronger foundation for subsequent automation.

References

- [1] C. Schaefer, “sUAS Infrared Image Dataset for Photovoltaic Module Fault Classification,” Zenodo, dataset, version 1, 2026. doi: 10.5281/zenodo.20860870.
- [2] L. Bommers, T. Pickel, C. Buerhop-Lutz, J. Hauch, C. Brabec, and I. M. Peters, “Computer vision tool for detection, mapping, and fault classification of photovoltaics modules in aerial IR videos,” *Progress in Photovoltaics: Research and Applications*, vol. 29, no. 12, pp. 1236–1251, 2021. doi: 10.1002/pip.3448.
- [3] Raptor Maps, “InfraredSolarModules,” GitHub repository, dataset, 2020. [Online]. Available: <https://github.com/RaptorMaps/InfraredSolarModules>.
- [4] B. Wang, Q. Chen, M. Wang, Y. Chen, Z. Zhang, X. Liu, W. Gao, Y. Zhang, and H. Zhang, “PVF-10: A high-resolution unmanned aerial vehicle thermal infrared image dataset for fine-grained photovoltaic fault classification,” *Applied Energy*, vol. 376, Art. no. 124187, 2024. doi: 10.1016/j.apenergy.2024.124187.
- [5] M. W. Akram, G. Li, Y. Jin, X. Chen, C. Zhu, and A. Ahmad, “Automatic detection of photovoltaic module defects in infrared images with isolated and develop-model transfer deep learning,” *Solar Energy*, vol. 198, pp. 175–186, 2020. doi: 10.1016/j.solener.2020.01.055.
- [6] W. Ahmed, A. Hanif, K. D. Kallu, A. Z. Kouzani, M. U. Ali, and A. Zafar, “Photovoltaic panels classification using isolated and transfer learned deep neural models using infrared thermographic images,” *Sensors*, vol. 21, no. 16, Art. no. 5668, 2021. doi: 10.3390/s21165668.
- [7] R. H. Fonseca Alves, G. A. de Deus Júnior, E. G. Marra, and R. P. Lemos, “Automatic fault classification in photovoltaic modules using convolutional neural networks,” *Renewable Energy*, vol. 179, pp. 502–516, 2021. doi: 10.1016/j.renene.2021.07.070.
- [8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [9] B. W. Matthews, “Comparison of the predicted and observed secondary structure of T4 phage lysozyme,” *Biochimica et Biophysica Acta*, vol. 405, no. 2, pp. 442–451, 1975. doi: 10.1016/0005-2795(75)90109-9.

- [10] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960. doi: 10.1177/001316446002000104.
- [11] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *Journal of the American Statistical Association*, vol. 22, no. 158, pp. 209–212, 1927. doi: 10.1080/01621459.1927.10502953.
- [12] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982. doi: 10.1148/radiology.143.1.7063747.
- [13] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, Art. no. e0118432, 2015. doi: 10.1371/journal.pone.0118432.